To Stand the Test of Time

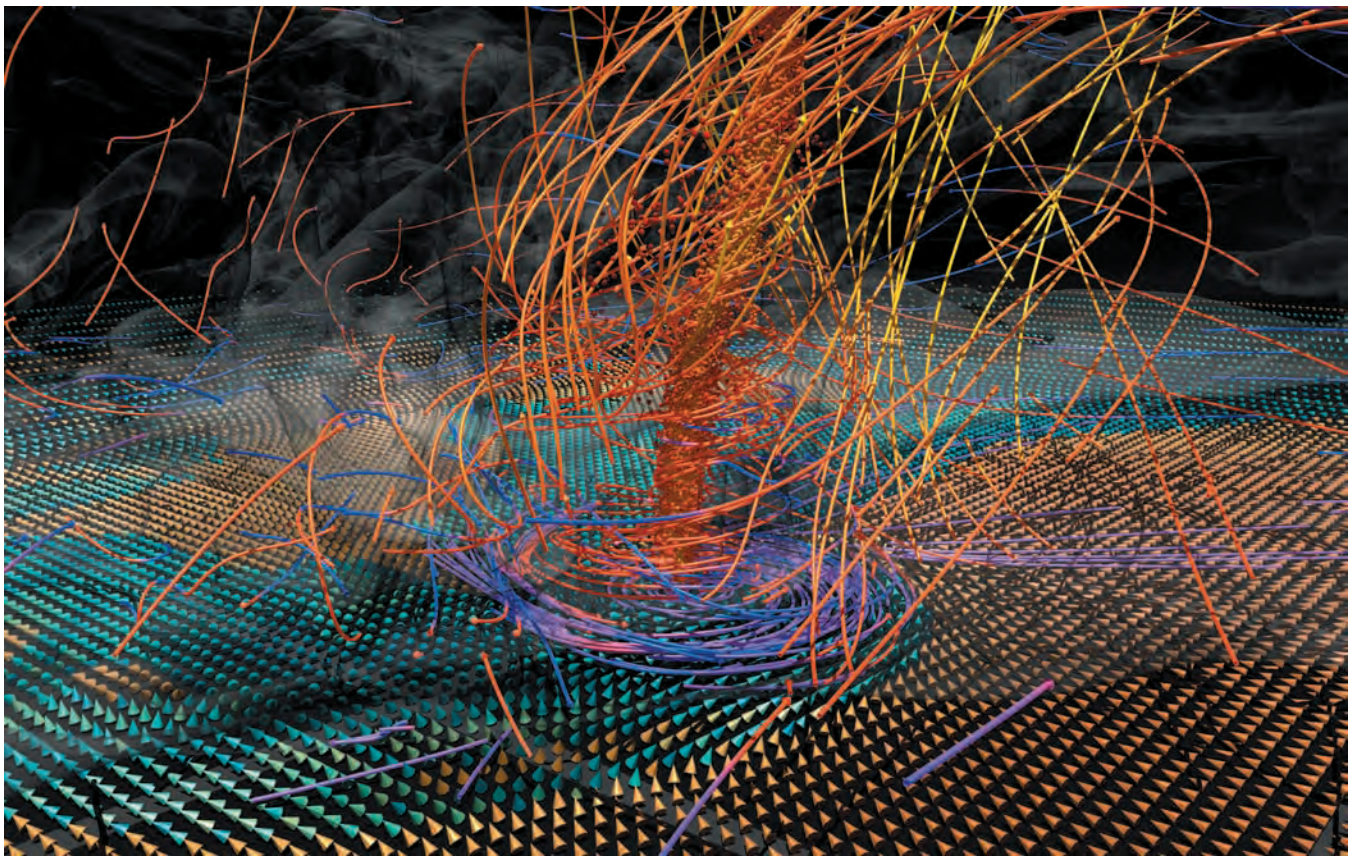
Long-term Stewardship of Digital Data Sets in Science and Engineering

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

September 26–27, 2006
Arlington, VA

"Nature, to be commanded, must be obeyed."

Attributed to Francis Bacon (1561–1626)
Novum Organum, bk.1, aph. 129 (1620)



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

September 26–27, 2006

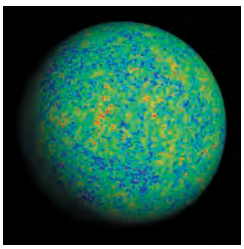
Arlington, VA

Support for this workshop was provided by the National Science Foundation

Disclaimer

The views and opinions expressed herein represent a summary of the consensus of the workshop participants and do not necessarily reflect those of the US Government, the National Science Foundation, or the Association of Research Libraries.

Layout and Design by Lee Anne George, ARL.



Contents

List of Figures

List of Boxes

Photo Credits

Cover: Tornado Simulation

This visual was created from data generated by a tornado simulation calculated on the National Center for Supercomputing Applications (NCSA) IBM p690 computing cluster. High-definition TV animations of the storm produced at NCSA were included in an episode of the PBS TV series NOVA called "Hunt for the Supertwister." The tornado is shown by spheres that are colored according to pressure: orange and blue tubes represent the rising and falling airflow around the tornado. (Date of Image: Sept. 5, 2004)

Credit: Bob Wilhelmson, NCSA and the University of Illinois at Urbana-Champaign; Lou Wicker, National Oceanic and Atmospheric Administration's National Severe Storms Laboratory; Matt Gilmore and Lee Cronce, University of Illinois atmospheric science department. Visualization by Donna Cox, Robert Patterson, Stuart Levy, Matt Hall and Alex Betts, NCSA.

Courtesy: National Science Foundation

Page 5: Universe in a Sphere

Our entire observable universe is inside this sphere, with a radius 13.3 billion light years, with us at the center. Space continues outside the sphere, but an opaque glowing wall of hydrogen plasma hides it from our view. This is our best image so far of what this plasma sphere looks like after cleaning out Galactic radio noise from the WMAP satellite observations. This image is from Phys. Rev. D, 68, 123525, supported by National Science Foundation grants AST 00-71213, AST 01-34999 and AST 02-05981. (Date of Image: March 4, 2003)

Credit: Credit Max Tegmark, Angelica de Oliveira-Costa, and Andrew Hamilton

Courtesy: National Science Foundation

Page 9: Hubble Exoplanet Search Field in Sagittarius

This is an image of one-half of the Hubble Space Telescope field of view in the Sagittarius Window Eclipsing Extrasolar Planet Search (SWEEPS). The field contains approximately 150,000 stars, down to 30th magnitude. The stars in the Galactic disk and bulge have a mixture of colors and masses. The field is so crowded with stars because Hubble was looking across 26,000 light-years of space in the direction of the center of our galaxy.

Credit: NASA, ESA, K. Sahu (STScI) and the SWEEPS Science Team

Page 11: Sunset at the Palomar Samuel Oschin Telescope

For the past three years, astronomers at the California Institute of Technology's Palomar Observatory have been using the High Performance Wireless Research and Education Network (HPWREN) as the data transfer cyberinfrastructure to further our understanding of the universe. Recent applications include the study of some of the most cataclysmic explosions in the universe, the hunt for extrasolar planets, and the discovery of our solar system's tenth planet. HPWREN work is supported under National Science Foundation grants 00-87344 and 04-26879. (Date of Image: 2003-06)

Credit: Caltech, Palomar Observatory

Courtesy: National Science Foundation

Pages 15, 27, 39, 47, and 57: Margaret Murnane and Ultraviolet Light Source

Researcher Margaret Murnane of JILA (a joint institute of the University of Colorado and the National Institute of Standards and Technology) and the extreme ultraviolet light source she developed with Henry Kapteyn. This laser-like beam of light at wavelengths from 10 to 100 times shorter than visible light will enable researchers to "see" tiny

features and measure the fastest reactions in the microscopic world, with important applications in the development of ultrafast computers as well as nanoscale devices. (Date of Image: 2000)

Credit: Courtesy of the John D. and Catherine T. MacArthur Foundation

Courtesy: National Science Foundation

Page 65: Schlieren Texture of a Nematic Film

Liquid crystals are not quite liquid and not quite solid. Physically, they are observed to flow like liquids, but they have some properties of crystalline solids. Liquid crystals can be considered crystals that have lost some or all of their positional order, while maintaining full orientational order. Nematics are polarizable rod-like organic molecules on the order of 20 Angstroms in length. Because of their tendency to organize themselves in a parallel fashion, they demonstrate interesting and useful optical properties. Digital watches for example, functioned using nematic liquid crystals.This image was taken as part of research performed at the Science and Technology Center for Advanced Liquid Crystalline Optical Materials (ALCOM), which was located at the Liquid Crystal Institute, Kent State University (supported under National Science Foundation grant DMR 89-20147). (Date of Image: 1993)

Credit: Microphotograph courtesy Oleg D. Laventovich, Liquid Crystal Institute, Kent State University

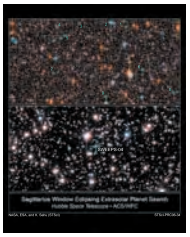
Courtesy: National Science Foundation

Pages 71, 75, 78, 104, 118, and 155: Earth's Magnetic Field

A geodynamo simulation computed on Pittsburgh Supercomputer Center's LeMieux supercomputer shows inward-directed magnetic field lines (blue) and outward-directed lines (red). The complicated field of the core becomes smoother at Earth's mantle. Gary Glatzmaier of the University of California, Santa Cruz and his colleague Paul Roberts of UCLA developed the first computational model of these geodynamic processes that evolves on its own self consistently. This model has successfully simulated many features of Earth's magnetic field, including magnetic-field reversal, a recognized phenomenon that has happened many times over Earth's history. (Date of Image: Sept. 5, 2004)

Credit: Gary Glatzmaier, University of California, Santa Cruz; Paul H. Roberts, UCLA; Darcy E. Ogden, UC Santa Cruz

Courtesy: National Science Foundation



Acknowledgements

Organizing an event such as this draws on the help and skills of many people and organizations. In particular, the Association of Research Libraries would like to acknowledge the logistical and IT support provided by the National Science Foundation (NSF), the accommodations provided by the Westin Hotel-Arlington, and the staff support of ARL itself. This project was funded by the NSF under Grant No. OCI-0638866 with additional support provided by ARL. ARL greatly appreciates the work of Francine D. Berman, Director, San Diego Supercomputer Center, and Wendy Lougee, University Librarian, University of Minnesota, who co-chaired this event. ARL thanks Robert Hanisch, Space Telescope Science Institute; Rick Luce, Vice Provost for Academic Information, Emory University Libraries; Brian E.. C. Schotttlaender, University Librarian, University of California, San Diego; and Chris Rusbridge, Director, Digital Curation Center, who graciously agreed to co-lead break-out groups; and Clifford Lynch, Executive Director, CNI, for summarizing key themes of the workshop. The workshop was organized and managed by Prudence Adler, ARL Associate Executive Director, with additional support from Mary Jane Brooks, ARL Executive Officer, Julia Blixrud, ARL Assistant Director, External Relations, Heather Joseph, Executive Director, SPARC, Sarah Segura, Executive Assistant, Crystal Blue, Projects Coordinator, and Lee Anne George, Publications Program Officer, of ARL, and Amy Friedlander and Amy Harbur of Shinkuro, Inc.

The organizers are profoundly grateful to the workshop participants who gave generously of their time and expertise to further the goal of building a sustainable framework for the future stewardship of digital scientific and engineering data.

The report was written by Amy Friedlander and Prudence Adler with contributions by the co-chairs and workshop participants. ARL is deeply indebted to Amy Friedlander for her assistance and support throughout the project.

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 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.

The background of the workshop is described in Chapter I. Descriptions of the discussions of the three major topics from the three breakout groups and in plenary sessions are provided in Chapters II, III, and IV, and Chapter V discusses additional topics raised in the plenary sessions and final recommendations. 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.*

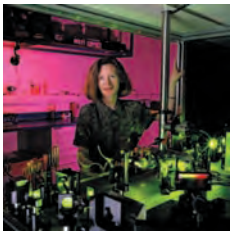
3. *NSF should support the development of usable and useful tools and automated services (e.g., metadata creation, capture, and validation) which make it easier to understand and manipulate digital data. Incentives should be developed which encourage community use.*

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.*



I. Introduction: Five Years and Five Centuries of Thinking

This is a report of a workshop that examined the new partnerships, infrastructure and sustainable economic models required to support long-term curation and management of scientific and engineering data as a critical component of the nation's cyberinfrastructure. It builds upon five years of careful thought and analysis by many research communities and reflects a concerted effort by scientists and librarians to evolve ways of collaborating to achieve the critical goal of digital data stewardship. This overarching goal is fundamental to the future of the scientific and engineering research enterprise and hence to innovation and competitiveness.

We are living through a revolution in the conduct of science and engineering, enabled by advances in computing and information technologies. It goes without saying that evidence based on both theory and data is fundamental to research. Reconciling the tension between theory and observation forms one of the major themes in the Scientific Revolution in the West, which may be said to have begun with the publication of Copernicus' *De Revolutionibus* in 1543.[1]

In our own time, many have pointed to computational and information technologies as having led to an advance in the scientific method through techniques such as simulation and visualization. Large-scale investigations such as those in genomic sequencing and protein folding and astronomical sky surveys have created data sets of a magnitude and granularity well beyond what might have been accommodated by paper and analog photography. Likewise, large databases have supported disparate and highly heterogeneous data for studies of history and culture, climate, geography, ecology and weather. In addition, expanding digitally based communication systems permits remote analysis and collaboration by distributed teams of investigators and new forms of dissemination within and across disciplines as well as to the public (Figure I-1). But with the reliance on digital data for scientific and engineering research, and the likelihood that such collections will proliferate, comes the need to manage the data, both to support the verification of published findings as well as to enable re-use of collections. Hence the need for sustainable economic and organizational models to support stewardship of digital data.

The National Science Foundation (NSF) has played, and continues to play, a major role in this intellectual revolution from supporting the Internet and supercomputing centers to sponsoring basic research. Over the past five years, the agency has convened numerous workshops[2] across the scientific disciplines to examine the notion of a "cyberinfrastructure," that is, the integrated "hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools."[3] The Foundation recognizes that the implications of providing this

infrastructure extend beyond physical facilities and software tools:

> Investments in interdisciplinary teams and cyberinfrastructure professionals with expertise in algorithm development, system operations, and applications development are also essential to exploit the full power of cyberinfrastructure to create, disseminate, and preserve scientific data, information and knowledge.[4]

These collaborations are perceived broadly to encompass stakeholders across government, the private sector, higher education and the research enterprise in the United States and internationally. A frequent theme in many of the workshop reports, as well as in the previously cited vision statement from the NSF Cyberinfrastructure Council, is the notion of data and its preservation. For example, as early as May 2001, NSF and the Office of Naval Research sponsored a workshop on marine geology and geophysics in La Jolla to examine issues related to data management.[5] The outcomes and recommendations are prescient in that they call for coordination of distributed centers (rather than centralization) and collaboration among the investigators and those who manage the collections. Thus, the challenges of managing highly heterogeneous digital data arise not only from the different disciplines and under-

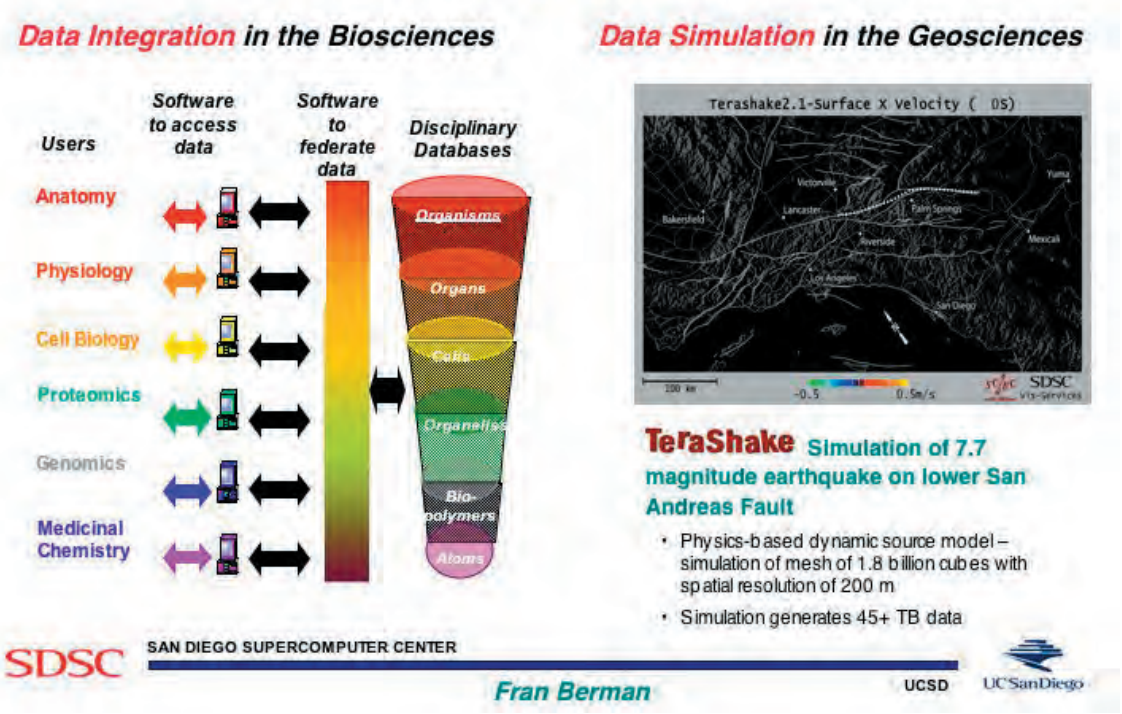


Figure I-1.
Working with Data: Data Driving New Discoveries in Research and Education
Source: F. Berman, Preserving Digital Collections for Research and Education, September 26, 2006

lying heterogeneity of the data—from sensors to censuses—but also from the different creator and user communities.

Prior workshop participants have grappled with these issues at some length, examining how cyberinfrastructure allows them to advance their research objectives as well as how best to manage the infrastructure itself.[6] On the basis of the past five years of study, the NSF Cyberinfrastructure Council has set a two-fold, five-year goal for data, data analysis, and visualization:

- To catalyze the development of a system of science and engineering data collections that is open, extensible, and evolvable; and

- To support development of a new generation of tools and services facilitating data mining, integration, analysis, and visualization essential for turning data into new knowledge and understanding.[7]

"The resulting national digital data framework," the Council's report continues, "will consist of a range of data collections and managing organizations, networked together in a flexible technical architecture using standard open protocols and interfaces, and designed to contribute to the emerging global information commons." As envisioned in the report, the national data framework will:

- Promote interoperability between data collections supported and managed by a range of organizations and organization types;

- Provide for appropriate protections and reliable long-term preservation of digital data;

- Deliver computational performance, data reliability and movement through shared tools, technologies and services; and

- Accommodate individual community preferences.[8]

This requires an over-arching coherent organizational framework engaging both collections and managing organizations, a flexible technical architecture, and coherent data policies.

The consequences for the research enterprise are profound. As Hey and Hey have recently observed in their paper on the "e-Science revolution" and its implications for stakeholders, including libraries:

> Increasingly academics will need to collaborate in multidisciplinary teams distributed across several sites in order to address the next generation of scientific problems. In addition, new high-throughput devices, high-resolution surveys and sensor networks will result in an increase in scientific data collected by several orders of magnitude. To analyze, federate and mine this data will require collaboration between scientists and computer scientists; to organize, curate and preserve this data will require collaboration between scientists and librarians.[9]

And, they continue, "A vital part of the developing research infrastructure will be digital repositories containing both publications and data." Thus, the transformation extends from the investigation, encompassing data collection and analysis, through communication of those results and the organizational settings in which these functions will be performed.

Data in Organizations

Curation and preservation (Box I-1) in the analog world was largely handled by libraries, archives, and museums, some specialized and others more general. The system was robust in the sense that there was substantial overlap and complexity in the underlying business models

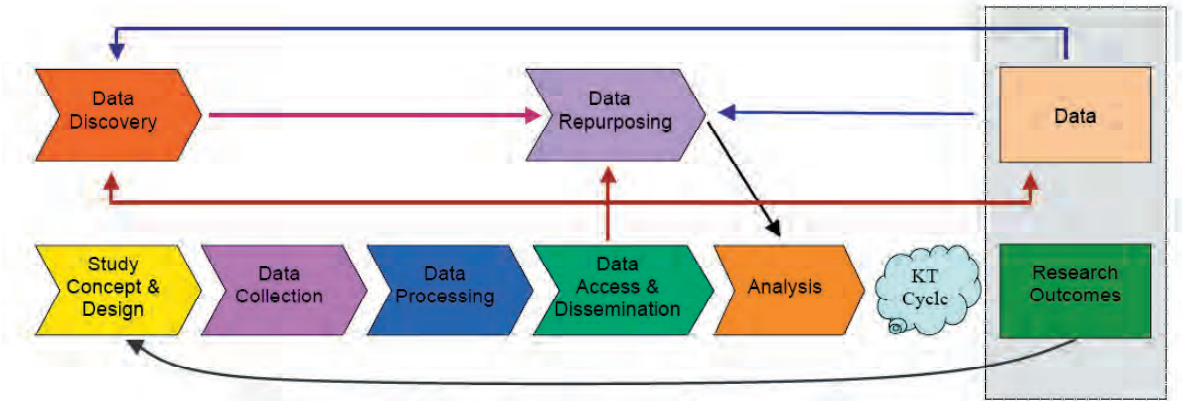
Box I-1. Stewardship, Preservation, and Curation
Charles Humphrey, University of Alberta

Stewardship of digital resources involves both preservation and curation. What is the distinction between these terms? Are these concepts describing the same thing? If not, how do curation and preservation differ? Both concepts have been borrowed from other fields and applied to the realm of research data. Curation has its roots in museum management, while preservation traces its origins to archivists. The Digital Curation Centre (DCC) in the U.K. defines digital curation as "maintaining and adding value to a trusted body of digital information."[10] DCC documents frequently make reference to "curation and preservation." That is, they treat these concepts as functionally different.

What are the functions that one would attribute to curation that differ from preservation? Preservation consists of (a) the management practices based on standards that guide and build metadata and data throughout the research life cycle and of (b) the subsequent long-term care for these digital products. The outputs of (a) are copies of the metadata and data in discipline-acknowledged standards best suited for (b), their long-term care, access, migration and refreshment.

Curation involves ways of organizing, displaying, and repurposing preserved data collections. Along the lines of the DCC definition, curation functions add value to a collection of preserved data by organizing and displaying the data through analyses of the collection's metadata or through the creation of new data from the preserved collection.

From the perspective of the life cycle of research data,[11] preservation occurs through the stages of data production and the creation of research outputs represented on the bottom row. Long-term preservation in this model consists of the practices followed in caring for the data, which is represented by the box on the right side of the figure. Data curation is characterized on the top row by the stages of data discovery and data repurposing, which make use of the preserved data. The activities of these two functions bring new value to the collection through analyses of the metadata, which display aspects of the collection in new light, and the creation of new data from the existing data collection.



and sources of support. However, the costs were frequently absorbed by organizations and entities that did not create or actually use the information.

Preservation of digital data, on the other hand, with its requirements for active management has forced a re-examination within the library/archives community of inherited assumptions about responsibility, use, oversight, and cost. Thus, long-term preservation and curation are understood not merely as preservation of bits and the ability to decode them but also as a system that requires both cooperation across a diversity of organizations, uses, and stakeholders and sustainable models of technical and economic support. In a nutshell, preservation is an organizational as well as a technical challenge and the responsibility, as has been widely recognized, spans a broad range of stakeholders.[12]

According to the definition of digital data adopted by the NSF Cyberinfrastructure Council, data are partitioned into three major categories: research collections, which are generally individualized, project-based, and perhaps not even candidates for long-term preservation; resource collections, which are community-based and mid- to long-term in anticipated longevity; and reference collections, which serve large segments of the research community, conform to agreed-upon standards, and require substantial and very long-term support.[13]

Chapter III of the draft report acknowledges the need for organizational and technical frameworks that reflect the diversity of data and of managing entities that can evolve as the data and technology evolve. These entities must maintain sufficient consistency, coherence, and interoperability to enable wide present and future use. More specifically, Berman has mapped a tri-partite data framework to the pyramid first proposed by Lewis Branscomb, enabling her to devise a data pyramid that links facilities, communities, and collections (Figure I-2).

The boundaries, particularly between the middle and top levels are understandably fluid. But, she points out, the framework allows funding sources, particularly public sources, to be targeted appropriately. Thus, Berman argues, commercial interests might serve the needs of small collections and "future NSF researchers and educators may request budgets for project-oriented storage services in the same way they currently request budgets for project-oriented personal computers." The middle level often involves partial federal investment, but invites "creative private/public/academic partnerships" like katrinasafe.com, a joint effort involving the International Red Cross, Microsoft, SDSC, and others. Finally, the top level, where national collections required to advance research and education are categorized, requires federal and sustained investment, again, in partnership with universities and other organizations. Even in this case, however, Berman outlines a case for interagency support and multiple funding strategies.[14]

Berman's model is an example of one way to parse the substantial challenge of massive, highly heterogeneous collections, diverse communities of creators and users, rapidly evolving technologies, and disparate organizations. Data are collected across a wide range of instrumentation and methodologies as well as across different disciplinary cultures. Consistently, those engaged in preservation of digital information have called for distributed yet federated collections that recognize heterogeneity yet preserve high-level coherence and support interoperability. Managing that tension is a fundamental requirement for the technical infrastructure and mandates partnerships across a range of stakeholder groups.[15]

To this effort requiring fusion of individuals, organizations, technology, and collections, libraries bring to bear not only their experience in managing physical collections but also their long experience building partnerships and

meeting the needs of diverse user communities. Indeed, the academic and research library community constitutes an information and social infrastructure that can be leveraged to support the needs of managing appropriate parts of this information infrastructure. Thus, while the workshop participants repeatedly called for active participation by scientists in the management of the data, creative partnerships between librarians and scientists will be critical for professional data stewardship in future science and engineering efforts. This workshop, therefore, assembled a range of investigators and librarians, and this report summarizes their deliberations and their recommendations.

Diversity in the Data Landscape

The data generated and used to support science and engineering research exist in many forms. In order for digital data to be analyzed, managed, curated, shared, or preserved, it is necessary to have a contextual understanding of the data concerning its capture, acquisition, or generation. Workshop participants acknowledged the importance of understanding the environment in which data are gathered and used in order to address how support for the digital data might be sustained over the long term.

Each discipline or sub-discipline has its own set of data characteristics and the type of research project will determine the level of complexity of the data. Some of the characteristics include the nature of the data (e.g., raw numbers such as time, position, temperature, calibration; images; audio or video streams; models; simulations or visualizations; software; algorithms; equations; animations), whether they can be reproduced, and whether they have

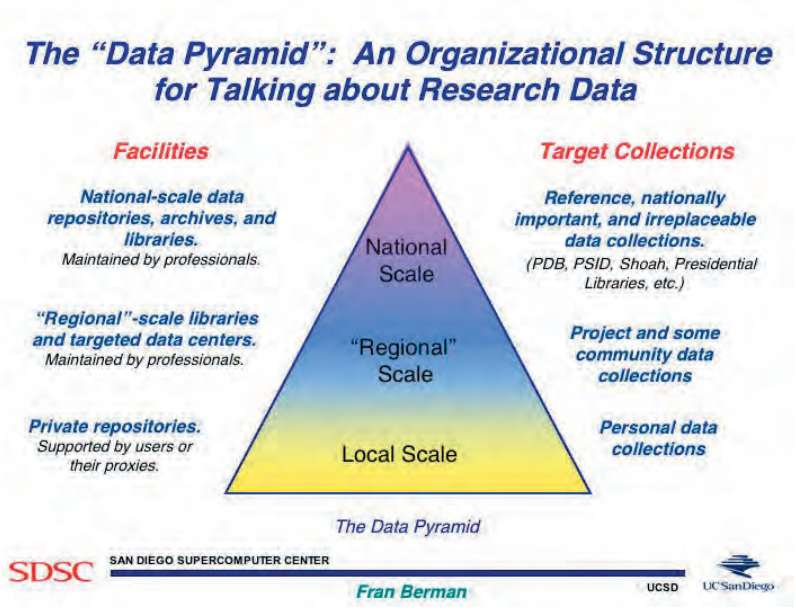


Figure I-2.
The Data Pyramid
Source: Francine Berman, Towards a Cost-Effective National Data Preservation Plan, June 2006, p. 2

been processed or interpreted by some means.

The size of the project also determines the heterogeneity of the data. Small projects can be managed with raw data in simple flat files and coding familiar only to the project investigator. Large projects must agree on data standards and curation procedures to make any progress.

However, the convergence of communication and computing technologies is providing new methodologies through which researchers gather and share larger amounts of instrumented and captured data. David Messerschmitt characterized what he called digital science as having "five complementary elements:

- Collection of data from the physical world (using distributed sensors and instruments);
- Distributed, organized repositories of such data;
- Computation using theoretical models and experimental data;
- Presentation of results for scientific visualization and interpretation; and
- Support for collaboration among scientists."

He goes on to write, "[d]igital science and engineering research often involves close coordination of theory, experiment, and collaboration among digital scientists, so geographically distributed collaboration and access to geographically distributed sensor networks and instrumentation are crucial. Much of digital science is conducted by authoring (or in many cases executing existing) discipline-specific and generic software that automates data collection and capture, computational models and data analysis, visualization of the results, and collaboration. Software is a primary tool of a digital scientist, just as microscopes and telescopes and pencil and paper are tools of experimental and theoretical scientists."[16]

As projects become larger and more complex, the data collected is several orders of magnitude higher and increases the need for coordination and data management. Databases and specially designed repositories are established to collect and make available the, in some cases, terabytes or petabytes of raw data gathered from such methods as sensor networks, satellite surveys, or supercomputer simulators.

According to *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, "[d]ata can also be distinguished by their origins—whether they are observational, computational, or experimental. Observational data, such as direct observations of ocean temperature on a specific date, the attitude of voters before an election, or photographs of a supernova are historical records that cannot be recollected."[17]

Computational data are usually the result of executing a computer model or simulation and generate outputs that can be reproduced (provided the model and its associated descriptive information about the hardware, software, and input data are preserved). Experimental data includes such things as measurements of patterns of gene expression, chemical reaction rates, or engine performance. In order to share data, they must be preserved unless determined to be cost-effective to reproduce the experiment.

> The experimental process is the origin of another distinction, in this case between the intermediate data gathered during preliminary investigations and final data. Researchers may often conduct variations of an experiment or collect data under a variety of circumstances and report only the results they think are the most interesting.

> Selected final data are routinely included in data collections, but quite often the intermediate data are either not archived or are inaccessible to other researchers.
>
> Processing and curatorial activities generate derivative data. Initially, data may be gathered in raw form, for instance as a digital signal generated by an instrument or sensor. [These] raw data are frequently subject to subsequent stages of refinement and analysis, depending on the research objectives. There may be a succession of versions. While the raw data may be the most complete form, derivative data may be more readily usable by others.[18]

All of the work to analyze and process the data is domain or application specific throughout the life of the project, but at the project's completion those data and their context, including provenance, need to be clearly articulated if the data are to be retained or shared.

Domains and applications often differ significantly, consequently, the increase in multidisciplinary projects and the associated use of pre-existing data highlight the importance of explicit descriptive information about data to optimize the investment made in data acquisition or generation. In a digital environment, data exist as bits or bytes, but without context they cannot be used. The data need descriptive information, tools, and frameworks in order to be fully useful for repurposing.

Although there are many similarities in the landscape of digital data, the specifics of managing data will depend significantly on the nature of the primary discipline with which the data are associated. An understanding of how the data are produced and structured for use prompts the methods available to curate and preserve the data for the long term.

Heterogeneous Data and Systems, Interoperability, and Metadata

As previously observed, scientific and engineering data sets are highly heterogeneous. This heterogeneity arises from multiple sources: formats, size, collection technique, coding, use of instrumentation. In addition, the delivery and storage formats and media are also variable—capturing data by satellite telemetry is fundamentally different from social science recorded from telephone surveys. Astronomical imagery looks little like a genomic database and both differ substantially from census and polling data. Other collections are quite small but no less valuable. Moreover, the structure of the data sets reflects differences in the disciplinary cultures in which they were created, including expectations about future use. Thus, the small scale project level collection of recordings of indigenous language might never have been intended for inclusion in a larger anthropological reference collection yet the rarity of the small collection as well as the endangered language itself might change the circumstances of the long term use. As several participants pointed out, the confidentiality that surrounds respondents in a social science survey is not an issue in astronomical imagery but both require systems that ensure the long-term integrity of the data. Moreover, the career goals of individual investigators, who legitimately need to protect their use of the data for some period of time, must be reconciled with its potential long-term value to the community and to future researchers.

This example, the tension between the prestige systems that surround individual investigators and the long-term importance of a collection, is very familiar to archivists, in particular. It is not uncommon that a small, important collection of important material must be integrated into a larger collection. Yet, the conditions under which the creator (or author or investigator) worked and the expectations

that future users—perhaps users a century off—may bring to bear are substantially different. Some of these tensions can be resolved through embargoes on use of the collections or uses of some parts of the collections. However, digital data are fragile. Thus, the management of the data requires active engagement by the investigators as well as the information managers. How this might work is open to discussion, but it is widely recognized that collaborations require recognition of the institutional practices within which collections may have been created as well as the requirements imposed by management itself. Thus, Green and Gutmann have argued a life cycle approach to data that understands reciprocal relationships and information flows between investigators and managers, enabling both sides to see where the hand-offs occur and how the cultural conditions of the respective environments can been respected and accommodated.[20]

At the technical level, designers of repository architectures, the umbrella term used to embrace the conceptual structure of a data storage system and the components required to organize, manage, and provide access to them, have focused on ways to make heterogeneous data and systems interoperate. The goal is not to compel homogeneity but rather to devise layers on top of the data in their native or "raw" form (perhaps as transmitted by the instrumentation or collected by the investigator). This "layer" is the metadata, or data about data. The most familiar example is the bibliographic record in a card catalog, but computational metadata is much more extensive and can include information about formats, structure, computer language, operating system, experimental environment, and so on. Thus, interoperability among data, collections, systems, and institutions is inevitably linked to metadata.

Metadata standards are a critical part of digital information. Given what some have called a "deluge" of digital information, substantial effort must be invested in automatic metadata creation, particularly when it is a by-product of the data collection workflow itself (as in astronomy, for example). Others have focused on the role of investigators who might be incentivised to undertake initial metadata creation according to an agreed upon community standard; both metadata and incentives for investigators to participate in long-term curation systems are among the themes that permeate the position papers submitted by participants in this workshop. Addressing this topic is critical since, as librarians and archivists have long argued, metadata in the digital environment must begin at creation precisely to address questions that might be most easily and accurately addressed at that point (for example, the nature of the instrumentation or the time of day for some types of data). In this regard, many involved in the creation of digital libraries and archives have pointed to the need for automatic metadata generation, which is, in many classes of scientific data (for example, sensor data or digital imagery), inherent in the process of data collection.[19]

Workshop Goals and Description

Based on the framework established by prior work for the Office of Cyberinfrastructure, the workshop outlined three major 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 identifying 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 manage-

ment of digital data in distributed yet federated collections, recognizing the rapid pace of technological change and the need for unfettered access.

Threaded through these goals were a series of questions:

- What role do the research and academic libraries envision for themselves and do scientists envision for librarians in a digital data framework that provides for preservation of digital publications, digital data, and the links between them?
- What partnerships/coalitions among research and academic libraries and with other sectors (government, international, non-profit, and for-profit) could facilitate the creation of such a framework?
- How do new and emerging technologies affect organizational roles and responsibilities?
- Are there opportunities to test sustainable models for digital data preservation organizations, including consortia and partnerships?
- What resources would be required to enable such tests?

These goals and questions prompted a workshop structure based on three breakout groups with distinct charges:

- **Infrastructure.** How do we manage digital data now and migrate it successfully over future generations of technologies, standards, formats, and institutions?
- **Partnerships.** What mix of individuals and organizations should be involved in digital data preservation? What creative partnerships can be developed between the multiple sectors?
- **Sustainable Economic Models.** What models are required to sustain digital data management and preservation efforts over the long term?

Thirty-two individuals participated in the invitation-only, two-day workshop, held at NSF headquarters in Ballston, Virginia. A complete list of participants and the agenda are included among the appendices to this report. In advance of the meeting, workshop participants were asked to submit a brief statement describing their top three issues within the themes of the workshop. These statements were posted to the workshop Web site and analyzed in advance of the meeting as a point of departure for the deliberations. They are included in their entirety in Appendix E.

Participants were divided into three breakout groups, which articulated the issues and formulated recommendations. Workshop co-chairs were encouraged to ask their groups to outline well-formed, actionable recommendations and, where possible, to identify links with other topics and funding opportunities and implications. The first day was given over to plenary sessions and breakout group deliberations. On the second day, the breakout groups reconvened in plenary session to present their findings and entertain broad discussion.

Roadmap to this Report

This report is divided into three major chapters, each corresponding to the breakout groups. Each chapter is divided into three major sections: a brief summary of the issues and rationale for the topic, the discussions that surrounded the issues in the group and during the plenary session that followed on the second day, and then the recommendations in priority

order formulated by the breakout group. The final list of recommendations is presented in Chapter V, Summary, Conclusions, and Recommendations, together with a summary of the more general discussion.

Endnotes

[1] Marie Boas Hall, Introduction: The Scientific Revolution, in Marie Boas Hall, ed., *Nature and Nature's Laws; Documents of the Scientific Revolution* (New York: Harper & Row, 1970), 2.

[2] For a list of the workshops and workshop reports, see National Science Foundation, Reports and Workshops Relating to Cyberinfrastructure and Its Impacts, updated, July 12, 2006; http://www.nsf.gov/od/oci/reports.jsp. These are also identified in Appendix VI, Selective Bibliography.

[3] NSF Cyberinfrastructure Council, NSF Cyberinfrastructure Vision for the 21st Century Discovery (draft), Version 7.1, July 20, 2006, p. 6.

[4] Ibid., p. 6.

[5] Data Management for Marine Biology and Geophysics, Tools for Archiving, Analysis and Visualization, Workshop Report, La Jolla, California, May 14-16, 2001; http://humm.whoi.edu/DBMWorkshop/data_mgt_report.hi.pdf.

[6] See for example, Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences, F. Berman and H. Brady, available at www.sdsc.edu/sbe/.

[7] NSF Cyberinfrastructure Council, Vision for the 21st Century Discovery (draft), p. 20.

[8] Ibid. p. 21.

[9] Tony Hey and Jessie Hey, E-Science and its Implications for the Library Community, pp. 1–2.

[10] See "What is Digital Curation" on the Digital Curation Centre Web site at http://www.dcc.ac.uk/about/.

[11] The "KT Cycle" in the diagram represents the processes of knowledge transfer. This life cycle diagram comes from Charles Humphrey, "e-Science and the Life Cycle of Research" (2006) available online at http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc.

[12] Library of Congress, Preserving Our Digital Heritage, Plan for the National Digital Information Infrastructure and Preservation Program, A Collaborative Initiative of the Library of Congress, October 2002, p. 3.

[13] NSF Cyberinfrastructure Council, Vision for the 21st Century Discovery (draft), p. 18.

[14] Berman, pp. 4–5.

[15] Library of Congress, Preserving Our Digital Heritage, p. 45; see also NSF Cyberinfrastructure Council, Vision for the 21st Century Discovery (draft), pp. 21–22.

[16] David G. Messerschmitt. "Opportunities for Research Libraries in the NSF Cyberinfrastructure Program," *ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC*, no. 229 (August 2003): 2. http://www.arl.org/newsltr/229/cyber.html.

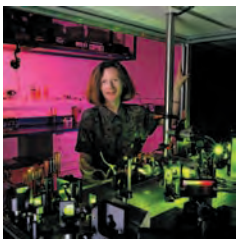
[17] National Science Board. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. (Arlington, VA: National Science Foundation, September 2005), 19. http://www.nsf.gov/pubs/2005/nsb0540/start.htm.

[18] Ibid.

[19] Metadata is typically defined as "data about data." The structure of metadata records has occasioned substantial research and discussion and various schema have been put forth; good introductions to these resources can be found at "Metadata and Resource Description" <http://www.w3.org/Metadata/>, a site maintained by the World Wide Web Consortium (W3C); "Understanding Metadata" <http://www.niso.org/standards/resources/UnderstandingMetadata.pdf>, a resource developed under the rubric of the National Institute of Standards; "Dublin Core Metadata Initiative" <http://dublincore.org/>; and "Introduction to Metadata: Pathways to Digital Information" <http://www.getty.edu/research/conducting_research/

standards/intrometadata/index.html>. Related to the formal structure of metadata is the notion of an ontology, which deals with the semantics of a description. Although it is a term that originated in philosophy, it has a formal and different definition in artificial intelligence where, in a general sense, an ontology can be understood to mean a common vocabulary with which queries and assertions are exchanged among agents. Thus, the purpose of creating ontologies is to enable knowledge sharing by establishing consistency. For an introduction, see Tom Gruber, "What is an Ontology?" <http://www.ksl.stanford.edu/kst/what-is-an-ontology.html>; the W3C's "Web-Ontology (WebOnt) Working Group" <http://www.w3.org/2001/sw/WebOnt/>; and the W3C's "Web Ontology Language (OWL)" <http://web4.w3.org/2004/OWL/>. Both metadata and ontologies, as the workshop participants acknowledged, are important to resource description, management, discovery, and re-use.

[20] Ann Green and Myron P. Gutmann, Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives, http://hdl.handle.net/2027.42/41214.



II. Infrastructure

The breakout group on infrastructure took its charge to be as follows: What capacities should we build now to manage and migrate data over future generations of technologies, standards, formats, and organizational stakeholders?

This chapter summarizes the discussion of the issues undertaken by the breakout group and sets forth its recommendations.

Discussion of the Issues

As Berman pointed out in her introduction to the workshop, data management and preservation have multiple layers that require vertical integration between layers and horizontal coordination across collections, disciplines, and institutions, all to support active use of the information (Figure II-1).

These functions require specific attention to the technical issues associated with interoperability across heterogeneous networks, services, platforms, and data; metadata to support access, interoperability, and long-term management and curation (including rights management, security, privacy, and confidentiality); and institutional policies to support these functions as well as collaborations among institutions, disciplines, and investigator teams. Moreover, data preservation has distinctive requirements for resources, continuity, metrics of success, and funding:

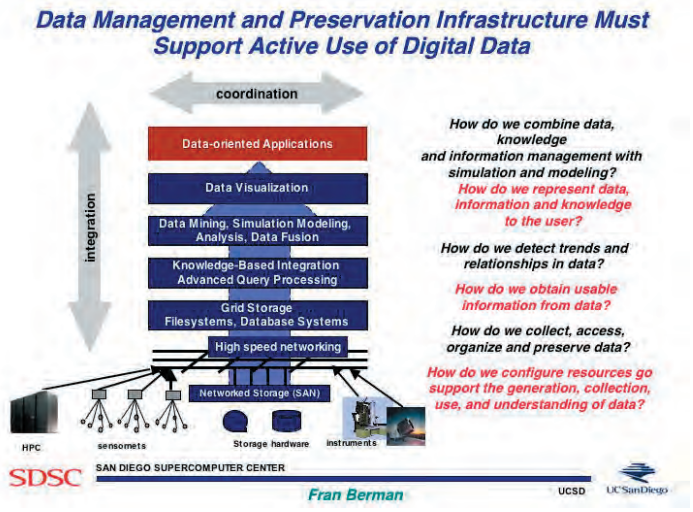


Figure II-1.
Data Management and Preservation Infrastructure Must Support Active Use of Digital Data

- Resources: Archival storage, network, systems, replication and backup, staff for maintenance, database management, and user services.

- Continuity: A key criterion. Data collections must migrate over new generations of technology seamlessly, without loss of information or interruption in service.

- Metrics of success: No serious loss of data, preservation of reference collections, appropriate research collections.

- Funding model: Funding commitment needs to address long-term consistency needed for data collections.

Within the breakout group, participants examined their assumptions, requirements, and values about the infrastructure—i.e., that shared layer that supports multiple and divergent collections and users within and across organizations. They concluded:

- Infrastructure is a shared common good; the data are a public good.

- Infrastructure must:

  - Support multiple representations of data by diverse users, now and in the future, representing diverse disciplinary cultures;

  - Support new service layers and reuse of components through data models and metadata as well as institutional policies governing use of collections and services;

  - Support appropriate security systems, privacy, confidentiality; and

  - Enable trusted relationships at multiple levels.

Heterogeneity of data and collections and the constraints potentially associated with using those collections (for example, rights management) were considered a particular challenge as was the need to incentivise investigators both to contribute material and to undertake some of the critical processing activity (e.g., preparation of metadata). Indeed, the question of how to construct appropriate incentives and the possible use of mandates struck a chord in the plenary session and will be discussed in more detail in Chapter V in the context of consensus recommendations.

The OAIS Architecture

The breakout group found the Open Archival Information System (OAIS) model[1] to be a useful mechanism for focusing and organizing the discussion. The reference model as developed over a number of years by the Space Science community and others has proven to be a useful mechanism for preservation within a number of communities. As pointed out by one participant, a key concept is the notion of a "designated community," which explicitly acknowledges the institutional culture within which data are created and used. This meets a critical need: namely cross-cultural communication among professions (library, archival, and research) institutions (libraries, archives, universities, professional societies, and so on) and disciplines (physical sciences, life sciences, computational sciences, and social sciences). Specifically, the notion of a designated community recognizes the specifics of a given discipline and its data requirements yet enables the community to conceptualize a common archival framework.

Two elements in the reference architecture proved especially useful: the environment

within which the archive (or curatorial facility) is located[2] (Figure II-2) and the major functions of the archive itself[3] (Figure II-3), which may be embodied in both centralized and distributed organizational and computational architectures. The focus on function thus enables discussion of capacity, independent of the specific institutional arrangements (the topic of the Partnerships breakout group) and recognizes that multiple instantiations may be possible.

Four functions were especially relevant to exploring data curation functions:

- Ingest: Which includes receiving information in a specified form, known as the "Submission Information Packages;" quality assurance; generating the form of the information that is actually stored in the archive, which is called the "Archival Information Packages;" extracting descriptive information; and coordinating updates.

- Archival storage: Which include services to support storage, maintenance, and retrieval of the archived information. Functions in-

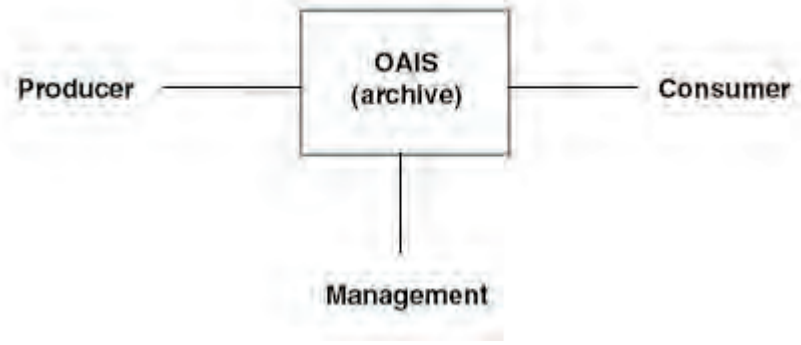


Figure II-2. Environment Model of an OAIS
Source: Consultative Committee for Space Data Systems, Figure 2-1, p. 2-2

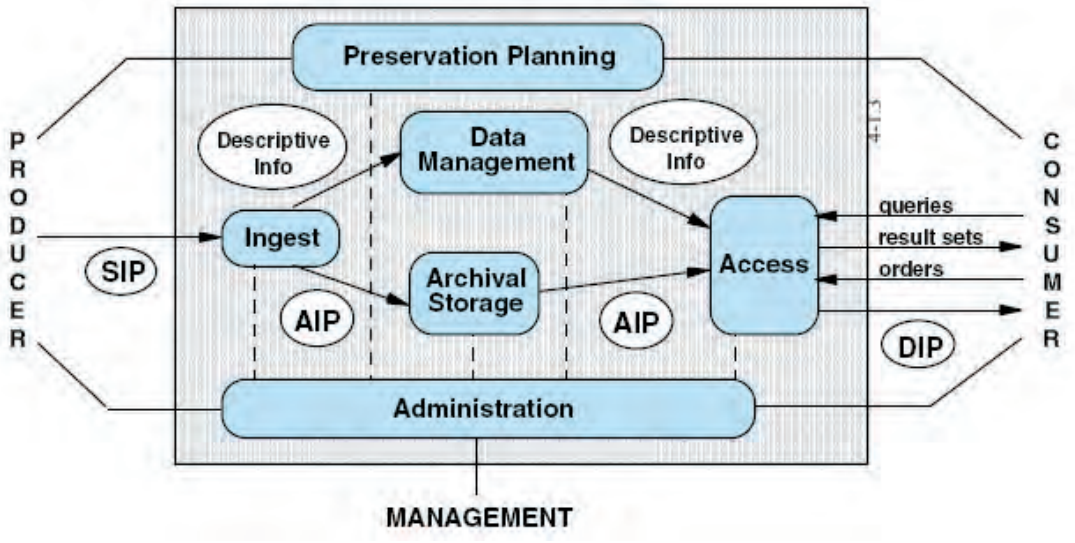


Figure II-3. OAIS Functional Entities
Source: Consultative Committee for Space Data Systems, Figure 4-1, p. 4-1

clude receiving the information that has been prepared for archiving, refreshing the media as needed, routine monitoring for errors, providing for disaster recovery, and fulfilling requests for access to the data.

- Data management: Which includes the services and functions for population, maintaining and accessing the descriptive information important for identifying and documenting the holdings and information necessary to manage the archive. Functions include database administration, performing updates, and querying the management data to generate reports required to manage the archive.

- Access: Which includes services and functions required to support users, enabling them to find information, including discovery, description, location, and availability.

Two other functions, preservation planning and administration, are included in the model but did not generate substantial discussion in the breakout group. They are included here for completeness.

- Preservation planning: Which includes services and functions required to monitor the environment (for example, format obsolescence, standards, and so on), enabling managers to ensure that the information remains accessible. Planning functions include developing migration plans and related prototyping and implementation.

- Administration: Which includes services and functions enabling overall operation of the archive.

Finally, all of these functions are supported by a set of common services. These include: operating systems, network services, and security (encompassing identification, authentication, access control, data confidentiality, data integrity, and non-repudiation).

Discussion of the OAIS Framework

Within this framework, the participants in the breakout group agreed that ingest and access were the critical functions to address. The issues include not simply what metadata are needed, but incentives to ensure the creation of necessary metadata. Storage, while important and challenging, did not present the same ambiguities and tensions associated with access or ingest. One participant observed, "Storage and computers are so cheap, I can buy several of them and then I can archive terabytes of data." Other disciplines like astronomy have far greater storage requirements and the scale of stored data presents challenges to access. The group agreed that issues related to storage (in addition to those identified in the OAIS reference implementation) include proliferation (storage is cheap, anyone can do it), scalability, heterogeneity in media and formats, and architecture (centralized versus distributed).

Metadata. Perhaps not surprisingly, metadata are critical not only to enabling access by consumers but also to the effective management of the data. Metadata schema are increasingly complex. For example, the schema for data preservation set forth by the National Library of New Zealand[4] has over 70 data elements, divided into four major categories or "entities" (Figure II-4).

Incentives. The ideal of capturing metadata at the time of creation requires both techniques of automatic metadata creation as well as the active engagement of the investigators themselves. The latter, in turn, requires both motivating the investigator (e.g., through increasing awareness of the imperative) as well as the development of tools to support the investiga-

tor's active participation. In addition, the OAIS model acknowledges that the object to be managed (the "package") might undergo transformations, so that the package of information submitted by the investigator might look different than the package that is managed inside the archive. Thus, the model does not require that the metadata that investigators might submit meet the standard imposed by the archive. Rather, the investigators need only meet minimal standards.

Repositories. Breakout group participants agreed that it is insufficient to outline incentives for investigators without ensuring that repository capacity is available to receive their data. Such repositories, largely for e-prints and publications, already exist in a number of disciplines (see Box II-1) although their existence tends not to be well known outside of their respective communities.

Such resources vary by discipline. In some research areas, one participant pointed out, access to shared repositories is central to the nature of the research. An obvious example is the Protein Data Bank (http://www.wwpdb.org/) which is an international collaboration and is sponsored in the USA by nine federal agencies. At the other extreme is the individualized approach characteristic of much social science research where data formats together with changes in the underlying phenomena (for example, geopolitical boundaries) may substantially complicate re-use of earlier data as well as justify a fresh round of data collection. Still, the ability to compare data over time, a hallmark of some questions in social research, would be greatly enhanced by continued access to and reuse of older data, not to mention potential efficiencies since it is generally known that data collection and cleaning can be extremely expensive. However, the requirements of man-

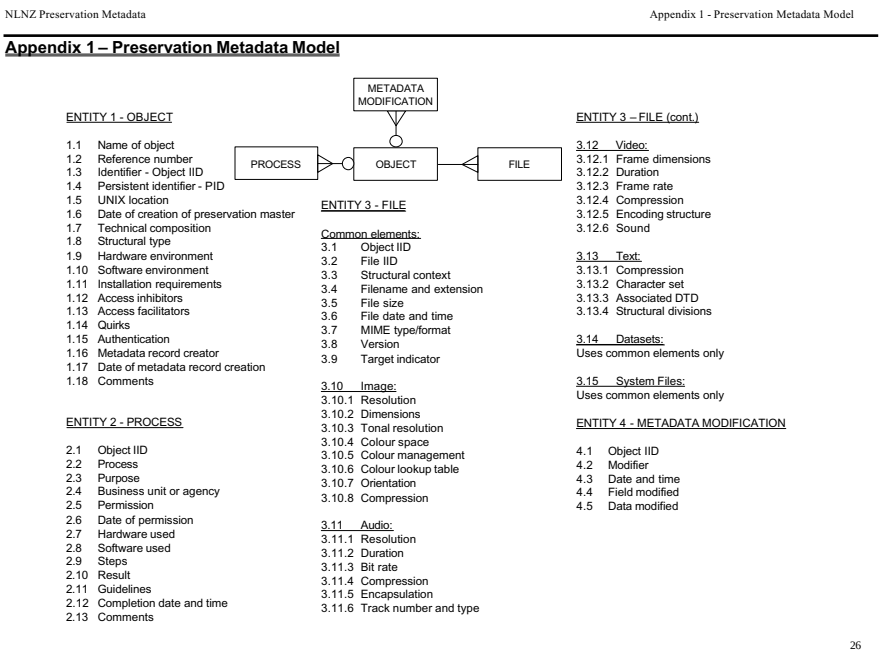


Figure II-4.
The National Library of New Zealand Preservation Metadata Model
Source: National Library of New Zealand. Metadata Standards Framework—Preservation Metadata (Revised), June 2003, Appendix 1

aging each of these types of data may be substantially different. Incentives may be useful in some cases, while insufficient in others.

The Importance of Prototypes: The National Virtual Observatory

Several of the position papers pointed to the importance of prototypes as a way of understanding aspects of curation ranging from interactions with the community, to business models, to technical architectures and services. One important example is the National Virtual Observatory Project (NVO), which is supported by the NSF and led by Johns Hopkins University, California Institute of Technology, and the Space Telescope Science Institute and engages numerous partners in academic institutions, publishing, and research centers in the US and internationally. Until recently, the project has not undertaken data preservation efforts, but a new grant from the US Institute of Library and Museum Services (IMLS) will enable it to develop prototype projects in data curation: capturing, curating, preserving, and providing access to the wealth of data that the NVO has accumulated and makes available to investigators. Goals for the new work include assessing scientific impact as well as working out sustainable business models (see Chapter V for more discussion about issues surrounding economic sustainability.).[5]

The NVO is a coalition of archives and research projects based at different facilities and includes diverse instrumentation. The core of the NVO's structure is interoperability enabled by common metadata standards that allow individual archives and collections to be made visible to the end user through a common portal or application. The underlying heterogeneity is thus masked to the user but preserved as

Box II-1. Data archives and repositories (not comprehensive, for illustration only)

Cambridge Structural Database [crystal structures]
http://www.ccdc.cam.ac.uk/products/csd/

CISTI Depository of Unpublished Data
http://cisti-icist.nrc-cnrc.gc.ca/cms/unpub_e.html

CrossFire Beilstein [chemistry]
http://www.mdl.com/products/knowledge/crossfire_beilstein/

Digital Archive Network for Anthropology and World Heritage
http://dana-wh.net/home/

Earth Observing System Data Gateway
http://redhook.gsfc.nasa.gov/%7Eimswww/pub/imswelcome/

Global Biodiversity Information Facility [prototype data portal]
http://www.europe.gbif.net:80/portal/index.jsp

Global Change Master Directory
http://gcmd.nasa.gov/

Goddard Earth Sciences Data and Information Services Center
http://daac.gsfc.nasa.gov/

Inter-university Consortium for Political and Social Research (ICPSR)
http://www.icpsr.umich.edu/

IRI/LDEO Climate Data Library
http://ingrid.ldgo.columbia.edu/

IRIS (Incorporated Research Institutions for Seismology)
http://www.iris.edu/data/data.htm

Land Processes Distributed Active Archive Center (LP DAAC)
http://edcdaac.usgs.gov:80/main.asp

Marine Geoscience Data System
http://www.marine-geo.org/

Multimission Archive at STSci [Astronomy]
http://archive.stsci.edu/

NASA's High Energy Astrophysics Science Archive Research Center
http://heasarc.gsfc.nasa.gov/

NASA Space Science Data Archives
http://science.hq.nasa.gov/research/space_science_data.html

National Center for Atmospheric Research & the UCAR Office of Programs
http://www.ucar.edu/tools/data.jsp

National Center for Biotechnology Information
http://www.ncbi.nlm.nih.gov/

National Center for Ecological Analysis and Synthesis (NCEAS) Data Repository
http://knb.ecoinformatics.org/knb/style/skins/nceas/index.html

National Oceanographic Data Center
http://www.nodc.noaa.gov/

National Virtual Observatory Project
http://www.us-vo.org/

Oak Ridge National Laboratories DAAC
http://www-eosdis.ornl.gov/holdings.html

Statistical Reference Datasets
http://www.itl.nist.gov/div898/strd/

Statistics Canada
http://www.statcan.ca/start.html

TranStats [Transportation]
http://www.transtats.bts.gov/DataIndex.asp

U.S. Department of Agriculture Economic Research Service Data Sets
http://www.ers.usda.gov/data/

U.S. Geological Survey, National Satellite Land Remote Sensing Data Archive
http://edc.usgs.gov/archive/nslrsda/

Worldwide Protein Data Bank
http://www.wwpdb.org/

relevant to the individual project or archive. Investigators publish their findings in a range of publications, which have actively cooperated in the NVO. While the journals do release a subset of the data (typically in images and tables) the underlying observational data rarely appear in the journals, in part because investigators themselves are reluctant to release work that is still in progress. Yet, these results, however preliminary, are clearly of use to future investigators, hence the need for a close coupling between the archive and the scholarly communication process (a point also made in Chapter IV, New Partnerships).

The architecture, which reflects both the roles of partners and the technological capacities, is shown in Figure II-5. In technical terms, the architecture combines NVO Web services with a distributed repository system based on Fedora in a framework that supports long-term digital archiving of astronomical derived data in a variety of formats (for example, tables, catalogs, spectra, images, and documents). The distributed nature of this preservation framework is significant because it acknowledges the importance of multiple parties being responsible for different functions, depending on their relative expertise. The modular nature of the technical system bolsters the ability to support different components and elements over time without requiring expensive or difficult overhauls of the entire system.

The architecture is clearly compatible with key elements of the OAIS architecture. It is based on a designated community, namely astronomers and related institutions (libraries, journals); it identifies an environment within which the specific curatorial and archival functions occur, namely the library (which is, in fact, a group of libraries that operate cooperatively); and it allows for coherent access by several stakeholder groups (the publishers, the scientists who deposit data, and the researchers who use data from the collections); and it provides for storage. The "storage" is actually a set of services and facilities that function over a set of repositories that accommodate disparate and heterogeneous data that share com-

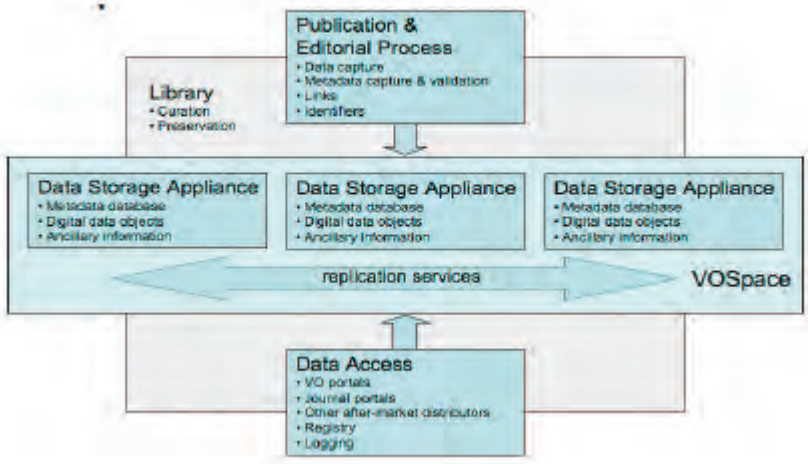


Figure II-5.
National Virtual Observatory Architectural Components
Source: S. Choudhury et al., Digital Data Preservation in Astronomy, 2nd International Digital Curation Conference: Digital Data Curation in Practice, Glasgow, UK, November 21–22, 2006.

mon metadata standards. The storage function is replicated to improve performance and enhance stability and security of the system.

Recommendations of the Breakout Group

The group recognized the need for a change in the way that NSF and other research agencies structure their calls; hence there is a need for a change in the culture of the funding agencies relative to data stewardship, as well as for changes in the culture of the research enterprise that these agencies support. One participant commented, "The large goal for NSF should be: find a way to make data description, integration, and archiving part of the scientific process. In proposals they should require, enforce, and fund a data management plan according to agreed upon standards." The group agreed on recommendations for future action and research that address the dominant challenges of cross-disciplinary cultures, interoperability across heterogeneous data, models and systems, and incentives. The primary targets for these recommendations fall in four areas:

Recommendation II. 1: Build capacity and models

Recommendation II. 2: Develop policy infrastructure to create a culture of sharing

Recommendation II. 3: Promote education

Recommendation II. 4: Promote research initiatives

These are briefly discussed in the following paragraphs.

Recommendation II. 1: Build capacity and models. This recommendation calls for supporting the development of *collaboratories* that model essential infrastructure for community-based, data-enabled research. Such collaboratories may be built upon existing resources or be newly conceived and could exist in a number of institutions and settings, incorporating 1) stand-alone institutions or organizations at local, regional, national, or international scales; 2) research and academic libraries, consortia, and special collections; and 3) data centers housed in universities, scientific and professional societies, and other heritage institutions. Such collaboratories could include projects that:

- Address small, medium, and large data collections
- Encourage collaborations among multiple stakeholders
- Instantiate data models, technical, and organizational architectures
- Reflect discipline-based or cross-disciplinary data collection, submission, and reuse
- Incorporate development of a sustainability plan

Specific research topics include: prototyping technical architectures under different organizational and collaborative arrangements; specifying ingest systems at different scales; deploying interoperable data models across organizations; developing tools for automatic metadata creation and for methods to "harvest" information about collections that might not be part of existing centers but might be of interest to investigators or potential candidates for future inclusion.

Recommendation II. 3: Develop policy infrastructure to create a culture of sharing. This recommendation calls for creating data management policies to ensure that the contribution of research data is considered a shared asset, enabling reuse in new research contexts as

shared public goods. This includes:

- Understanding how to motivate and incentivize researchers to contribute digital data to collaborative environments (deposit).
- Understanding the range of rights management, data confidentiality, security, and privacy concerns.
- Developing support structures and training programs that will encourage investigators to prepare archive-ready data and objects with appropriate metadata and formats.

Specific research topics include: surveys to understand patterns of use, deposit, and reuse of archival data across disciplines; work with journal editors, librarians, and others in the scholarly communication process to understand their requirements and how management of scientific data might be integrated into existing information flows; specifying the potential range of rights management, security, confidentiality, and privacy concerns among institutions, collections, and individuals, and ways for managing these concerns now and in the future.

Recommendation II. 3: Promote education. This recommendation calls for NSF among others to stimulate the development of expert data curators and an informed scientific community. Specific suggestions involved:

- Partnering with IMLS to support new programs in data science/curation.
- Supporting cultural change to encourage scientists and engineers to contribute to digital data collaboratories.
- Contributing to the development of curricula to create next generation scientists and information specialists.

Specific areas of research and funding might include: surveys of existing curricula to determine needs and support for model programs; and training programs, particularly those that might be aimed at graduate students and junior faculty, consistent with other programs for young researchers that NSF currently sponsors.

Recommendation II. 4: Promote research. Each of the previously described recommendations might support various research topics. In addition, the group identified a series of discrete problems related to infrastructure. These include creating and developing:

- Interoperability across institutions, collections, and heterogeneous data
- Reference tools to support specific kinds of collections and continued use (for example, gazetteers to map and monitor changes in geopolitical boundaries and terminologies)
- Architectural options to address risk management
- Specification of intellectual property and access rights
- Collaboration tools
- Security and trust
- Cross-disciplinary discovery
- Attributes of discipline-based repositories
- Appropriate standards
- Automatic generation of metadata

Endnotes

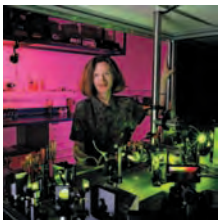
[1] Consultative Committee for Space Data Systems, Recommendation for Space Data System Standards, Reference Model for an Open Archival System (OAIS), CCSDS 650.0-B-1 Blue Book (January 2002). http://public.ccsds.org/publications/archive/650x0b1.pdf.

[2] Ibid., p. 2–2, especially Figure 2-1.

[3] Ibid., p. 4–1, especially Figure 4-1.

[4] National Library of New Zealand. Metadata Standards Framework—Preservation Metadata (Revised), June 2003, p. 4; http://www.natlib.govt.nz/files/4initiatives_metaschema_revised.pdf. On the mapping to the OCLC/RLG model, see Appendix 6 of this document.

[5] S. Choudhury et al., Digital Data Preservation in Astronomy, 2nd International Digital Curation Conference: Digital Data Curation in Practice, Glasgow, UK, November 21–22, 2006.



III. New Partnership Models

The importance of partnerships is the dominant critical theme in the participants' position papers. The charge to the breakout group asked the questions: What mix of individuals and organizations should be involved in data preservation? What creative partnerships can be developed between the multiple sectors? That is, who should be involved? What are their respective contributions, roles, and responsibilities?

Discussion of the Issues and Challenges

The breakout group outlined a series of challenges and issues primarily as they affected and were affected by research and academic libraries. Three major information processes: the scholarly communication process, the knowledge transfer process, and the life cycle of research formed a matrix within which the group considered the role of librarians, the requirements of preservation and curation facilities, and short- versus long-term needs. On the basis of this discussion, the group was able to characterize the relevant stakeholders and outline a series of considerations from which new partnerships might be modeled and eventually established.

Models, Processes, and Stakeholders. The three information processes, the scholarly communication process, the knowledge transfer process, and the life cycle of research, are shown in Figures III-1, III-2, and III-3. Figure III-1 shows how the traditional role of research

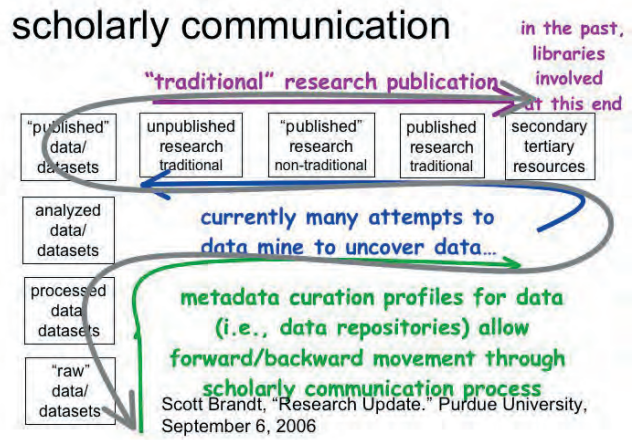


Figure III-1. Scholarly Communication
Source: Scott Brandt, Research Update, Purdue University, September 6, 2006

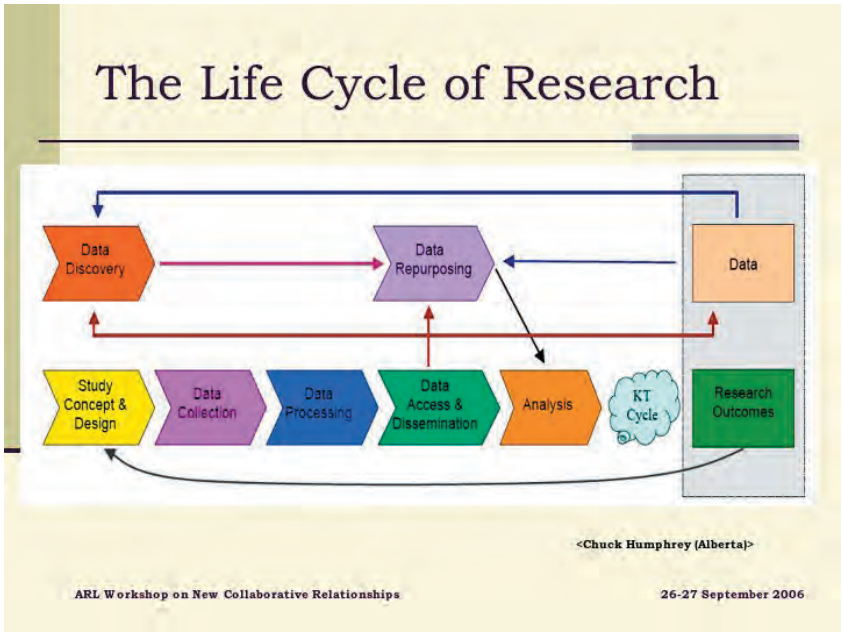


Figure III-2. The Life Cycle of Research
Source: Chuck Humphrey. The Role of Academic Libraries in the Digital Data Universe, September 26, 2006

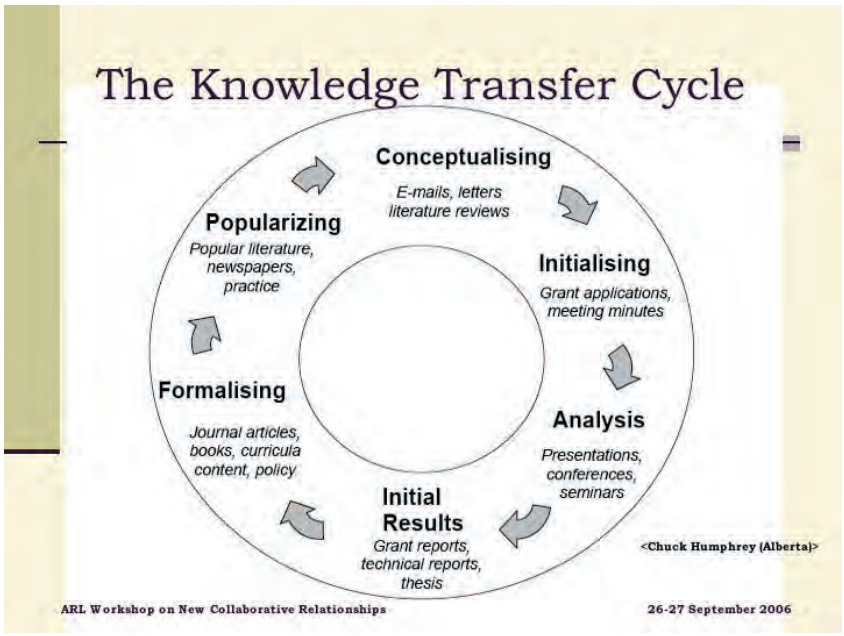


Figure III-3. The Knowledge Transfer Cycle
Source: Chuck Humphrey. The Role of Academic Libraries in the Digital Data Universe, September 26, 2006

and academic libraries is shifting from a focus on managing information in its published form to managing the digital data sets on which findings may be published. Figure III-2 illustrates the life cycle of research through which data are generated and Figure III-3 embeds the information in a social context, showing how information is used at different steps of the knowledge transfer process.

Some of these steps generate information that may be curated; in others, the information may be used and transformed. Each of these cycles has a series of participants and stakeholders on which partnerships may be based. Stakeholders include:

- Universities
- Libraries and librarians
- Domain specialists
- Computer scientists
- Standards-setting bodies
- Editors
- Professional societies
- Publishers
- Commercial and not-for-profit vendors
- Funding agencies

But as noted in many of the position papers, these stakeholder groups have different expertises, outlooks, assumptions, and motivations about the use of data, so that forging partnerships also requires transcending and reconciling cultural differences. For example, in one position paper the author commented, "Currently, the responsibility for management of data accompanying scientific publications falls to publishers in the form of supplemental data collections. Academic journals have little incentive to invest in the establishment and maintenance of digital data repositories that can be used for anything beyond minimal documentation of published reports."[1] Yet another author observed, "The traditional practice of gathering a paper trail of research outputs long after a scientific investigation has been concluded and depositing them with an archive is inapplicable in the digital era. Too much valuable research data are either at high risk of being lost or have been destroyed because of inappropriate practices that were carried over from a time when paper was the dominant medium. The challenge is to coordinate among partners the care of research data throughout the life cycle."[2]

Motivation and Heterogeneity. One of the most challenging issues is motivating individual investigators to deposit their data. Repeatedly, the position paper authors called attention to the need to motivate researchers. In part, this is a question of raising awareness of the future potential value of the data. But curation of individual investigator's data sets also requires resolution of the tension between the promotion and tenure system in which researchers understandably wish to protect their research asset in the form of their data set and the possible broader use of that data.

Moreover, as a practical matter individualized data sets can be highly idiosyncratic and difficult to integrate into larger collections, leading to the questions surrounding metadata discussed in the previous chapter and to issues of appraisal and management. That is, what is determined appropriate for inclusion in a collection and how are those decisions reached? (This is reflected in Figure III-2 by the arrow between data access and dissemination and data repurposing.) As one participant noted in the plenary discussions, "In conjunction with a study we did in Canada, we looked at the size of the problem in preserving research from humanities/social sciences. Between 505 and 595 out of every 1,000 funded projects resulted in the creation of data/databases. There's a lot of data being created. How significant is this?"

Distributed Systems and Hand-offs. The obvious solution is distributed systems in which

different disciplines and entities undertake responsibility for different "pieces" of the landscape. But working out the specifics is challenging and resources to support such facilities are required. If the universities undertake to support these activities, either by housing curatorial facilities on their campuses or by participating in consortia, there is concern that the less well-resourced institutions could be disadvantaged. Moreover, while it is agreed that there are multiple stakeholders and potential partners, difficulties may arise in determining respective responsibilities and where hand-offs occur. If the author of a paper is responsible for seeing to the disposition of the data, what is the role of the journal and the journal editor when an article is published based on part or all of the data? Should publishers take responsibility for maintaining complete runs of their journals and the data on which the articles are based? And how does this model compare with the morgue traditionally maintained by major newspapers, which comprises the "paper of record" and not the reporters' notes, polls, and opinion survey or other original data that contemporary newspapers frequently commission?

This problem of hand-offs (or more generally, of interfaces between complementary and potentially overlapping roles) is complicated by differing time scales. Current partnerships tend to focus on interoperability and integrated access but lack a long-term component, which is fundamental to partnerships whose mission is long-term data stewardship and to ensuring sustainability and hence confidence or trust in the system. So it becomes necessary to align short-term pressing needs with long-term goals. The time scale issue is compounded by tensions over open access, rights management, and policies governing protection of confidentiality, limitations on liability of data sources, use of data for humanitarian purposes, and definitions of appropriate use.[3]

The Role of Research and Academic Libraries. Piecing together compatible roles in new partnerships requires parties to think through their respective responsibilities and how these change in the digital environment. The National Virtual Observatory's new prototyping offers one example of how the research library can function in this environment, as specified in the OAIS model (see discussion in Chapter II). Librarians face both changing roles and changing perceptions of their roles relative to information (as implied in Figures III-1 and III-3). Traditionally, libraries have focused on information discovery, rather than information management, and, as one position paper author argued, the reference function continues to be an important function in the digital age. In addition, libraries have tremendous credibility on campus as shown in a survey of information use at four-year liberal arts colleges and Ph.D. granting public and private universities that was funded by the Andrew W. Mellon Foundation. Ninety-eight percent of the faculty, graduate students, and undergraduates included in this study agreed with the statement, "My institution's library contains information from credible and known sources."[4]

That said, the group found that research and academic libraries need to expand their portfolios to include activities related to storage, preservation, and curation of digital scientific and engineering data. This requires evaluating where in the research process chain (Figure III-2) curation and preservation activities (Figure II-2) should take place, what capacities should be built to support these activities (see Infrastructure), and who is best suited to undertake these activities. This begs three related questions: Where do partnerships come into play? Where are the hand-offs? How do we lower the barriers to participation? Experience suggests it is difficult for research and academic libraries to have the expertise to curate in every do-

main area. It must be a shared responsibility. Librarians can identify where skills lie across domains and to coordinate them so faculty can do discovery. Universities have played a leadership role in the advancement of knowledge and their libraries have shouldered substantial responsibility for the long-term preservation of knowledge. An expanded role in digital data stewardship for some of these universities and libraries, along with other partners, is a topic for critical debate and affirmation.

"It takes a research community to preserve its data."

The group agreed that given the scale of the challenge, the responsibilities should be distributed across multiple entities and partnerships that engaged their respective communities. During the plenary discussion, one participant noted that some people managing discipline-based archives are worried about institutions or library programs taking on this responsibility. So, while the specific responsibilities might be distributed, it would also require a tight partnership with the library or institution and the discipline. "As a researcher I'd love to have help of someone with expertise" to assist with curation and management. Others in the plenary discussion suggested looking to organizational models offered by museums and archives where the notion of the community served by the institution is not bounded by a single institution in the way that research and academic libraries are situated within universities and colleges.

In any case, whether a museum or research or academic library, it is important to build the necessary expertise and equally important to build an appreciation of the need for this expertise.

Outreach and Education. The group outlined a series of steps that would be required to address the challenges of building new partnerships. The first is to raise awareness and meet the needs in the research community. Outreach to the scientific community to raise awareness of the need for long-term data stewardship and to encourage researchers to reuse data was a theme that appeared in several of the position papers as well as in this session. Wrote one participant, "Many scientists continue to use traditional approaches to data, i.e., developing custom data sets for their own use with little attention to long-term reuse, dissemination, and curation. Although there has been considerable progress in data stewardship for 'big science' projects, even modest collaborative projects are inconsistent in their attention to data management and few individual scientists think beyond posting selected results and data on the Internet or submitting a final data product to a data archive if required to do so. Changing this sort of behavior will require a range of efforts, …perhaps most important of all, concerted efforts to educate current and future scientists to adopt better practices."[5] Librarians and "data scientists" also require new training. "We must ensure that the talent to preserve scientific data will be available," wrote one participant. "The preferred approach is to provide incentives for computer science and library science departments to include suitable disciplines in their curricula."[6]

Research: Metrics, Motivations, and Technical Requirements. Workshop participants called for more formal metrics and studies similar to market research. For example, it was suggested that when proposals are written to NSF, a data management plan could be written into the budget as part of the budget justification. When the final report is submitted, it might include information on the execution of the plan. Then, the Foundation might gather the data, as it does on scientific and technology and engineering indicators, and use this information to understand the patterns as well as to underline

the importance of the data management exercise. In short, this idea has a double resonance: it highlights data management as a metric and calls attention to its importance, precisely because it is a metric. By extension, such metrics provide incentives to research both as requirements and as a potential source of recognition.

Clearly, the viability of partnerships among institutions is based on people to staff them and the willingness of investigators to take advantage of them. Systematic information is lacking about the willingness of researchers outside of the fields, like astronomy or genomics, where the massive data sets are central to the research. Consistently, the position paper authors cited reluctance among many researchers to engage in partnerships to curate digital data. But at least one study of this sort is in progress at ICPSR. Statistical Disclosure Control: Best Practices and Tools for the Social Sciences, led by JoAnne McFarland O'Rourke and Myron Gutmann and funded by the National Institute of Child Health and Human Development (NICHD), constructed a national sample of National Institutes of Health (NIH) and NSF researchers who collected original social or behavioral between 1998 and 2001.[7] In addition to issues specifically about disclosure, the survey covered related topics, including data sharing practices, concerns about risks to data subjects' privacy in files that are shared, and procedures and resources for preparing files to be shared. It was fielded in April—July of 2006 and the results are not yet analyzed. However, it is expected that the results will help establish a climate for data sharing by developing an instrument that others can use and documenting some of the barriers investigators face, the extent to which they do share data, the ways in which they share, and the training in which they would be interested. One of the overall aims of the project is to develop tools to assist researchers in evaluating their data for risks of disclosure and provide guideline to help them determine appropriate measures to protect against those risks. Both the instrument, which will eventually be made available for broader use, and the results are critical to understanding the environment within which viable institutional partnerships might be built.

Having recognized the interest and the need for curatorial facilities by educating the research community, the next step is to understand and define the requirements for repositories (for example, granularity, metadata, rights management, and so on); many of these issues were addressed by the Infrastructure Group (see Chapter II). The management structure of the institutions in which these repositories might be located should distribute stewardship responsibilities across a body of responsible parties roughly equivalent in magnitude (i.e., size, capacity) to the magnitude of the collective data in need of stewardship and reflective of the disciplines that create data in the data repository and that might use that data. The management structure should also ensure that the work environments of those responsible parties are well supplied with curatorial tools that facilitate their carrying out their responsibilities. Clearly, a range of structures might be possible (a point the Infrastructure group also made), and a period of prototyping and testing would be required before new capacities were deployed. Key elements of any institution would be measurement and an appropriate business model (see Chapter IV).

Recommendations of the Breakout Group

As the last paragraph suggests, the structure of any partnership requires attention to the infrastructure, both the infrastructure of a given facility as well as the infrastructure more broadly that would enable cooperation across individual institutions. Indeed, the library and archives professions have been leaders in the development of organizational relationships that enable libraries and archives to function coherent-

ly across a broad range of specific institutional instantiations. That said, the New Partnerships group made one major recommendation and five narrower recommendations that built out aspects of the principal recommendation.

Overarching Recommendation

NSF should facilitate the establishment of a sustainable institutional framework for long-term data stewardship involving the players previously enumerated (e.g., libraries, universities, professional societies, scholarly journals/publishers, disciplines, etc.). This was the major, overarching recommendation put forth by this group. The framework must encourage the articulation of what constitutes "curation" in various disciplines, encourage a diversity of designs and approaches that are sympathetic to the needs, practices, and relationships within affected research communities, and encourage the development of distributed partnerships between research libraries and research institutions. As part of this exercise, it should be expected that there is a balance between prototypes and long-term commitments as well as the capacity to evaluate the success of different approaches and to migrate data sets collected from experimental to long-term facilities as different models evolve.

Moreover, these approaches, it was pointed out in the plenary session, should consider the model of other heritage institutions (archives, museums, scholarly societies) where communities of interest transcend university boundaries and merge across several disciplines. These groups have grappled with many of the issues that drive concerns associated with management of digital collections. For example, with respect to confidentiality and individual privacy, archivists are used to taking in items with use restrictions. That concept allows archivists to understand how to manage data sets, for example with researchers who want information embargoed for a certain period of time.

Pursuant to this overarching recommendation, the breakout group articulated five more recommendations that executed specific aspects of this broad idea. Namely,

> Recommendation III. 1: NSF should fund pilot projects/case studies that demonstrate the intersections between libraries, a limited number of scientific/research domains, and extant technologies bases.
>
> Recommendation III. 2: NSF should fund projects in which research libraries develop deep archives of irreplaceable data, assuring descriptions of these data at a minimal level (floor, not ceiling), and facilitating discovery and access to these data, according to prevailing community standards.
>
> Recommendation III. 3: NSF should require that data management plans submitted as part of the application process identify the players involved in the custodial care of data for the whole of its life cycle and should support training initiatives to ensure that the research community can fulfill this requirement. This would include promoting new curricula, developing new programs, and linking the training of domain scientists with library/information scientists.
>
> Recommendation III. 4: NSF should foster the training and development of a new workforce in data science.
>
> Recommendation III. 5: NSF should partner with IMLS to train information and library professionals (extant and future) to work more credibly and knowledgeably on data curation as members of research teams.

The plenary session generally concurred with these recommendations with two caveats: the

structure of incentives and the possible need for programs to reduce the potential disadvantage to smaller, less well-resourced institutions. Both of these topics are discussed in more detail in Chapter V, where summary conclusions and recommendations are presented.

Endnotes

[1] Todd Vision, Position Paper.

[2] C. Humphreys, Position Paper.

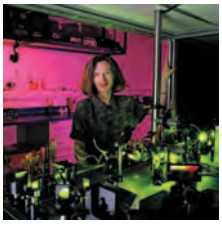
[3] R. Chen et al, Position Paper.

[4] A. Friedlander, Introduction, Dimensions and Use of the Scholarly Information Environment (Washington, DC: Council on Library and Information Resources, November 2002), p. 6.

[5] R. Chen et al., Position Paper.

[6] E. Van de Velde, Position Paper.

[7] Human Subject Protection and Disclosure Risk Analysis, http://www.icpsr.umich.edu/HSP/.



IV. Economic Sustainability

An economically viable system for data stewardship requires prevention rather than rescue of endangered materials. As one participant noted,[1] what is needed is a system of diagnoses of similar problems, treatment protocols and good practices, and criteria for making decisions. The charge to the Economic Sustainability group was: What models are required to sustain data management and preservation efforts over the long term? This broad question entailed consideration of many sub-questions and assumptions, among them: what does it cost to maintain, preserve, and manage collections including access to those collections? Who uses the collections? And who should support them?

Discussion of the Issues

This group covered a broad range of issues, including funding strategies, business models, the concept of public goods, and motivations. It proved to be a complex topic that embraced a broad range of topics and concerns with a particular focus on sustainability issues.

Overview and Framework. Rusbridge, one of the plenary speakers and a co-chair of the breakout group, outlined the range of questions that converge under the rubric of economic sustainability (Figure IV-1).

These topics fall roughly into seven broad categories:

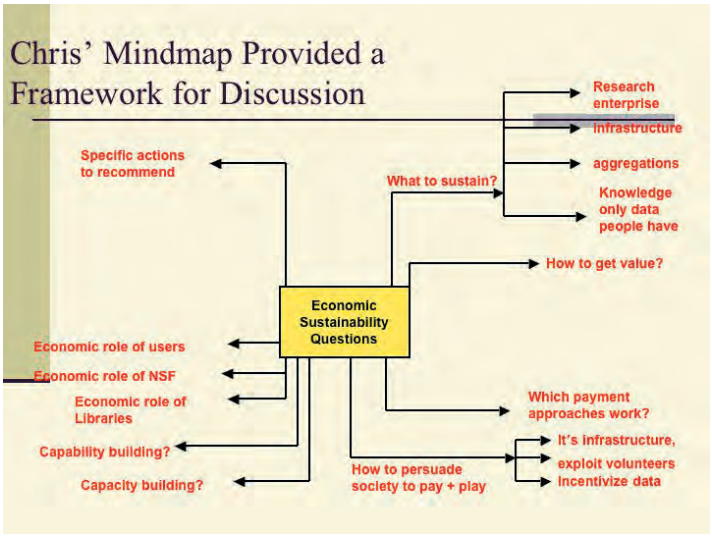


Figure IV-1. Economic Sustainability Questions
Source: Chris Rusbridge, September 26, 2006

1. What to sustain? The obvious answer is the data sets. Yet the "what" question also subsumes other forms of information and knowledge, a point also raised in the New Partnership Models breakout group in the context of the Knowledge Transfer Cycle (See Figure III-3). In addition, the model makes the important point that the economic sustainability of data stewardship also sustains the research enterprise itself.

2. How to get value? This question has several dimensions: At the level of data, it can mean the transformations required to make the data meaningful to researchers. It can also mean modeling the trade-offs required to determine priorities and decisions. Not all data need necessarily be afforded the same level of processing. Some data may be unique and therefore irreplaceable, but in other contexts, it may be more cost effective to replicate the experiment than to preserve the data. (The value can also accrue in other ways; the data set may be a source of academic credit in itself.)

3. What payment approaches work? The library community has been heralded for its ability to maintain large scale organizational coherence while maintaining robust heterogeneity in funding and business models at the institutional and local levels. Curating data requires a similar coherent yet diverse set of approaches that recognize cultural factors (some disciplines have been historically reluctant to pay for data access and use) as well as legacy arrangements (the university library is frequently funded centrally by the university and not by contributions from individual departments, and thus the cost of maintaining a relatively specialized data set for a given discipline would presumably come from the library's budget and not from the department or the investigators who created and might use the data.) Other models discussed by the group included:

   - Inter-University Consortium for Political and Social Research (ICPSR), which employs several strategies (subscription, user fees, federal, private funding) and is discussed in greater detail later in this chapter.

   - The Mormon Church, which combines tithing, user fees, and sales.

   - The Public Broadcasting Service, which assembles private donations, public funding at the federal and state levels for operations and for specific projects, volunteers, who contribute time and expertise, and sales.

   - Volunteer activity, which might take various forms, including perhaps archiving@home similar to SETI@home, or use of distributed commodity resources as in LOCKSS.

   - Markets, with a number of examples suggested, including DRI, data "futures," shares, and so on.

   - Hybrid funding, which comprises a mix of public and private funding from multiple sources. Public funding, it was noted, can take the form of contracts of several years that are re-competed so that the administration of a facility may be separated from the facility itself. Conceptually, the commitment to the data is separated from the commitment to the particular service organization.

The University of California, for example, has operated the Los Alamos Laboratory

for decades in a series of contracts and the NSF has similar arrangements for operation of its telescopes. This enables the threshold costs of construction to be separated from the operating costs and enables the government to revisit its share of operating costs and to obtain more efficient pricing for these services. With respect to the internal operations, this also means that the current holder of the contract is incentivized to practice reasonable efficiencies as well as to maintain the operations in a way that enables smooth transitions.

Several variations might be imagined. For example, a consortium that holds a multiyear contract to operate a regional data center might agree to hand off the actual operation each year to a different member of the team, an arrangement that provides for redundancy in expertise and enables transfer of experiential knowledge. Moreover, if the structure of the consortium required each member of the consortium to provide a percentage of matching funds, then shared matching costs might be equally distributed within the team by utilizing their respective capacities. This approach, however, presents many logistical and practical challenges.

4. **How to persuade society to pay?** This question is related to the preceding topic and resonates with the themes of outreach, communication, and education articulated by the New Partnerships breakout group. If the data sets are seen as an element of the infrastructure, then presumably it would be easier to obtain funding for their maintenance, either through public funding as a public good, which is discussed subsequently, or through user fees or other forms of payment by the research community. Like others, the Economic Sustainability breakout group saw incentivizing scientists as a critical element and also called for engaging volunteers. Certainly in some disciplines, notably archaeology and astronomy, volunteers have provided critical resources and the fields have benefited from the associated positive public profile, broad support, and direct contributions of time and effort.

5. **What capacities are required?** This question goes to some of the issues discussed in the Infrastructure group (See Chapter II) and are necessarily inputs into both economic and business models. It is important to note, though, that like the Infrastructure breakout group, the Economic Sustainability group, understood the notion of capacity as encompassing the human resources required to manage a data curation/stewardship center as well as the physical and software facilities.

6. **What capabilities are required?** This question addresses the services that a curation/stewardship facility might require to manage the internal processes as well as the external relationships with users, administrators (if the curation facility is housed within a larger institution), sponsors, and so on. These capabilities might be provided by humans but might also be automated and hence reliant on technology.

7. **Economic roles.** This topic embraces a number of issues relating to the economic roles of the users, of the investigators who deposit collections, of the NSF and other sources of public funding, and of the curation/stewardship facility itself. Ultimately, the facility represents an intellectual asset with economic value, although assessing this value is very challenging. Indeed, the economic value of intangible assets is an active area of

contemporary research and one that might become useful in future studies.[2]

As several of the group participants pointed out, there is an important distinction to be drawn between the economic model, which addresses questions that arise from the economic system, and the business model, which looks at how an individual institution remains solvent. The former looks at societal, macro, and microeconomic issues; the latter is highly context dependent. Thus, the economic models consider the role of the government based on its aggregate behavior and funding trends; the business model for a research or academic library might develop a plan based on its endowment, fund raising, alumni support, parent budget, and needs of its faculty, students, and research programs in which direct public funding from either state or local sources might be a relatively small component. Both perspectives are important, but they operate at different scales.

Public Goods. From the economic perspective, data and the curation facilities in which they are housed are typically viewed as a public good. A public good has been defined as "a good that must be provided in the same amount to all affected consumers."[3] In technical terms, this means that the good is both *nonrival*, meaning that one person's consumption of the good does not diminish the amount available to another consumer, and *nonexcludable*, meaning that one cannot exclude another from consuming it.[4] The classic examples are sidewalks, clean air, clean water, and national defense. However, Varian (among others) has argued that there is an important distinction between the two properties. Whereas nonrivalrousness inheres in the good itself, the nonexcludable property is socially constructed. Thus, providing clean water to all makes sense from a public health perspective but is, nonetheless, a decision. Information shares these public good properties. It is nonrival, in the sense that the cost of reproduction is very low (although the cost of initial production can be quite high) and it has been excludable as a result of intellectual property regimes.[5]

From the perspective of the data center, intellectual property rights management is less a source of income than a cost of operation. However, the notion of a public good highlights some of the challenges of building a business model. What constitutes the community of consumers? Moreover, public goods are vulnerable to the free rider problem, namely, when individuals allow others to provide the public goods that they then consume. This is a problem that is inherent in archives, where the consumer of the data 5, 10, or 100 years hence by definition will not have contributed to its production. This has been characterized by one participant as "public goods with an unknown future value."[6] There exists a substantial body of economic literature on the problems of information goods, public goods, and the associated problems, and the group believed that it would be useful to undertake systematic investigation of this literature and its relevance to modeling the economic sustainability of long-term stewardship of data.

Models and Examples. In her plenary remarks, Berman sketched out one model that linked the type of collection to an organizational structure and funding source (see Figure I-2). Small, local collections might be supported privately, perhaps by commercial entities offering "warehousing" space. Regional collections might be supported by regional libraries and data centers, and large national/international reference collections would be supported by national governments. In practice, though, data centers could employ a range of funding strategies, including relay funding, where a succession of grants is linked together; user fees, applied to both depositor and user; endowments; and

membership dues. At ICPSR, which is housed at the University of Michigan, approximately 25 percent of the annual budget comes from memberships in the consortium, 10 percent from fees from the summer school program, 33 to 40 percent from long-term projects to provide resources to the community, and the remainder from "project-based" funding, which is used sparingly to support infrastructure, as well as from federal agencies and private funders.

ICPSR is widely acknowledged as a success story, both in terms of its acceptance by the relevant communities and as a business venture. Some of the reasons for its success include:

- It provides a robust environment with low barriers to access. ICPSR accepts data in many formats subject to a well-defined and well-articulated set of criteria.

- It houses content which is of great value to its user communities and maintains that content in good form. Thus, ICPSR refreshes the media and migrates formats as necessary and appropriate.

- It has a business model and structure which reflect the culture of the domain and constituent users.

- It provides useful tools associated with data.

- It maintains a trusted repository. Users can be confident that the data they deposit will be maintained in good order, including appropriate access controls and hence appropriate levels of confidentiality, and they can also be confident that data they obtain from the repository will be authentic. That is, the data have not been compromised or tampered with as a result of storage or transmission within the repository. (Whether the data are *accurate* is a separate issue and not one with which the data center need be concerned.)

Costs: Facilities and Human Resources. A business model recognizes not only revenues but also costs, that is, the cost of managing the facility: staffing, resources, and so on. Less easy to quantify, but potentially far more costly, is the cost of labor, particularly the skilled labor associated with management and preservation. These activities include collection policies (including appraisal, weeding, deaccessioning, and so on); data clean-up, normalization, description, and preparation for submission (see Chapter II. Infrastructure for a discussion of the distinction between the submitted and archived form of the data in the OAIS framework); and collaboration with researchers around scholarly communication, best practices, and related topics.

Appraisal. Appraisal is a particularly challenging topic and bears on the mission of the data center (and hence on the question of partnerships) as well as on the underlying capacities (and hence on the question of infrastructure). As a practical matter, archivists and librarians are well aware that data is more abundant than time to process it. Indeed, the National Research Council's report on the Electronic Records Archive at the National Archives and Records Administration (NARA) predicted an "avalanche of digital materials" resulting from advances in computer technology, more finely-grained information and transactions, dynamic generation of information, communication technology that enables more frequent transactions and that generate preservable records (for example, text messages, e-mail messages, and so on), and the drop in the cost of storage.[7] The report called for multiple approaches to preservation, recognizing that technologies will continue to evolve and that different records require different treatment. From the perspective of evalu-

ating collections, workshop participants noted that small projects were at greater risk than large, since large collections tend to call attention to the data by virtue of their size and scale. The Human Genome Project, the Protein Data Bank, the National Virtual Observatory are all examples of large, valuable, well-recognized, and well-curated data collections. Moreover, because of their scale and innate complexity they have multiple sponsors and an inherent redundancy. (The storage structure of the National Virtual Observatory was discussed in Chapter II.)

Nevertheless, selection is necessary and selection policies require the skilled services of both information and domain professionals, who must address several major questions:

> What is the minimal level of processing required to preserve the collection in a way that it can be subject to more detailed and sophisticated processing at some future time?
>
> How are decisions made about which collections should be processed fully (within the constraints of existing technologies) and which should be set aside with only minimal processing sufficient to ensure stability and prevent degradation?
>
> Can we develop an understanding of how data might be re-purposed in ways totally unpredicted by its original creator/gatherer/researcher?
>
> What legal/ethical constraints encourage or discourage the long-term preservation of particular data sets?

Value Proposition. Although these questions are phrased in terms familiar to librarians and archivists who are responsible for collection development, these are also business questions that can be understood as a "value proposition." Value can be a difficult concept. A collection may possess value because it is rare and unique. Value can also be assessed as the cost to replace it, enabling trade-offs between the costs of preservation versus the cost of reconstituting the data. Finally, value can be understood as the value to the researchers, a measure of how much a future researcher may want the data and the ability it may afford to ask different kinds of questions precisely because longitudinal data may exist. These different notions of value all have economic and business consequences that are not fully understood. As put by one of the participants in her position paper, "If we take as a given that not all data are created equal and that we will not be able to afford to keep everything, how do we decide where to invest in preserving data? How do we make effective economic decisions in the face of uncertainty about the supply of data and the future demand? Are there any economic models or research issues that provide insights into comparable problems? What happens when the future value of a particular set of data is contingent upon its relationship to other data that have been preserved, and can therefore be aggregated? At what level of granularity do we make selection (e.g., investment) decisions, given that deciding what to preserve is a very labor-intensive and expensive process."[8]

Demand. Running through these discussions, but not explicitly stated, is the problem of demand. There is a tendency to look at infrastructure primarily from the producer side. Indeed, in the formal definition of a public good, demand is assumed to be roughly uniform; problems and disparities arise when it is not, as evidenced by the free rider problem. In practice, however, notions of value as well as decisions about collections, partnerships, the supporting infrastructure, and hence the economic and business models all eventually face the

question of demand. Who will want to use the collections? Who will be willing to pay even a nominal user fee? And how does demand affect value? To some degree, these questions are inherently unknowable. However, it is possible to parse knowable aspects through the kind of market research and programs of public education and awareness that the partnership group discussed.

Recommendations of the Breakout Group

The group formulated eight recommendations. Many resonate with recommendations made by the other groups. They are discussed in the following paragraphs.

> Recommendation IV. 1: Involve economics and social science experts in developing economic models for sustainable data preservation; this research should ultimately generate models which could be tested in practice.

Economists have begun to make progress in addressing various contributing elements of the data curation/data preservation problem: better understanding of information goods and markets, the economics of public goods and their relationship to market mechanisms; and the valuation of intangibles, including information assets. These developments, if systematically examined from the curation/preservation point of view, could be brought usefully to bear in formulating testable models. A second set of questions arises from the demand side, understanding motivations of the various stakeholders. A third set of questions examines the spectrum of potential business models that take into account the scope of collections (local, regional, national, and international); the range of organizational arrangements, as set forth by the New Partnerships group; and the infrastructure requirements, as outlined by the Infrastructure group.

> Recommendation IV. 2: Set up multiple repositories and treat them as experiments.

It is evident that multiple strategies will be required to meet the circumstances created by different disciplines, collections, and partnerships. Repository experiments should be required to develop plans to address key issues such as transition between media/formats/institutions, self-sustainability, exit strategy, etc.

> Recommendation IV. 3: Develop usable and useful tools for automated services and standards which make it easier to understand and manipulate data.

Automated tools to optimize the labor inputs of the professionals who staff curation facilities, as well as to make the researchers more effective, are critical. Such tools not only render the operations of the data stewardship facilities potentially more efficient but they also lower the barrier to participation by the researchers themselves and can potentially lead to better science. This has the positive effect of building momentum behind stewardship but also creates demand for the data stewardship facilities' services as both a depository and a source of data. Both can potentially generate income for the facility by becoming sources of user fees. Many communities are currently unfamiliar with the notion of paying to use data, but others are not, as the example of ICPSR illustrates. Thus, an education program coupled with successful stewardship facilities that meet the pent-up demand can create a so-called "virtuous circle" to support preservation, curation, and stewardship in the concerned disciplines.

> Recommendation IV. 4: Require a data sharing plan in proposals that has practical value (and appropriate support). Plans for resource and reference data should contribute to community data stewardship.

Augmenting the NSF proposal process prompted vigorous discussion during the closing plenary and will be discussed in greater detail in Chapter V. The basic idea, as presented in this breakout group report, was to include a section analogous to sections in the existing budget justification in which the proposer explains certain elements of the budget request. The rationale for including the provision for data is that it calls attention to the data collected as part of the project, raises its visibility for both the proposer and the reviewer, and begins to attack the problem of what data should be preserved, by whom, and for how long. Thus, an investigator may very well argue that data collected as part of the project does not merit long term preservation, but if it does, then the justification begins to contribute incrementally to broad criteria for preservation within a discipline and also creates a demand for the curatorial facility itself. Such requirements and facilities, for example, have long been a feature of archaeological investigations. Archaeological projects, whether funded as a research project or as part of the national environmental review process, are required to cull the collection recovered from the excavation and deposit it at a certified repository, typically maintained by the state according to agreed-upon federal guidelines.

> Recommendation IV. 5: Create and enforce data sharing policies among NSF awardees (e.g., the final report may not be accepted unless the awardee is compliant with the stated data management plan.)

Like the preceding recommendation, this proposal resulted in active discussion during the plenary session about burdens potentially placed on investigators. The intent, though, is to raise the profile of digital data stewardship among domain scientists and funding agencies, create resources for future users, and meet the demand and need for these services.

> Recommendation IV. 6: Use the NSF program process to help the research and academic library community take more responsibility for the stewardship of scientific and engineering research data (potentially with other funders).

Encouraging investigators to deposit their data sets in curatorial facilities is meaningful if such facilities exist. Although maintaining such digital depositories has not historically been part of the research and academic libraries' portfolio of responsibilities, some have argued, including those who participated in the New Partnerships group, that it is a reasonable extension of those libraries' mission. These libraries generally enjoy reputations as trusted facilities on campus and elsewhere, and funding and support from NSF and other agencies would assist them to undertake this responsibility by enhancing their capacities in specific instances and thus enhancing their credibility in this realm.

> Recommendation IV. 7: Use the NSF program process to facilitate cultural change in the research community.

This recommendation expands on the more specific recommendations above, which tied data management requirements to the proposal and reporting processes. More generally, it was intended to encourage NSF to think about a variety of strategies in which the importance of preservation/curation/stewardship might be made more visible. For example and as discussed by the New Partnerships group, contributions to data curation facilities might be included in the various reports of science and technology indicators, and credit might be given to creators of data sets through citation in NSF-funded reports, while institutions recognize this value through promotion and tenure reviews.

Recommendation IV. 8: Undertake capacity and capability building activities.

This recommendation covers a range of specialized activities, including software tools to support automatic or semi-automatic metadata creation, reference tools and ontologies, monitoring, and so on. It also addresses the need for training for domain specialists, data scientists, librarians, and archivists, compilation of best practices, and consideration of appropriate data management policies and procedures.

Endnotes

[1] M. Hedstrom, position paper.

[2] See, for example, Feng Gu and Baruch Lev, Intangible Assets; Measurements, Drivers, Usefulness; April 2001. http://pages.stern.nyu.edu/~blev/intangible-assets.doc.

[3] Varian, Intermediate Microeconomics (4th edition; New York: W. W. Norton & Company, Inc., 1996), p. 606.

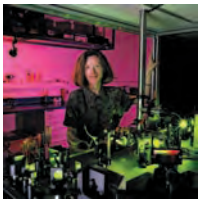
[4] Varian, Markets for Information Goods, April 1998; revised October 16, 2998; p. 6. http://www.ischool.berkeley.edu/~hal/Papers/japan/japan.pdf.

[5] Ibid. p. 7.

[6] M. Hedstrom, Position paper.

[7] National Research Council, Building an Electronic Records Archive at the National Archives and Records Administration (Washington, D.C.: National Academies Press, 2005), pp. 19–20.

[8] M. Hedstrom, Position paper.



V. Summary, Conclusions, and Recommendations

The stewardship of digital data is fundamental to the research enterprise. The data's utility ranges from the ability to replicate an experiment, to the efficiencies associated with reusing data, to the potential to ask new kinds of questions through new capabilities to integrate and manipulate data. So the challenge becomes, how do we as a society collectively gather, store, manage and make these data sets available while respecting a multitude of legitimate and sometimes competing interests? As the system of higher education itself suggests, an ecology of organizations will be required. But the organizational ecology that currently characterizes higher education (including heritage and cultural institutions with which they are affiliated) evolved over the course of decades, built on centuries of tradition and practice. And the research enterprise is also dynamic as new institutions and disciplines enter the system reflecting the ongoing needs of the population for new fields of study, continuing education, and new arrangements.

Digital information is fragile and we do not have the luxury of letting time take its course. The self-organizing archives already identified in this report demonstrate that investigators will establish and utilize these types of collections but the patchwork of archives and data centers, many of them supported entirely by volunteer efforts, is not sufficient in many disciplines to sustain the data over decades with enough confidence to inspire widespread use. For any system or set of systems to be successful, the users must trust it. The prestige of the NSF is a good start but not sufficient in and of itself. Potential users must see the value in the system and the system must function; it must be able to provide services reliably. Within the framework of a national system of data centers called for by the NSF, workshop participants outlined the challenge, asking a very simple question: "What does it take to get there?"

Summary of Plenary Discussions: Bridging Cultures and Creating Incentives

The discussions and recommendations set forth by the breakout groups converged albeit from their respective perspectives. Nearly all of the groups called for a system of data curation and/or stewardship facilities reflecting new kinds of partnerships. These should be approached experimentally at first through prototypes and temporary arrangements to study successful models and abstract lessons learned and best practices. There was also wide consensus for

- cross-disciplinary research and the technical capabilities and tools to support long-term preservation of the data;
- appropriate access monitoring and access controls to protect confidentiality and personal privacy as well as system security and risk mitigation more generally;
- automatic creation of metadata; and
- interoperability over heterogeneous data and systems.

In addition, the need for current and longer term research that would bring to bear the expertise of allied disciplines was identified. For example, this could include a focus on business and economics issues relating to some of the problems raised by the establishment of these data facilities. Thus, any facility, operational or prototyped, should be sufficiently flexible to accommodate changes in technology and organization, including hand-offs to successive operators in a relay structure.

Workshop participants also agreed that education and outreach to scientists, librarians, and the public on the topic of data stewardship was vital. This would include curricula to train a new kind of information professional, building on the traditions of librarians and archivists, as well as strategies to educate scientists on the value of digital curation and the possibilities for research that such data may offer in those domains where reuse of data is not common.

Finally, all of the groups acknowledged the challenge of communicating and working across disciplinary and institutional cultures and the burden that placed on developing appropriate incentives for individuals and management policies for institutions and collections. This theme echoed through the position papers and plenary sessions as well as in the group discussions. The way forward is inevitably through a mix of cross-institutional, cross-disciplinary structures that can take multiple forms and can range from the national and international organizations such as the Protein Data Bank and similar data sets to smaller regional centers that may be co-located with existing centers to specialized collections housed on individual campuses or within existing museums and libraries that serve a well-defined community. One such example is the Field Museum with a specialty in anthropology. Incentivizing participation is key and is complementary to the call for data stewardship facilities, since the existence of the facility provides opportunities for scientists to deposit their collections as well as experiment with the type of data such facilities offer. At the same time, such participation justifies the facilities and creates demands for curatorial services.

Two dimensions of the discussions about cross cultural challenges and incentives stand out: one concerns the culture within the NSF, and the second the culture in the research enterprise as it affects individual researchers and mechanisms for motivating their participation in digital data stewardship. In summary comments, it was noted that digital data stewardship challenges the NSF culture as a basic research agency. Specifically, projects that represent applied research, which may be highly relevant to the research required to build prototype data curation and/or stewardship facilities, can be difficult to fund through the standard review process, which reflects the values associated with basic research. It will be important for the NSF to address this concern.

From the academics' perspective, systems of prestige, embodied in promotion and tenure review, may act as disincentives to participation in long-term curation by tacitly encouraging investigators to retain control of their data and by discouraging them from allocating their time, a very scarce resource, to even minimal processing of the data. While the scarcity of time together with the challenges of extensive metadata argue for tools to assist researchers as well as for automatic metadata creation, NSF could also facilitate change by coming up with ways to recognize creation of the data sets and to ascribe authorship, in some manner. For example, deposit into a certified institution might count toward qualifying publications in a proposal submission or data curation might become one of the science and technology indicators that NSF compiles. Thus, data management and curation at the individual level become effectively congruent with existing prestige systems.

A number of ideas were identified to advance the NSF process of funding and reporting that could provide incentives to researchers, a stick as well as a carrot. Several draft recommendations called for attaching a requirement for a data management plan to the budget justification, which is an element of NSF submissions, and to the reporting required at the conclusion. The latter had the additional advantage of providing a corpus of information that NSF might use to understand patterns of data creation, deposit, and reuse.

The concept of data management plans is not new to NSF. The National Science Board report, "*Long-Lived Data Collections: Enabling Research and Education in the 21st Century*," stated, "individual or teams of researchers who will author and curate the data…need to have a strategy for dealing with data from their inception to their demise." The NSB recommendation provides further detail and guidance to the current NSF requirements as presented in the Grant Proposal Guide (NSF-04-2). While participants agreed with the strategy requiring a data management plan, a few concerns were voiced. Some participants suggested that requiring proposers to submit a data management plan, however brief, was an additional burden on the investigators' already stretched resources. An opposing viewpoint was that developing and elucidating a plan for valuable research data is part of responsible scholarship, but without the mandate that a data management plan be included in the proposal submission (in the same way that a budget justification or previous work is included), the culture around stewardship and preservation would be less likely to change.

Recommendations

The recommendations of each breakout group have been presented in the context of descriptions of the discussions that took place within those groups. Those recommendations have been refined based on plenary discussions and by eliminating redundancy. This process resulted in one over-arching recommendation that reflects the consensus of the Workshop participants. In addition, three general recommendations emerged from the group discussions that amplify the overarching recommendation. Finally, six targeted recommendations build on the more general recommendations.

Overarching Recommendation

NSF should facilitate the establishment of a sustainable framework for long-term stewardship of digital 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 digital 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.*

This overarching recommendation recognizes the simultaneous and mutually dependent need for both capacity (facilities and resources) and motivation, for supply and demand. It also recognizes that the capacities do not yet fully exist (although there are many examples of approaches within many disciplines) and require

additional prototyping and research into relevant technical, organizational, behavioral as well as economic and business issues. It was recognized that substantial effort in outreach and education is required to create an environment and mindset conducive to curation, preservation, and in some cases, stewardship of digital data. These efforts must be taken within NSF and other agencies, as well as within the cultures of the respective disciplines and organizations, including professional societies; libraries, archives and other heritage institutions; publishers, and universities. Finally, just as the stewardship of the data requires cross-disciplinary collaborations, so too must the responsibilities for effecting these goals transcend the mandate of the NSF. Hence this recommendation includes an interagency element not unlike the Digital Libraries Initiative, which assembled resources from multiple agencies in an integrated research program that pursued shared goals.

Three general recommendations emerged from the three groups along the following themes:

1. ***Fund projects that address issues concerning ingest, archiving, and reuse of digital 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, and evolving information technologies and institutions.

2. ***Foster the training and development of a new workforce in data science***. This could include supporting 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 digital data;*
    - *digital data registration;*
    - *reference tools to accommodate ongoing evolution 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.*

NSF should facilitate the establishment of a sustainable framework for long-term stewardship of federally funded research in science and engineering. To this end, funded projects should include multiple partners from key stakeholder communities including; universities, academic and research libraries, domain specialists, computer scientists, professional societies, standard-setting bodies, publishers, commercial and not-for-profit vendors, and funding agencies. The inclusion of multiple partners and interests is reflected in the statement, ''it takes a research community to preserve its data." Given the scale of the challenges, the projects should reflect the complexities of the stewardship of digital data within the research enterprise, the distributed nature of use, diverse responsibilities, and varied partnership interests

and needs.

The challenges of cross discipline collaboration are evident in a number of places, notably in the heterogeneity of the data sets and the cultural, organizational and technical frameworks within which the data are collected and recorded. This creates a very practical problem for curatorial facilities and their staffs. There was consensus that distributed approaches are required. Yet effective distributed organizations and technical architectures require shared tools, standards, protocols and procedures to enable efficient and interoperable systems and collaborations. These needs are particularly obvious with respect to the processes associated with ingest (submission), archiving (including storage and management), and reuse (retrieval of relevant information in a form useful to the future investigator). There has been work in various elements (for example, metadata and ontologies); less well-understood is the flow within organizations and across organizations, particularly where interdisciplinary collaborations are desired. Thus, prototyping these flows enables researchers to identify what works and what does not, where the hand-offs occur, and how to improve both the interfaces and the tools that support specific steps.

2. *NSF with other federal agencies 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 digital data stewardship—data curation, management, and preservation—as members of research teams.*

It is widely acknowledged that a new kind of professional is required whose expertise is critical to the successful stewardship of digital data resources. The digital environment requires new tools and skill sets. For example, scientists are typically not trained to manage the data sets, so there is an equivalent need within the disciplines to extend the current methodologies that are focused on collecting and analyzing data to include management of the data. Data management affects not only disposition after research is concluded but, with potential reuse of data, also requires that the scientific user more fully understand how data from an archive may have been stored and managed. It is the equivalent of understanding the potential effect of the instrumentation on the observation. The social sciences are already attuned to examining bias in data; extending that notion to understanding the implications for management of data is the logical next step. Thus, both the investigators and the data managers have a shared interest in managing data more effectively. Hence there is a broad need in many communities to understand data management, either as a future manager or a future consumer of the information.

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.*

Recommendation 1 calls for prototypes and models of the entire flow of information in the science and engineering arena. This recommendation builds on it by calling attention to several specific research areas that are subsumed into the flows described earlier. All of these are well-known problems in digital information management, and although there have been numerous research projects, the problems are far from solved. Moreover, they have not been solved with reference to integrated solutions that take into account distributed organizations and multiple communities. Thus, projects funded under this recommendation would comple-

ment work undertaken in the earlier recommendation and might attack a specific problem directly. This might include: data registration, automatic metadata creation, reference tools to accommodate ongoing evolution of commonly used terms and concepts and rights management.

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.*

Technical and economic sustainability were identified as critically important issues. Both are essential to engendering trust, which is a necessary precondition to viable curation facilities and to the larger framework of data stewardship of which each facility is a part. A number of approaches were discussed, combining local, regional and national/international scales and disciplinary and cross disciplinary content. However, it was generally agreed that developing good examples requires studying relevant organizational and behavioral issues as well as leveraging the research that has already been done in information economics, public goods, and infrastructure investments. Some of the topics that might be addressed include valuing intangibles, modeling collaboration, examining value proposition under various assumptions, motivations, incentives and prestige systems, and engendering trust.

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.*

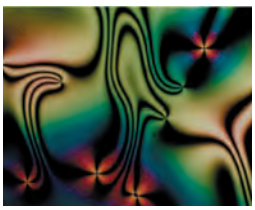
These recommendations seek to leverage the NSF processes to raise awareness of data curation in the research community, to acknowledge the need for the services of data curation and/or stewardship facilities, and to motivate researchers to participate through the requirements of the funding and reporting processes. Several key elements integral to a data management plan were identified. These include:

- how the data will be managed, by whom, and by what mechanism;
- whether data will be shared with the community (and if not, why not);
- whether the data will be preserved for the future, and if so, how; and
- how the data management, specifying preservation, curation, and/or stewardship, will be supported during and after the proposal is funded, as appropriate.

The information recovered through the reporting processes could be useful to the Foundation as it evolves its policies on data management.

Importantly, NSF should support training initiatives to ensure that the research community can fulfill this requirement while providing sufficient incentives to the community to ensure compliance. At the same time, it was acknowledged that there is a potential for such requirements to disadvantage applicants from less well-resourced universities. Thus, such requirements should be coupled with programs similar to those already in place, such as EPSCoR that seek to redress such imbalances.

The NSF has an opportunity to model internal business processes that other funding agencies might adopt. NSF has long been a leader among the basic research agencies. The workshop participants urged the Foundation to yet again to show leadership in the data management arena. As the Foundation has already recognized, it is fundamental to science and engineering research and the education enterprise in the digital age.



Selective Bibliography

General/Overview

Foster, I. (2005). Service-Oriented Science. Science, 308(5723), 814–817.

Hey, T., & Trefethen, A. (2003). E-Science and its Implications. Philosophical Transactions of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences, 361(1809), 1809-1825. RECOMMEND: Section 1—Introduction, Section 2—Technology Drivers for e-science and grids, Section 6—Scientific metadata, information and knowledge, and Section 7—Conclusions.

Towards 2020 Science(2006). Microsoft Corporation. http://research.microsoft.com/towards2020science/downloads/T2020S_ReportA4. pdf RECOMMEND: Summary -page 8, Section 1—Laying the Ground pages 14-20, Section 4—Conclusions and Recommendations pages 70–74.

Secondary Readings in General/Overview

Newman, H. B., Ellisman, M. H., & Orcutt, J. A. (2003). Data-intensive e-science frontier research. Communications of the ACM, 46(11), 68–77.

Vinge, V. (2006). 2020 Computing: The creativity machine. Nature, 440(7083), 411.

The Grid and the Semantic Web

De Roure, D., & Hendler, J. A. (2004). E-Science: the grid and the Semantic Web. IEEE Intelligent Systems, 19(1), 65–71.

Berman, F., Fox, G., and Hey. T., editors, Grid Computing: Making the Global Infrastructure a Reality," 1st Edition, John Wiley and Sons, LTD, England, 2003.

Gagliardi, F. (2005). The EGEE European grid infrastructure project. High Performance Computing for Computational Science-Vecpar 2004, 3402, 194–203.

Gagliardi, F., & Begin, M. (2005). EGEE—providing a production quality grid for e-science. Local to Global Data Interoperability—Challenges and Technologies, 2024 June 2005, 88–92.

Hendler, J. (2003). Science and the Semantic Web. Science, 299(5606), 520.

Cyberinfrastructure

Hey, T., & Trefethen, A. E. (2005). Cyberinfrastructure for e-Science. Science, 308(5723), 817–821.

Hey, T., & Tony Hey and Trefethen, A.E.,The Data Deluge: An E-Science Perspective; http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf#search=%22dat a%20curation%20and%20conference%20and%20e-science%22

Almes, G., Birnholtz, J. P., Hey, T., Cummings, J., Foster, I., & Spencer, B. (2004). CSCW and cyberinfrastructure: opportunities and challenges. Computer Supported Cooperative Work Conference Proceedings, 6–10 Nov. 2004, 270–273.

Berman, F., Berman, J. , C. Pancake, and Wu, L., A Process-Oriented Approach to Engineering Cyberinfrastructure, http://director.sdsc.edu/pubs/ENG/report/EAC_CI_Report-FINAL.pdf.

Moore, R., Berman, F., Schottlaender, B., Rajasekar, A., Middleton, D., JaJa, J., "Chronopolis—Federated Digital Preservation Across Time and Space", IEEECS International Symposium on Global Data Interoperability Challenges and Technologies, June 2005.

National Science Foundation Cyberinfrastructure Council. (2006). NSF's Cyberinfrastructure Vision for the 21st Century Discovery National Science Foundation. http://www.nsf.gov/od/oci/CI-v40.pdf.

National Science Foundation, Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, (the "Atkins Report"), http://www.nsf.gov/cise/sci/reports/atkins.pdf.

National Science Board, Long-Lived Data Collections Enabling Research and Education in the 21st Century, the National Science Board, http://www.nsf.gov/pubs/2005/nsb0540/start.jsp.

National Science Foundation

Over the past three years, a number of reports and papers on cyberinfrastructure and its

impact on research and education have been compiled. Links to a sample of some of the reports and papers are listed below.

Building a Cyberinfrastructure for the Biological Sciences; workshop held July 14–15, 2003. http://research.calit2.net/cibio/archived/CIBIO_FINAL.pdf http://research.calit2.net/cibio/report.htm.

CHE Cyber Chemistry Workshop; workshop held October 3–5, 2004. http://bioeng.berkeley.edu/faculty/cyber_workshop.

Commission on Cyberinfrastructure for the Humanities and Social Sciences; sponsored by the American Council of Learned Societies; seven public information-gathering events held in 2004; report in preparation. http://www.acls.org/cyberinfrastructure/cyber.htm.

Computation as a Tool for Discovery in Physics; report by the Steering Committee on Computational Physics. http://www.nsf.gov/pubs/2002/nsf02176/start.htm.

Cyberinfrastructure for the Atmospheric Sciences in the 21st Century; workshop held June 2004. http://netstats.ucar.edu/cyrdas/report/cyrdas_report_final.pdf.

Cyberinfrastructure for Engineering Design; workshop held February 28–March 1, 2005; report in preparation.

CyberInfrastructure and the Next Wave of Collaboration, D. E. Atkins, Keynote for EDUCAUSE Australasia, Auckland, New Zealand, April 5–8, 2005.

Cyberinfrastructure for Engineering Research and Education; workshop held June 5–6, 2003. http://www.nsf.gov/eng/general/Workshop/cyberinfrastructure/index.jsp.

Cyberinfrastructure for Environmental Research and Education (2003); workshop held October 30–November 1, 2002. http://www.ncar.ucar.edu/cyber/cyberreport.pdf.

CyberInfrastructure (CI) for the Integrated Solid Earth Sciences (ISES) (June 2003); workshop held on March 28–29, 2003, June 2003. http://tectonics.geo.ku.edu/ises-ci/reports/ISES-CI_backup.pdf.

Cyberinfrastructure and the Social Sciences (2005); workshop held March 15–17, 2005. http://www.sdsc.edu/sbe/.

Cyberlearning Workshop Series; workshops held Fall 2004–Spring 2005 by the Computing Research Association (CRA) and the International Society of the Learning Sciences (ISLS).
http://www.cra.org/Activities/workshops/cyberlearning.

Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences, F. Berman and H. Brady.
http://vis.sdsc.edu/sbe/reports/SBE-CISE-FINAL.pdf.

Geoinformatics: Building Cyberinfrastructure for the Earth Sciences (2004); workshop held May 14–15, 2003; Kansas Geological Survey Report 2004-48.
http://www.geoinformatics.info/.

Geoscience Education and Cyberinfrastructure, Digital Library for Earth System Education, (2004); workshop held April 19–20, 2004.
http://www.dlese.org/documents/reports/GeoEd-CI.pdf.

Identifying Major Scientific Challenges in the Mathematical and Physical Sciences and their CyberInfrastructure Needs, workshop held April 21, 2004.
http://www.nsf.gov/attachments/100811/public/CyberscienceFinal4.pdf.

Materials Research Cyberscience enabled by Cyberinfrastructure; workshop held June 17–19, 2004.
http://www.nsf.gov/mps/dmr/csci.pdf.

Multiscale Mathematics Initiative: A Roadmap; workshops held May 3–5, July 20–22, September 21–23, 2004.
http://www.sc.doe.gov/ascr/mics/amr/Multiscale%20Math%20Workshop%203%20%20Report%20latest%20edition.pdf.

An Operations Cyberinfrastructure: Using Cyberinfrastructure and Operations Research to Improve Productivity in American Enterprises"; workshop held August 30–31, 2004.
http://www.optimization-online.org/OCI/OCI.doc.
http://www.optimization-online.org/OCI/OCI.pdf.

Planning for Cyberinfrastructure Software (2005); workshop held October 5–6, 2004.
http://www.nsf.gov/od/oci/ci_workshop/index.jsp.

Preparing for the Revolution: Information Technology and the Future of the Research University (2002); NRC Policy and Global Affairs, 80 pages.
http://www.nap.edu/catalog/10545.html.

Polar Science and Advanced Networking: workshop held on April 24–26, 2003; sponsored by OPP/CISE.
http://www.polar.umcs.maine.edu/.

Research Opportunities in Cyberengineering/Cyberinfrastructure; workshop held April 22–23, 2004.
http://129.25.60.81/%7Eworkshop/.

Revolutionizing Science and Engineering Through Cyberinfrastructure: report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure; Daniel E. Atkins (Chair), January 2003.
http://www.nsf.gov/od/oci/reports/atkins.pdf.

A Science-Based Case for Large-Scale Simulation; workshop held June 24–25, 2003.
http://www.pnl.gov/scales/docs/volume1_72dpi.pdf.
http://www.pnl.gov/scales/docs/SCaLeS_v2_draft_toc.pdf.

Summit on Digital Tools for the Humanities; workshop held September 28–30, 2005.
http://www.iath.virginia.edu/dtsummit/SummitText.pdf.

Supplement to the President's Budget for FY 2006; Report by the Subcommittee on Networking and Information Technology Research and Development (NITRD), February 2005.
http://www.nitrd.gov/.

Trends in IT Infrastructure in the Ocean Sciences (2004); workshop held May 21–23, 2003.
http://www.geo-prose.com/oceans_iti_trends/oceans_iti_trends_rpt.pdf.

Data

Almes, G., Birnholtz, J. P., Hey, T., Cummings, J., Foster, I., & Spencer, B. (2004). CSCW and cyberinfrastructure: opportunities and challenges. Computer Supported Cooperative Work Conference Proceedings, 6–10 Nov. 2004, 270–273.

Choudhury, S., et al. (2006) Digital data preservation in astronomy. 2nd International Digital Curation Conference: Digital Data Curation in Practice, Glasgow, UK, November 21–22, 2006.

Humphrey, C. (2004). Preserving research data: a time for action. Symposium of the Canadian Conservation Institute on the Preservation of Electronic Records: New Knowledge and Decision-making, 83–90.

Humphrey, C., & Jacobs, J. (2004). Preserving research data. Communications of the ACM, 47(9), 27–29.

Muggleton, S. H. (2006). 2020 Computing: Exceeding human limits. Nature, 440(7083), 409–410.

Szalay, A., & Gray, J. (2006). 2020 Computing: Science in an exponential world. Nature, 440(7083), 413–414.

Rusbridge, C., & McHugh, A. (2005). Saving for the nation. [Electronic version]. Information Scotland, 3(4). see http://www.slainte.org.uk/publications/serials/infoscot/vol3(4)/vol3(4)article6.htm.

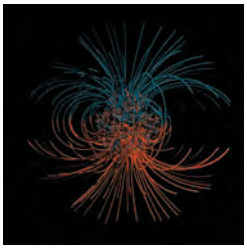
Virtual Organizations (secondary reading)

Camarinha-Matos, L. M. (2003). Infrastructures for virtual organizations—where we are. 2003 IEEE Conference on Emerging Technologies and Factory Automation. Proceedings, 16–19 Sept. 2003, vol. 2 405–414. RECOMMEND Section V—Support for Remote Operation and E-Science page 412.

Conference: Data Curation

1st Digital Curation Conference: An overview of the 1st Digital Curation Conference can be found online at http://www.ariadne.ac.uk/issue45/dcc-1st-rpt/.

Compiled by Kristi Jenkins, Physics and Astronomy Librarian, University of Minnesota.



Appendix A. List of Participants

Daniel Atkins
Director, Office of Cyberinfrastructure
The National Science Foundation
e-mail: datkins@nsf.gov

Francine Berman, co-chair
Director, San Diego Supercomputer Center University of California, San Diego
e-mail: berman@sdsc.edu

Henry E.. Brady
Director of the Survey Research Center
University of California, Berkeley
e-mail: hbrady@berkeley.edu

Suzanne Carbotte
Program Director
Marine Geoscience Data System
Lamont-Doherty Earth Observatory
Columbia University
e-mail: carbotte@ldeo.columbia.edu

Robert Chen
Interim Director and Senior Research Scientist
CIESIN, The Earth Institute,
Columbia University; CODATA
e-mail: bchen@ciesin.columbia.edu

Sayeed Choudhury
Associate Director for Library Digital Programs
Johns Hopkins University
The Milton S. Eisenhower Library
e-mail: sayeed@jhu.edu

Paul Constantine
Associate Dean of University Libraries for Research and Instructional Services
University of Washington Libraries
e-mail: pjc6@u.washington.edu

Peter Cornillon
Professor of Oceanography
University of Rhode Island
Graduate School of Oceanography
e-mail: pcornillon@gso.uri.edu

Bernard Dumouchel
Director General
Canada Institute for Scientific and Technical Information
National Research Council Canada
e-mail: bernard.dumouchel@nrc-cnrc.gc.ca

Amy Friedlander
Consultant
Shinkuro, Inc.
Washington, DC
e-mail: amy@shinkuro.com

Sara Graves
Director, Information Technology and Systems Center
The University of Alabama in Huntsville
e-mail: sgraves@itsc.uah.edu

Christopher Greer
Program Director, Office of Cyberinfrastructure
Cyberinfrastructure Advisor, Office of the Assistant Director for Biological Sciences
The National Science Foundation
e-mail: cgreer@nsf.gov

Myron Gutmann
Director, Inter-university Consortium for Political and Social Research
ICPSR - University of Michigan
e-mail: gutmann@umich.edu

Stephanie Hampton
Deputy Director
National Center for Ecological Analysis and Synthesis
University of California, Santa Barbara
e-mail: hampton@nceas.ucsb.edu

Robert Hanisch
Senior Scientist
Space Telescope Science Institute
Baltimore, MD
e-mail: hanisch@stsci.edu

Margaret Hedstrom
Associate Professor
School of Information
University of Michigan
e-mail: hedstrom@umich.edu

Charles (Chuck) Humphrey
Academic Director, Research Data Centre and Head, Data Library
University of Alberta
e-mail: humphrey@datalib.library.ualberta.ca

John King
Vice Provost for Academic Information
University of Michigan
e-mail: jlking@umich.edu

Wendy Lougee, co-chair
University Librarian, McKnight Presidential Professor
University of Minnesota Libraries
e-mail: wlougee@umn.edu

Barbara Lust
Professor
Cornell University
e-mail: bcl4@cornell.edu

Janet McCue
Director, Mann Library
Cornell University
e-mail: jam7@cornell.edu

James L. Mullins
Dean of Libraries
Purdue University Libraries
e-mail: jmullins@purdue.edu

Frank Rack
Executive Director, ANDRILL
ANDRILL Science Management Office
University of Nebraska-Lincoln
e-mail: frack2@unl.edu

Mark Sandler
Director, CIC-CLI
Committee on Institutional Cooperation
University of Michigan
e-mail: msandler@uiuc.edu

Rick Luce
Vice Provost and Director of the University Libraries
Emory University Libraries
Robert W. Woodruff Library
e-mail: rluce@emory.edu

Clifford Lynch
Executive Director
Coalition for Networked Information
Washington, DC
e-mail: cliff@cni.org

Don Middleton
Manager, Visualization and Enabling Technologies Section
NCAR
Boulder, CO
e-mail: don@ucar.edu

James Myers
Associate Director, Cyberenvironments
NCSA, University of Illinois at Urbana-Champaign
e-mail: jimmyers@ncsa.uiuc.edu

Chris Rusbridge
Director
Digital Curation Centre
University of Edinburgh
e-mail: C.Rusbridge@ed.ac.uk

Brian E. C. Schottlaender
University Librarian
University of California, San Diego, Libraries
e-mail: becs@ucsd.edu

MacKenzie Smith
Associate Director for Technology
MIT Libraries
e-mail: kenzie@mit.edu

Alex Szalay
Alumni Centennial Professor
Department of Physics and Astronomy
The Johns Hopkins University
e-mail: szalay@jhu.edu

Eric Van de Velde
Director Library Information Technology
California Institute of Technology
e-mail: evdv@library.caltech.edu

Todd Vision
Associate Director of Informatics
The National Evolutionary Synthesis Center
Department of Biology
University of North Carolina at Chapel Hill
e-mail: tjv@bio.unc.edu

Tyler Walters
Associate Director for Technology and Resource Services
Georgia Institute of Technology, Library and Information Center
e-mail: Tyler@gatech.edu

ARL Participants:

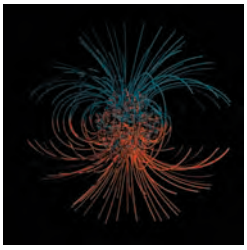
Prudence Adler
Associate Executive Director
Association of Research Libraries
e-mail: prue@arl.org

Julia Blixrud
Assistant Executive Director, External Relations
Association of Research Libraries
e-mail: jblix@arl.org

Heather Joseph
Executive Director
SPARC
e-mail: heather@arl.org

Consultant:

Amy Harbur
Shinkuro, Inc.
e-mail: harbur@shinkuro.com



Appendix B. Agenda

ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe

September 26–27, 2006
National Science Foundation
Room 595 Stafford II

Agenda

Tuesday, September 26

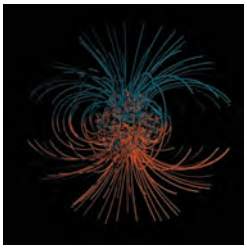
8:00a.m.–8:45a.m.	**Breakfast Room 595 Stafford II**
8:45a.m.–9:15a.m.	**Welcome** *Chris Greer, Office of Cyberinfrastructure, NSF*
9:15a.m.–9:45a.m.	**Introductions and Goals of the Workshop** *Fran Berman, SDSC and UCSD; Wendy Lougee, University of Minnesota; Prue Adler, ARL*
9:45a.m.–10:15a.m.	**Infrastructure** *Fran Berman, SDSC and UCSD*
10:15a.m.–10:30a.m.	**Break**

10:30a.m.–11:00a.m.	**New Partnership Models** *Bob Hanisch, STSI*
11:00a.m.–11:30a.m.	**Sustainable Economic Models** *Chris Rusbridge, University of Edinburgh*
11:30a.m.–12:00p.m.	**Review Key Themes from Position Papers Goals of the Break-out Sessions** *Amy Friedlander, Shinkuro, Inc.*
12:00p.m.–1:15p.m.	**Lunch at NSF**
1:30p.m.–5:15p.m.	**Three Concurrent Break-out Sessions with Mid-Afternoon Break Rooms 525, 545, and 585** **Infrastructure, Room 525** **New Partnership Models, Room 545** **Sustainable Economic Models, Room 585**
5:30p.m.–7:00p.m.	**Reception NSF Atrium, First Floor**

Wednesday, September 27

8:00a.m.–8:30a.m.	**Breakfast Room 595, Stafford II**
8:30a.m.–9:30a.m.	**Infrastructure Discussion and Recommendations** *Wendy Lougee, University of Minnesota and Rick Luce, Emory University*
9:30a.m.–10:30a.m.	**New Partnership Models Discussion and Recommendations** *Bob Hanisch, STSI and Brian E. C. Schottlaender, UCSD*
10:30a.m.–10:45a.m.	**Break**
10:45a.m.–11:45a.m.	**Sustainable Economic Models Discussion and Recommendations** *Fran Berman, SDSC and UC, and Chris Rusbridge, University of Edinburgh*
11:45a.m.–12:15p.m.	**Lunch**

12:15a.m.–12:30p.m.	**Break**
12:30p.m.–2:30p.m.	**Summary of Recommendations and Findings** *Group Discussion*
2:30p.m.–3:00p.m.	**Closing and Wrap up** *Clifford Lynch, CNI*



Appendix C. Plenary Papers

1. Chris Greer

   Welcome

2. Francine Berman

   Preserving Digital Collections for Research and Education

3. Robert Hanisch

   Digital Data Discovery and Federation in Astronimy: A Partnership of the Virtual Observatory, Scholarly Publishers, and Research Libraries

4. Chris Rusbridge

   Sustainability Issues

5. Amy Friedlander

   Workshop Participant's Position Papers: Overview and Charge to Breakout Groups

"*Sometime in the 2010s, if all goes well, the Large Synoptic Survey Telescope (LSST) will start to bring a vision of the heavens to Earth. Suspended between its vast mirrors will be a three billion-pixel sensor array, which on a clear winter night will produce 30 terabytes of data. In less than a week this remarkable telescope will map the whole night sky …. And then the next week it will do the same again … building up a database of billions of objects and millions of billions of bytes.*"

Nature 440:383

The Fragility of Memory in a Digital Age

"In 1964, the first electronic mail message was sent from either MIT, the Carnegie Institute, or Cambridge University. The message does not survive, however, and so there is no documentary record to determine which group sent the pathbreaking message."

Report of the Task Force on Archiving of Digital Information
Commission on Preservation and Access and the Research Libraries Group

Data Strategic Planning Group

- Sylvia Spengler (Chair)
- Deborah Crawford
- Cheryl Eavey
- James French
- Chris Greer
- Elizabeth Lyons
- David Lightfoot
- Fillia Makedon
- Jose Munoz
- Dan Newlon
- Nigel Sharp

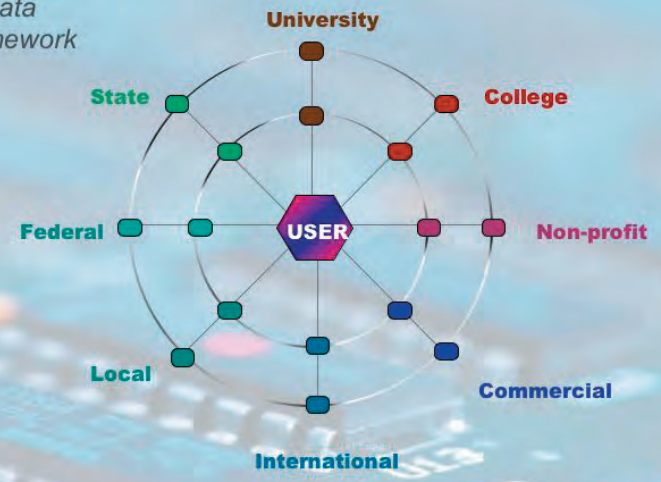
Vision:

"… a vision in which science and engineering digital data are routinely deposited in well-documented form, are regularly and easily consulted and analyzed by specialists and non-specialists alike, are openly accessible while suitably protected, and are reliably preserved."

Goals:

- To catalyze the development of a system of science and engineering data collections that is open, extensible and evolvable.
- To support development of a new generation of tools and services facilitating data acquisition, mining, integration, analysis, and visualization.

National Digital Data Framework



User-centric, Multilevel, Nimble, Sustainable, Reliable

The Challenge

"If we are effectively to preserve for future generations the …. corpus of information in digital form that represents our cultural record, we need … to commit ourselves technically, legally, economically, and organizationally to the full dimensions of the task."

Report of the Task Force on Archiving of Digital Information, 1996

The Universities

"*Ever since their inception, universities have been occupied with the fundamental elements of what we now call 'knowledge management', i.e. the creation, collection, preservation and dissemination of knowledge.*"

A. Oosterlinck, Knowledge Management in Post-Secondary Education: Universities

The Academic Libraries

"*It is to the research library community that others will look for the preservation of ... digital assets, as they have looked to us in the past for reliable, long-term access to the 'traditional' resources and products of research and scholarship.*"

Association of Research Libraries (ARL) Strategic Plan 2005-2009

The Challenge

"*If we are effectively to preserve for future generations the .... corpus of information in digital form that represents our cultural record, we need ... to commit ourselves technically, legally, economically, and organizationally to the full dimensions of the task.*"

Report of the Task Force on Archiving of Digital Information, 1996

Preserving Digital Collections for Research and Education

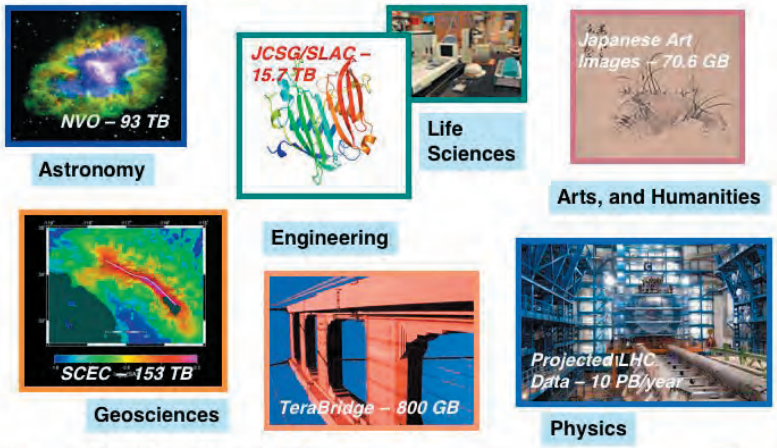
ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, September 2006

Dr. Francine Berman

Director, San Diego Supercomputer Center

Professor and High Performance Computing Endowed Chair, UC San Diego

Digital Data is Fundamental to 21st Century Research and Education



NSF Proposed 5 Year Goal

"to catalyze the development of a system of science and engineering data collections that is **open, extensible, and evolvable** … a **national digital data framework** consist[ing] of a range of data collections and managing organizations … simultaneously **local, regional, national, and global** in nature."

NSF CI Vision for 21st Century Discovery

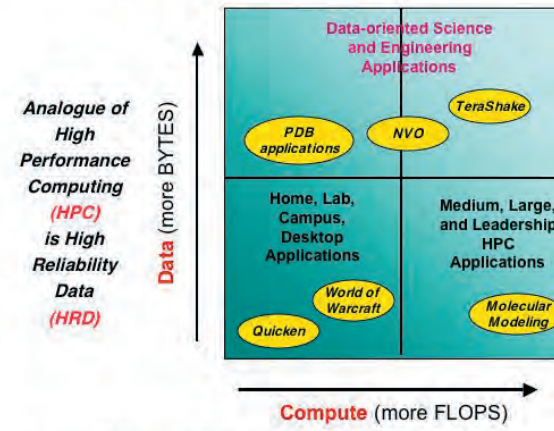
Workshop Exercise: How to Develop a National Framework for Digital Data Stewardship

- **Infrastructure.** How do we manage data now and migrate it successfully over future generations of technologies, standards, formats, and institutions?
- **Partnerships.** What mix of individuals and organizations should be involved in data preservation? What creative partnerships can be developed between the multiple sectors?
- **Sustainable Economic Models.** What models are required to sustain data management and preservation efforts over the long term?

Infrastructure

How do we manage data now and migrate it successfully over future generations of technologies, standards, formats, and institutions?

Today's Applications Cover the Spectrum

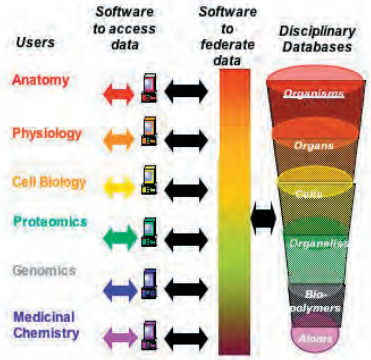


Large-scale data required as input, intermediate, output for many modern HPC applications

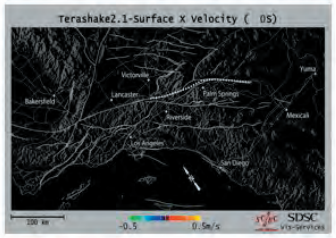
Applications vary with respect to how well they can perform in distributed mode (grid computing)

Working with Data: Data Driving New Discovery in Research and Education

Data Integration in the Biosciences



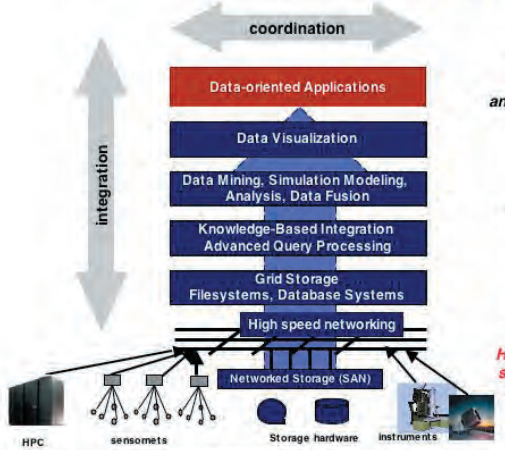
Data Simulation in the Geosciences



TeraShake Simulation of 7.7 magnitude earthquake on lower San Andreas Fault

- Physics-based dynamic source model – simulation of mesh of 1.8 billion cubes with spatial resolution of 200 m
- Simulation generates 45+ TB data

Data Management and Preservation Infrastructure Must Support Active Use of Digital Data



How do we combine data, knowledge and information management with simulation and modeling?

How do we represent data, information and knowledge to the user?

How do we detect trends and relationships in data?

How do we obtain usable information from data?

How do we collect, access, organize and preserve data?

How do we configure resources go support the generation, collection, use, and understanding of data?

National Research Data Repository: SDSC DataCentral

- **Community data collection and database hosting**
  - Collection management services
- First broad program of its kind to support research and community research data collections and databases
- **Comprehensive resources**
  - ***Disk:*** 400 TB accessible via HPC systems, Web, SRB, GridFTP
  - ***Databases:*** DB2, Oracle, MySQL
  - ***SRB:*** Collection management
  - ***Tape:*** 6 PB, accessible via file system, HPSS, Web, SRB, GridFTP
  - ***24/7 operations, collection specialists***

New Allocated Data Collections include

- **Bee Behavior** (Behavioral Science)
- **C5 Landscape DB** (Art)
- **Molecular Recognition Database** (Pharmaceutical Sciences)
- **LIDAR** (Geoscience)
- **AMANDA** (Physics)
- **SIO_Explorer** (Oceanography)
- **Tsunami and Landsat Data** (Earthquake Engineering)
- **Terabridge** (Structural Engineering)

DataCentral infrastructure includes: Web-based portal, security, networking, UPS systems, web services and software tools

Public Data Collections Hosted in SDSC's DataCentral

Seismology	3D Ground Motion Collection for the LA Basin
Atmospheric	Sciences50 year Downscaling of Global Analysis over California Region
Earth Sciences	NEXRAD Data in Hydrometerology and Hydrology
Elementary Particle Physics	AMANDA data
Biology	AfCS Molecule Pages
Biomedical Neuroscience	BIRN
Networking	Backbone Header Traces
Networking	Backscatter Data
Biology	Bee Behavior
Biology	Biocyc (SRI)
Art	C5 landscape Database
Geology	Chronos
Biology	CKAAPS
Biology	DigEmbryo
Earth Science Education	ERESE
Earth Sciences	UCI ESMF
Earth Sciences	EarthReLorg
Earth Sciences	ERDA
Earth Sciences	ERR
Biology	Encyclopedia of Life

Life Sciences	Protein Data Bank
Geosciences	GEON
Geosciences	GEON-LIDAR
Geochemistry	Kd
Biology	Gene Ontology
Geochemistry	GERM
Networking	HPWREN
Ecology	HyperUer
Networking	IMDC
Biology	Interpro Mirror
Biology	JCSG Data
Government	Library of Congress Data
Geophysics	Magnetics Information Consortium data
Education	UC Merced Japanese Art Collections
Geochemistry	NAVDAT
Earthquake Engineering	NEESIT data
Education	NSDL
Astronomy	NVO
Government	NARA
Anthropology	GAPP

Neurobiology	Salk data
Seismology	SCEC TeraShake
Seismology	SCEC CyberShake
Oceanography	SIO Explorer
Networking	Skitter
Astronomy	Sloan Digital Sky Survey
Geology	Sensitive Species Map Server
Geology	SD and Tijuana Watershed data
Oceanography	Seamount Catalogue
Oceanography	Seamounts Online
Biodiversity	WhyWhere
Ocean Sciences	Southeastern Coastal Ocean Observing and Prediction Data
Structural Engineering	TeraBridge
Various	TeraGrid data collections
Biology	Transporter Classification Database
Biology	TreeBase
Art	Tsunami Data
Education	ArtStor
Biology	Yeast regulatory network
Biology	Apoptosis Database
Cosmology	LUSciD

Preservation Infrastructure for Research and Education Data

- **Data repositories must**
  - ***Be trusted facilities***
  - *Be highly reliable*
  - *Provide high capacity and state of the art storage.*
  - *Be able to sustain data over generations of media, users, owners, institutions, etc.*
  - *Have a 5 year, 50 year, 100+ year plan*
  - *Etc.*

Issues for research and education collections:

- Who is responsible for **selecting** community reference and research collections for preservation?
- Who is responsible for **preserving** the collections and for how long?
- Which group is responsible for **ensuring the integrity** of the collection as it evolves over time?
- What are the **community's responsibilities** with respect to funding the collection?
- Who makes the decision to **de-select** or downgrade the level of service in maintaining the collection?

Partnership

What mix of individuals and organizations should be involved in data preservation?

What creative partnerships can be developed between the multiple sectors?

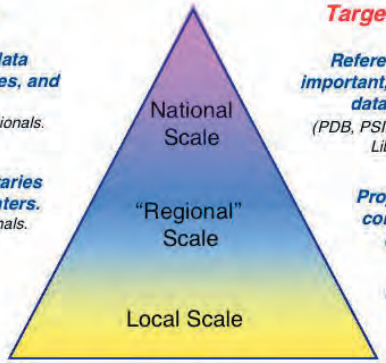
The "Data Pyramid": An Organizational Structure for Talking about Research Data

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.



Target Collections

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

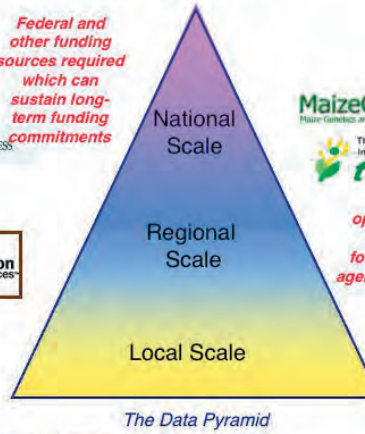
Project and some community data collections

Personal data collections

The Data Pyramid

Opportunities for Partnerships throughout the "Data Pyramid"

Federal and other funding sources required which can sustain long-term funding commitments



Creative partnership opportunities at the mid-range: Universities, foundations, government agencies, private sector, etc.

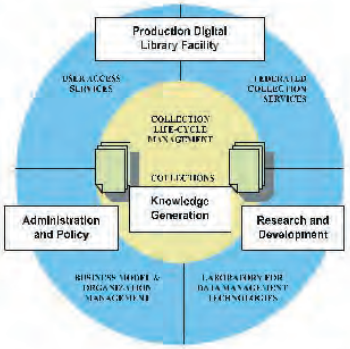
Commercial opportunities at the low end

The Data Pyramid

Chronopolis™: A Partnership Approach to Long-term Digital Preservation

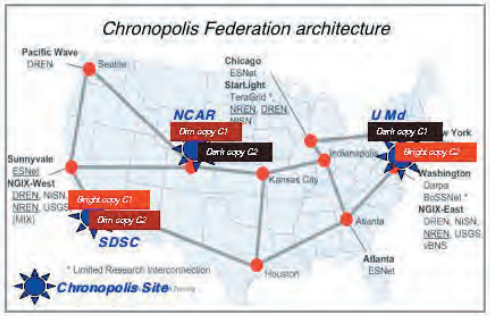
SDSC , the UCSD Libraries, NCAR, UMd, NARA, NSF working together on long-term preservation of digital collections

- **Chronopolis™ provides a partnership approach to infrastructure for long-term preservation integrating**
  - **Collection ingestion**
  - **Access and Services**
  - **Research and development** for new functionality and adaptation to evolving technologies
  - **Business model, data policies, and management** issues critical to success of the infrastructure



Chronopolis™ – Replication and Distribution

- **3 replicas of valuable collections considered reasonable mitigation for risk of data loss**
- **Chronopolis Consortium will store 3 copies of preservation collections:**
  - **"Bright copy"** – Chronopolis site supports ingestion, collection management, user access
  - **"Dim copy"** – Chronopolis site supports remote replica of bright copy and supports user access
  - **"Dark copy"** – Chronopolis site supports reference copy that may be used for disaster recovery but no user access
- **Each site may play different roles for different collections**



Economic Sustainability

What models are required to sustain data management and preservation efforts over the long term?

Providing Sustainable and Reliable Data Infrastructure Incurs Real Costs

Entity at risk	Size	What can go wrong	Frequency	Minimum number of replicas needed to mitigate risk	Administrative support FTEs
File	~2 MB	Corrupted media, disk failure	1 year	2 copies in single system	System Admins
Tape	~200 GB	+ Simultaneous failure of 2 copies	5 years	3 homogeneous systems	+ Storage Admin
System	~10 TB	+ Systemic errors in vendor SW, or Malicious user, or Operator error that deletes multiple copies	15 years	3 independent, heterogeneous systems	+ Database Admin + Security Admin
Archive	~1 PB	+ Natural disaster, obsolescence of standards	50 - 100 years	3 distributed, heterogeneous systems	+ Network Admin + Data Grid Admin

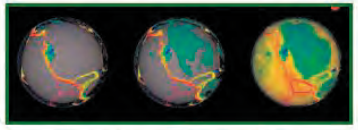
Less risk means more replicants, more resources, more people

Making the Case: What Happens if We Don't Preserve Our Most important Digital Data?

Life sciences research would have the resources available in roughly the 1970's – no PDB, no Swiss-Prot, no PubMed, Etc.

What if digital history was only as old as the current storage media?

Long-lived federal, state, and local records would need to remain on paper.



New discoveries from climate and other predictive simulation models which utilize longitudinal data would dramatically slow

iTunes would store only current music, NetFlix would provide only current movies

Data Preservation Infrastructure Requires a Different Model than Computing Infrastructure

	Supercomputing	Data Preservation
Resources	Large-scale HPC system (compute, storage, networking, etc.), staff for maintenance, user services, etc.	Archival storage, networking, web access, plan/resources for replication/backup, staff for maintenance, database management, user services
Resource Refresh	**Large-scale capability/capacity** key -- New HPC system not required to be compatible with old although there should be application transition path	**Continuity** key -- Data collections must migrate over new generations of technology and standards without disruption for users
Metrics of Success	(**Deep impact**) Newsworthy/Breakthrough science and engineering; (**Competitiveness**) Good ranking on the Top500 list; (**Broad impact**) number of users, etc.	**No serious loss of data**; adequate preservation of reference collections, appropriate research collections
Funding Model	New **"one time" funding** for each HPC resource	Funding commitment needs to address **long-term consistency** needed for data collections

Making Infinite Funding Finite

- **Difficult to support infrastructure for data preservation as an infinite, increasing mortgage**
- Creative solutions can be used to create sustainable economic models

User fees, recharges

Relay Funding

Endowments

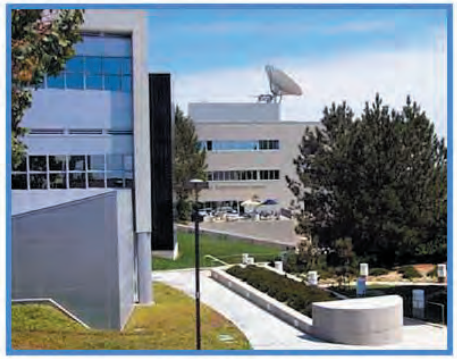
Consortium support

Hybrid solutions

An Opportunity to Pioneer

- **NSF-ARL Workshop provides the opportunity to help NSF provide community leadership** for research and education digital data collections
- Workshop recommendations will provide valuable community **input to future programs and funding**
- **Workshop recommendations should help focus the discussion** on infrastructure, partnerships, and sustainability for digital data management and preservation
- **Thanks to NSF for their vision in this area and their support of this workshop.**

Thank You



berman@sdsc.edu
www.sdsc.edu

Digital Data Discovery and Federation in Astronomy: A Partnership of the Virtual Observatory, Scholarly Publishers, and Research Libraries

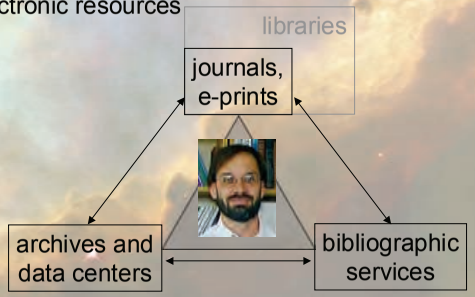
Robert Hanisch
US National Virtual Observatory
Space Telescope Science Institute
Baltimore, MD

Electronic information in astronomy

- Astronomy was one of the first scientific disciplines to pioneer e-publishing (ApJLett 1995, ApJ and AJ 1996)
- Astronomy has comprehensive e-abstract and bibliographic services
  - Astrophysics Data System, SIMBAD, NED
- Astronomy makes extensive use e-preprints on arXiv.org
- Astronomy data is archived and is generally publicly accessible
  - NASA mission archives
  - ground-based observatories (U.S., Europe, Australia, etc.)
  - data centers (catalogs, tables, value-added services)

Electronic information in astronomy

- E-journals link to underlying data, and data archives link to e-journals, through a system of persistent, unique identifiers
- Astronomers interact with a set of connected electronic resources



The Virtual Observatory

- The *Virtual Observatory* is a framework for providing access to distributed data, distributed services. The VO is about *data discovery, access, and integration*, and combining data with computational services.
- Motivation:
  - The data deluge. Needs tools to locate and sift through immense collections and to correlate data from many resources. ~500 TB of data currently available.
  - Scientific discovery opportunities exist at the intersections of diverse data sets.
- Astronomy, of course! Space science, solar physics, aeronomy, seismology, oceanography, hydrology, biology, genomics, medicine. […]ology and […]onomy.
- Keywords: *Metadata, interoperability*

Data/Information in the VO

- Basic data
  - digital images, spectra, time series, catalogs, tables
- Simulations
  - models (results, computer codes, computational services)
  - virtual observations

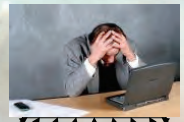
not discoverable through text-based search engines

- Analysis and interpretation
  - journals, e-preprints
  - reprocessed and enhanced data
- Name-resolution services
  - "Andromeda Galaxy", "Messier 31", "M31", "NGC 224", "UGC 454", etc. ==> ra 00h 42m 44s, dec +41° 16' 08"
  - Geographic equivalent of "Glenn Dale, MD", "20769", "Prince George's County" ==> 76° 48' 19 □, 38° 58' 36" N

The Virtual Observatory in Astronomy

- The Virtual Observatory enables new science by greatly enhancing access to data and computing resources. The VO makes it easy to locate, retrieve, and analyze data from archives and catalogs worldwide.
- The VO is NOT a huge centralized data repository.
- The VO provides standard protocols for obtaining data from *distributed collections.*
- The VO is *national* (US NVO) and *international* (IVOA).
  - US National Virtual Observatory is partnership of (real) observatories, universities, IT/CS groups. NSF-funded.
  - International VO Alliance is self-organized collaboratory and standards body (□ 3C-like).

□ ithout VO



n services,
n interfaces

archive 1
archive 2
archive 3
survey 1
survey 2
survey 3
service 3
service 2
service 1

□ ith VO

n services,
"1" interface

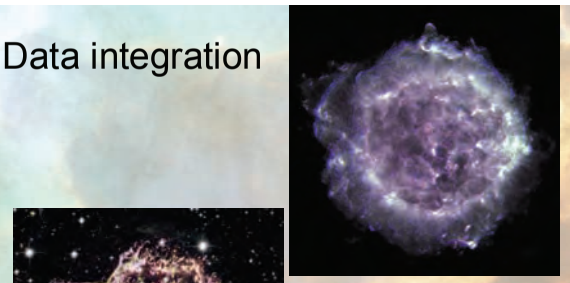
archive 1
archive 2
archive 3
survey 1
survey 2
survey 3
VO
service 3
service 2
service 1

Data integration

Cas A supernova remnant

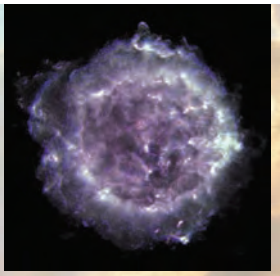
optical (HST)

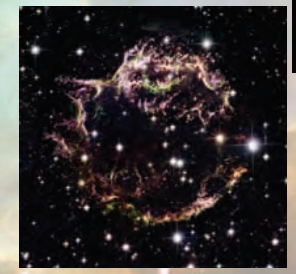
Data integration

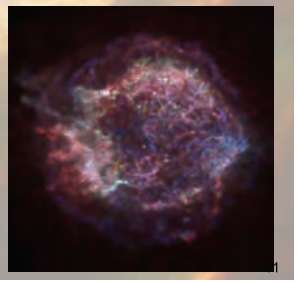


radio (VLA)

Data integration

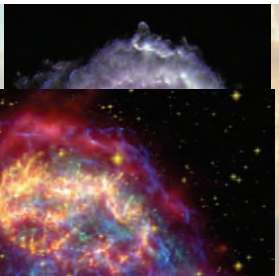






x-ray (Chandra)

Data integration



x-ray (Chandra)

Data integration

The Key to the VO: Interoperability

- Metadata standards
- Data discovery
- Data requests
- Data delivery
- Database queries, responses
- Distributed applications; web services
- Distributed storage; replication
- Authentication and authorization

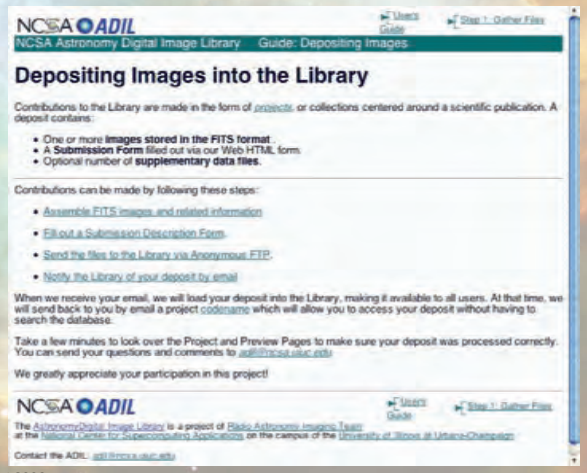
The data preservation problem

- Research communities publish peer-reviewed journal papers that describe highly processed data.
- Long-term preservation and curation systems for digital journal content are not currently in place; *only the graphical representations of data are being saved*.
- The research cannot be verified and the results cannot be easily compared to other data in order to broaden impact.
- Public funds invested in scientific research do not have maximum return on investment. Essential legacy datasets are being lost.

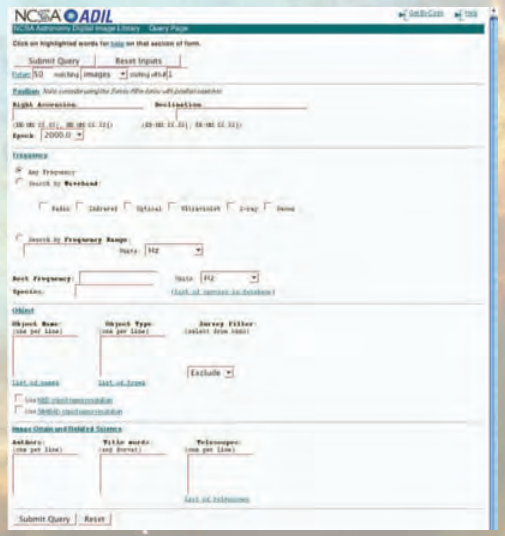
Approach

- Integrate digital data management into the publication process (data capture, review, metadata tagging and validation, storage).
- Exploit emerging information technology standards for managing distributed data collections, including digital journals.
- Provide multiple access methods to digital data to maximize visibility and re-use.
- Exploit information management and curation experience in the university libraries and build on long-term institutional commitments to preservation.

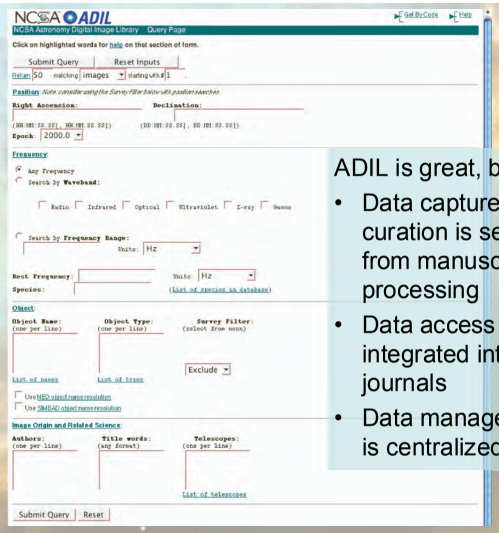
Astronomy Digital Image Library



ADIL query



ADIL query



ADIL is great, but…

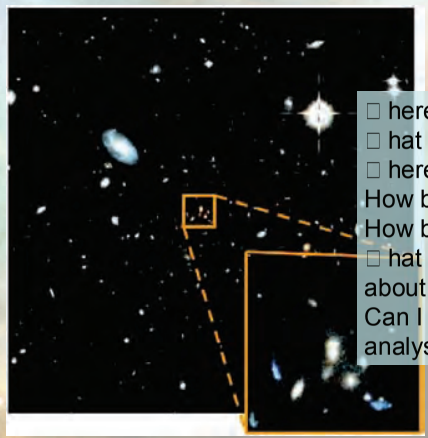
- Data capture and curation is separate from manuscript processing
- Data access is not integrated into the journals
- Data management is centralized

Storyboard

Storyboard

Hubble Space Telescope image.
Most distant cluster of galaxies known.
□ hat more can I find out?

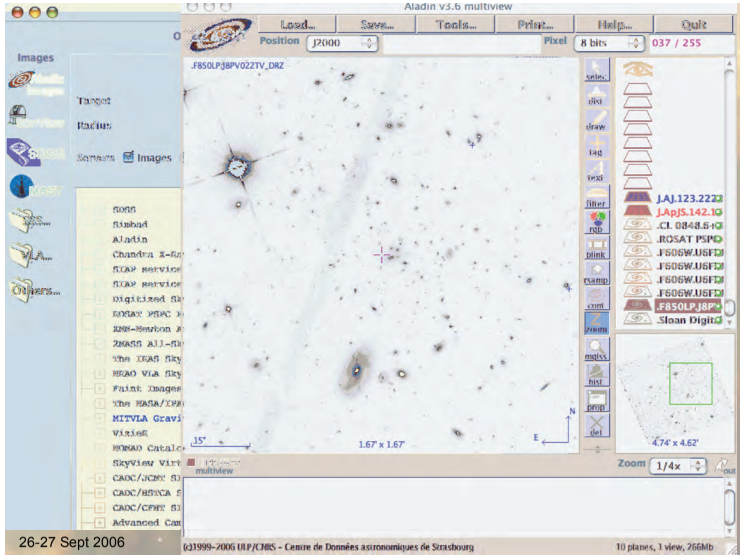
Storyboard

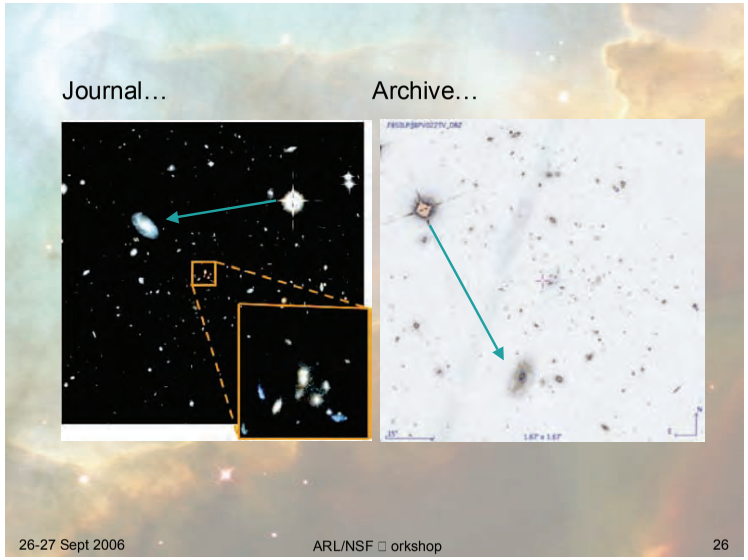


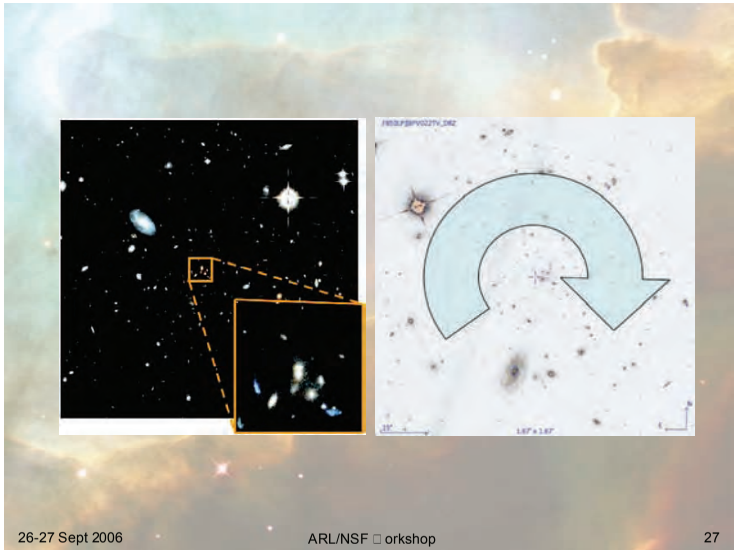
□ here is this?
□ hat is the image scale?
□ here is north?
How bright is the star?
How bright is the galaxy?
□ hat else is known about this region?
Can I trust the data analysis in this paper?

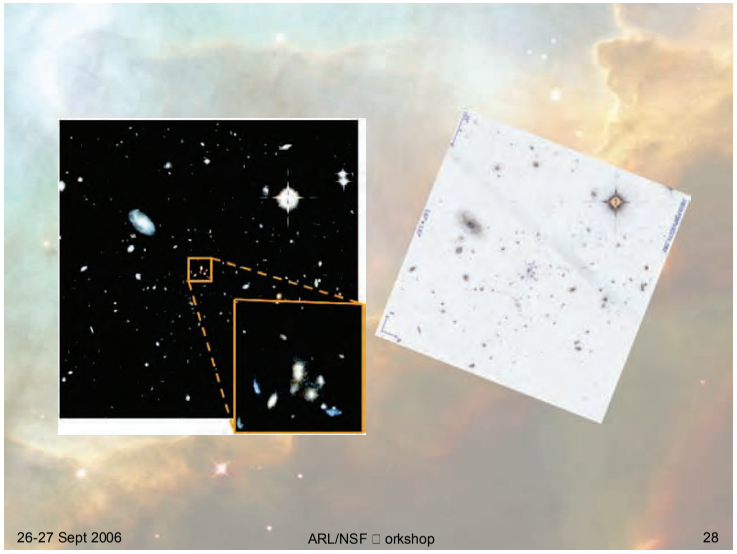
Storyboard

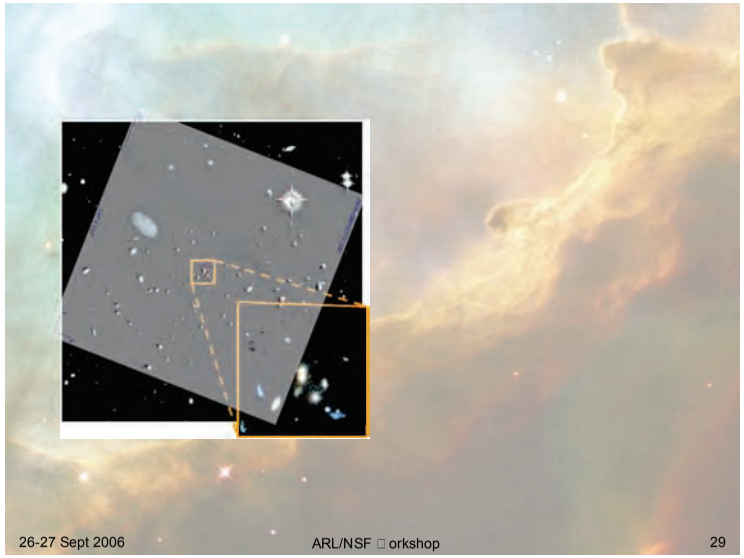
Save file
Copy to my VOSpace
Display and
ompare

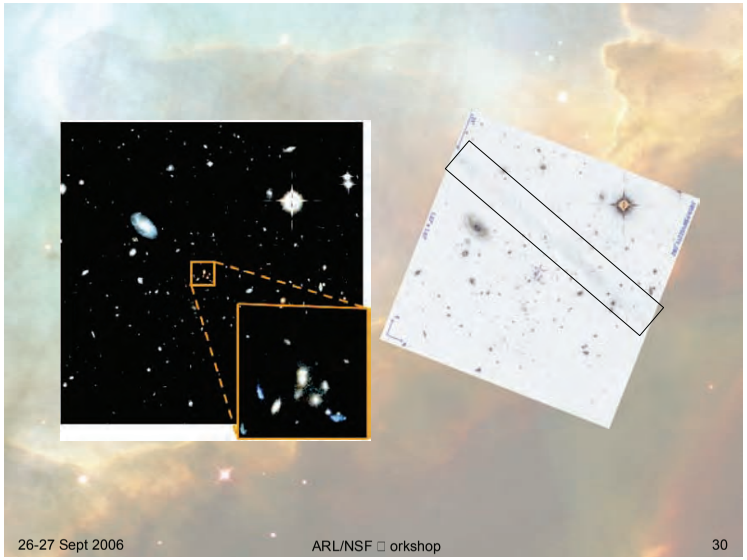

Aladin v3.6 multiview
Load...
Save...
Tools...
Print...
Help...
Quit
Position
J2000
Pixel
8 bits
037 / 255
Images
Target
Radius
Zoom
1/4x
multiview
1.67 x 1.67
4.74' x 4.62'
(c)1999-2006 ULP/CNRS - Centre de Données astronomiques de Strasbourg
10 planes, 1 view, 266Mb


Journal…
Archive…

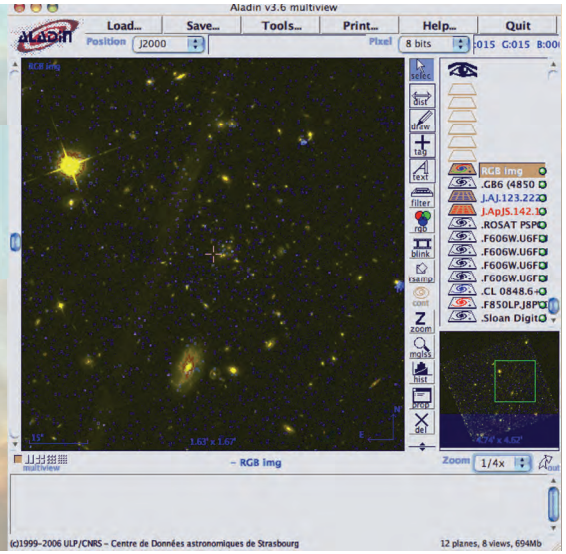




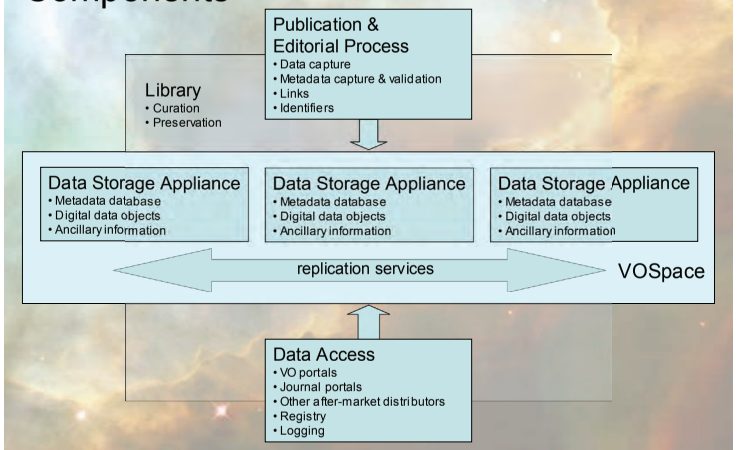




Is there any X-ray emission from this cluster of galaxies?



Components



Data preservation tasks & partners

- Tasks (partners)
  - Metadata definition (VO, library)
  - Content management tool evaluation/selection (Fedora) (VO, library)
  - Physical storage and replication (VO, library, publisher)
  - Publication process revisions and testing (publisher, editorial staff)
  - Policy development (editorial staff, professional society)
  - Business model development (publisher, professional society)

Digital data discovery and access is essential for the research community

- Data re-use, with provenance
- Optimization of public investment in science
- Increasing the discovery space
- Creation of a research legacy
- Integrity in scientific publication

Success requires cooperation among providers (individual and institutional), publishers, curators, and preservationists

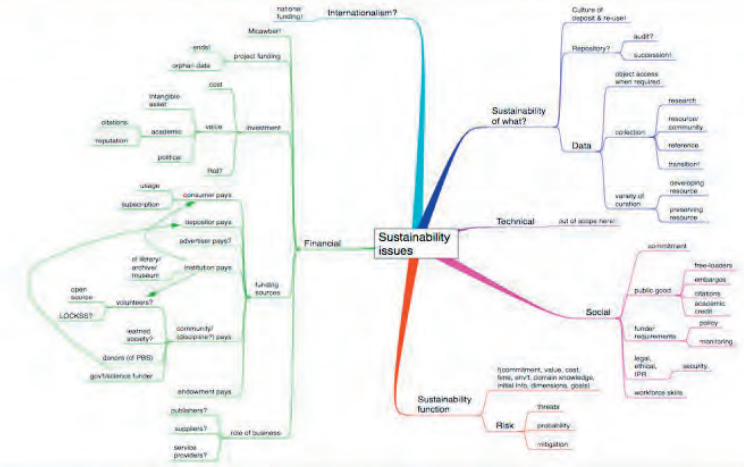
D|C|C a centre of expertise in data curation and preservation

ARL NSF workshop: Sustainability Issues

Chris Rusbridge
(not an economist!)

Funded by: JISC e-Science

D|C|C a centre of expertise in data curation and preservation



ARL NSF Curation Sustainability

D|C|C a centre of expertise in data curation and preservation

Sustainability: the emerging worry

- Sustainability work package in DCC (new grant!)
- JISC/NDIIPP meeting addressed it
- AHRC report draft soon
- Research Information Network report draft
- JISC study on sustainable IT systems for HE
- This workshop, NSF strategy

ARL NSF Curation Sustainability

D|C|C a centre of expertise in data curation and preservation

Sustainability of what?

- Repository as an organisation
- Repository as a service
- Repository as a system
- Repositories as a network (federation?)
- Collections and objects supported by repositories

- Commit to collection: contract the manager!

ARL NSF Curation Sustainability

DCC a centre of expertise in data curation and preservation

Sustainability of what?

- Curation service
  - Separate service from collection
  - Funding always finite: 5 + 5 then re-compete?
  - Relay approach: hand on in good order
  - Succession! Start with the plan for your own end…
- Culture of deposit & re-use!
  - One of the most important social dimensions, but out of scope here…

DCC a centre of expertise in data curation and preservation

Sustainability of what?

- Data
  - Digital object access when required (for long future time)
  - Collection: (LLDDC classification)
    - Research (local)
    - Resource (community)
    - Reference (global)
  - Transition from one class to another… and misclassification!

ARL NSF Curation Sustainability

DCC a centre of expertise in data curation and preservation

Sustainability for what?

- Variety of curation approaches
  - Developing resource
  - Preserving resource
- Significant properties have a big impact
  - Produce bit stream as ingested?
    - All the work for the consumer
  - Produce full look and feel as ingested? Expensive!
    - May also be unfamiliar for future consumer
  - Somewhere between?
    - Depends on goals…

ARL NSF Curation Sustainability

DCC a centre of expertise in data curation and preservation

Social factors

- Commitment essential… much more than anything else (cf persistent identifiers)
- Funder requirements express social determination
  - Policy & grant application forms, selection criteria
  - Monitoring essential
- Legal, ethical, IPR impacts all significant
- Public good questions
  - Academic credit (citations?)
  - Free-loaders (embargos?)
  - Disciplines are different!
- Workforce skills: researcher, data librarian/scientist

ARL NSF Curation Sustainability

Sustainability a function of...

- Commitment
- Goals
- Value and cost
- Business model
- Time
- Environment
- Domain knowledge and information
- Dimensions (how much stuff)
- Technical approaches
- Usage

DCC a centre of expertise in data curation and preservation

Risk

- Risk affects the cost of sustainability
  - Threats
  - Probability
  - Mitigation
- Sustained care from creation
- Single point of loss; correlated failures
- Network dependencies
- Policy change!
- Independence, federation, redundancy!
  - Expensive!

ARL NSF Curation Sustainability

DCC a centre of expertise in data curation and preservation

Financial sustainability 1

- Micawber: "... that if a man had twenty pounds a year for his income, and spent nineteen pounds nineteen shillings and sixpence, he would be happy, but that if he spent twenty pounds one he would be miserable."
- "There's plenty of money… there's just not plenty of money for everything!" (Courant)

ARL NSF Curation Sustainability

DCC a centre of expertise in data curation and preservation

Financial sustainability 2: projects

- Traditional research project approach:
  - Produces unsustainable resources
  - PIs focus on next project proposal
  - RAs focus on next job application
  - Result: no metadata, orphan data

ARL NSF Curation Sustainability

Financial sustainability 3: investment

- How you justify a long-term spend: persuasive? No!

```
Return on investment = value - cost
```

  - Intangible asset: hard to value; situated, multi-scaled
  - Aggregate rather than individual
  - Academic value is key
    - Citations support this: needs work
    - Reputation is the target currency
    - But dollars pay the bills!

DCC a centre of expertise in data curation and preservation

Value

- "… the by-products of our research may be more significant than our soon dated theoretical insights." (Seeger 2004, quoted by Kevin Bradley, APSR)
- "I think I would be safe in saying that worldwide hundreds of millions of dollars' worth of crystallographic data is lost each year. For spectra and synthetic chemistry it will be at least 10 times greater. Many synthetic chemists say they are interested in failed reactions - and these are almost never published!" (Murray-Rust blog)
- Value of curation service can grow from re-use promotion & community proxy activities (eg BADC & CF conventions, ICPSR & DDI)
- (But the value of data is easily negated, at creation and after)

ARL NSF Curation Sustainability

DCC a centre of expertise in data curation and preservation

Financial sustainability 4: the 8 pillars of wisdom?

- Someone has to pay…
  - Consumer pays: subscription or usage?
  - Depositor pays (ie grant or institution)?
  - Institution pays (IR, cf library/archive/museum)
  - Community (discipline repository?) pays
    - Government, or science funder
    - Learned society?
    - Volunteers (cf open source, social computing, LOCKSS)?
  - Side effect (advertiser) pays (unlikely for much data?)
  - Endowment or donor pays…
- Diversity?

ARL NSF Curation Sustainability

DCC a centre of expertise in data curation and preservation

Role of business?

- Publishers?
  - Traditional role in related areas, under threat…
  - May want a piece, do we want them?
  - "Give me your content, then rent it back"
- Search engines? GoogleInChi (PMR blog)
- Suppliers?
  - Big impact on costs
- Service providers?
  - Bring their own risk issues

ARL NSF Curation Sustainability

D|C|C a centre of expertise in data curation and preservation

Role of libraries

- 2-4% of university budgets (see Courant quote)?
- Traditional role in sustaining the raw material of scholarship
  - Looking for new roles in the digital world?
  - Many unsaid assumptions from publishing paradigm?
  - Domain knowledge: wide but not deep
  - Involvement in data creation low

D|C|C a centre of expertise in data curation and preservation

International scale science & data?

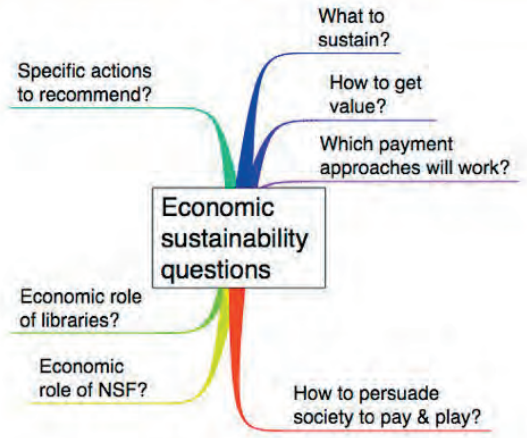
- National funding! (really agency funding)
  - Negotiations, deals, MoUs, Concordats, "treaties"

D|C|C a centre of expertise in data curation and preservation

So, sustainability…

- Digital data repositories already sustained > 30 years
  - How?
  - Vision, leadership, commitment
- Libraries, archives, museums sustained 100s of years
  - How?
  - Aggregate value proposition
  - Perception now under threat!
- Our job to help identify the next steps toward digital data sustainability

D|C|C a centre of expertise in data curation and preservation



Workshop Participants' Position Papers: Overview and Charge to Breakout Groups

Amy Friedlander
Shinkuro, Inc.

ARL Workshop on New Collaborative Relationships:
The Role of Academic Libraries
in the Digital Data Universe

September 26, 2006
[Rev. September 25, 2006]

Three questions:

- What did you say?
- What was common and what was not?
- How can this information contribute to the goals of the workshop? To enable the breakout sessions to formulate recommendations to the NSF and potentially to other funders

The Waterfront:

Seventeen unique papers on inter-related topics: Institutions, collections and individuals

- Partnerships (institutional, individual, cultural challenges)
- Funding and sustainable economic models including human resources
- Collections (integrated/federated yet heterogeneous and distributed data sets, publications, "grey literature")
- Access ("open" policies, metadata, ontologies, discovery)
- Policies (institutional, rights management, security, confidentiality, privacy, life cycle [creation/capture/acquisition/management/ appraisal/ retention], incentives)
- Technical infrastructure and tools
- R & D (including prototypes)
- Science education and communication/outreach (scholarly and public, including professional training)

A Public policy framework

Getting to action; making sense of scale and stakeholders

- Institutions
- Collections
- Individuals

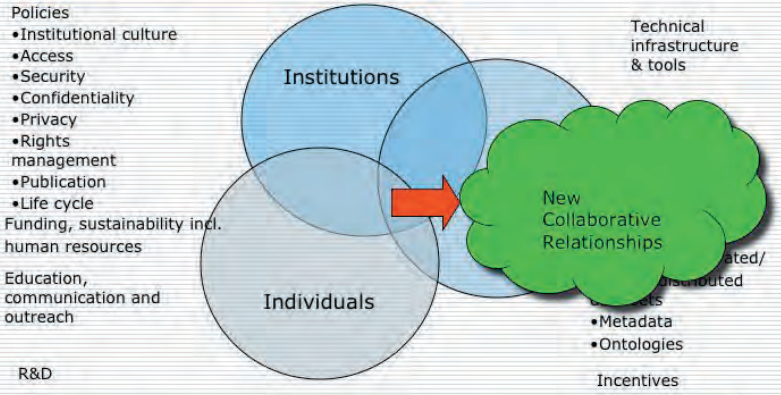
- Authority
- Responsibility
- Accountability
- Resources

Test: Do they align?

Evolving partnerships in a public policy framework:

Who does what to/for whom?

[Authority, responsibility, accountability, and resources]

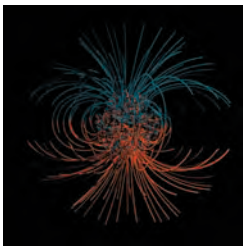


Charge to the Breakout Groups

- **Identify key issues**
- **Articulate well-formed, actionable recommendations with attention to implications beyond the core topic of the breakout groups (infrastructure, sustainable economic models, partnerships)**
- **Questions:**
  - Within the breakout topic, what are the most important challenges that must be met to create a reliable framework for preservation and access to digital data?
  - What are the critical steps necessary to meet those challenges?
  - What are the roles of each of the various sectors (e.g. academic, government, non-profit, commercial, international) in taking those critical steps?
  - What resources (funds, people, expertise, infrastructure, etc) are needed to take those steps?
  - What new technologies and new research are needed to take those steps?
  - What benefits would accrue from meeting the challenges and what are the costs of failure?

Infrastructure: Parting Observation

- Organization embodied in social, logical and physical structures
- Build it so that it is trustworthy
- That it does what you tell it to do
- That it doesn't do what you tell it not to do
- That it becomes a platform for great things



Appendix D. Breakout Session Reports

1. Infrastructure Breakout

   Wendy Pradt Lougee and Richard Luce

2. The Role of Academic Libraries in the Digital Data Universe

   Bob Hanisch and Brian Schottlaender

3. Summary of Economic Sustainability Models Breakout Session

   Chris Rusbridge and Fran Berman

Infrastructure Breakout

What capacities should we build now to manage data and migrate it over the future generations of technologies, standards, formats, and institutions?

Co-leads: Wendy Pradt Lougee & Richard Luce
NSF Digital Data Workshop
September 2006

Process

- Articulation of assumptions/principles for data infrastructure
- OAIS model: framework for exploring data infrastructure
- Recommendations

Assumptions

- Infrastructure = technology, people, instruments, data
- Tension: infrastructure requires a shared layer and needs to be discipline agnostic, while supporting discipline requirements
  - Culture prevails; must address incentives and process
  - Coordination problem - the universe will be split up
  - Both federated and central models; incentivize and leverage local investment
- Collaboration critical model
  - The role of library in the curatorial function, e.g., governance, standards, collection development, privacy, etc.
  - Imperative of stakeholder representation

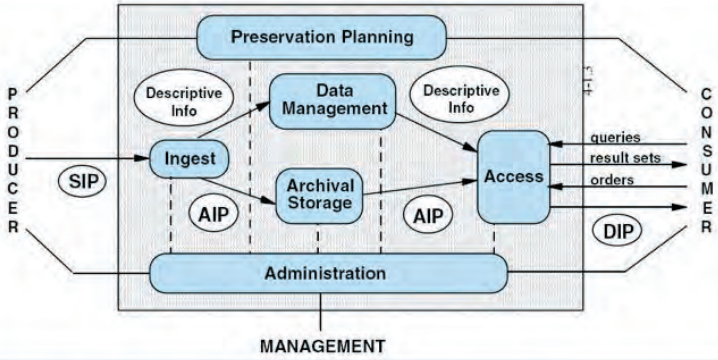
Assumptions

- Infrastructure is a shared common good - data is a public good
- Infrastructure must:
  - support multiple representations of data
  - Layering and building / reuse of components
  - Support appropriate security systems, privacy, confidentiality, and enable trusted relationships

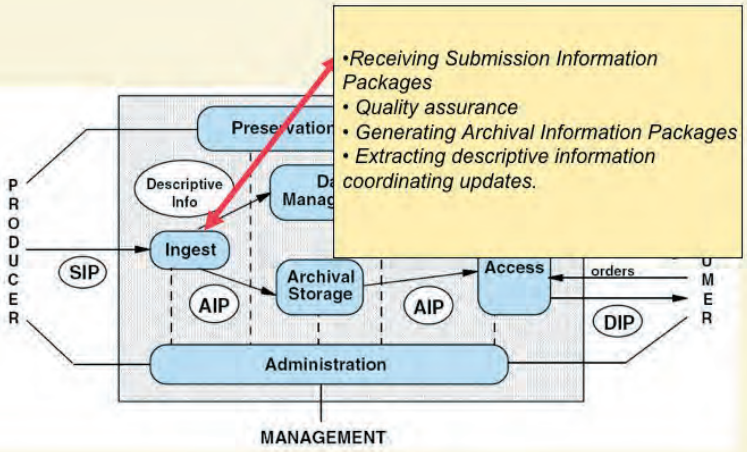
Structuring the Problem

- Functions (OAIS model):
  - Ingest & submission
  - Access
  - Archive
  - Management (preservation & storage)

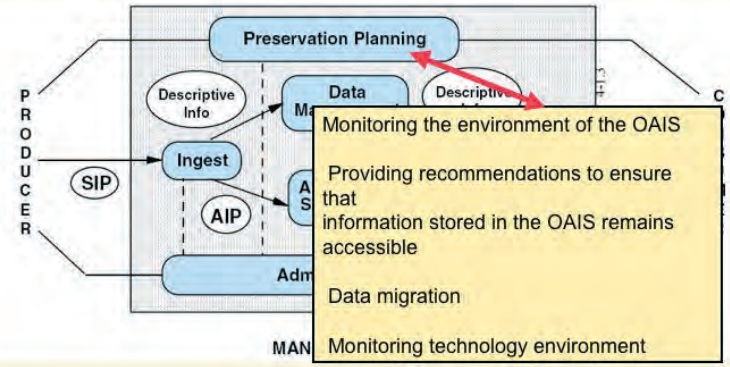
OAIS Reference Model

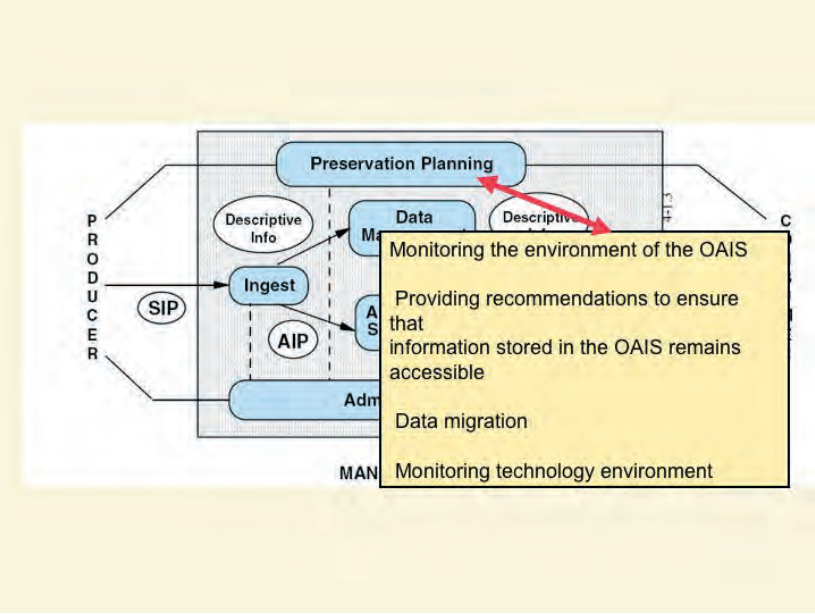


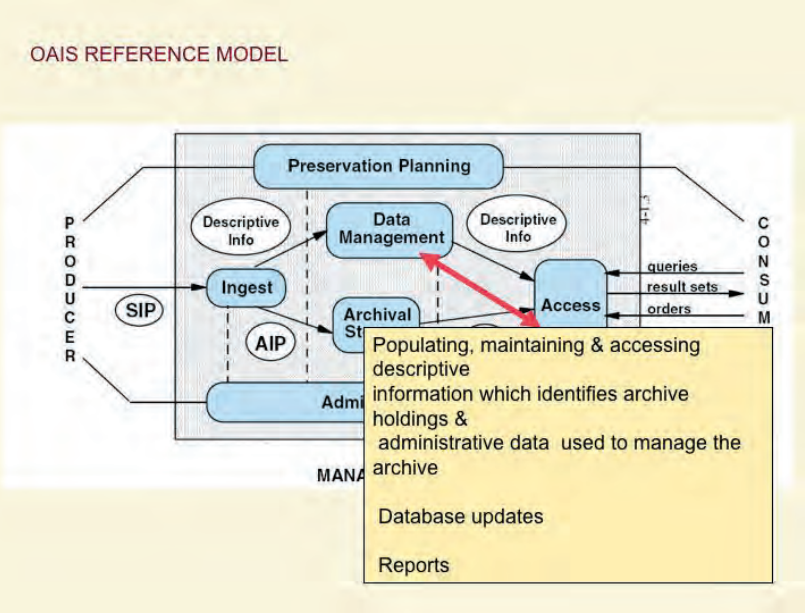
http://public.ccsds.org/publications/archive/650x0b1.pdf

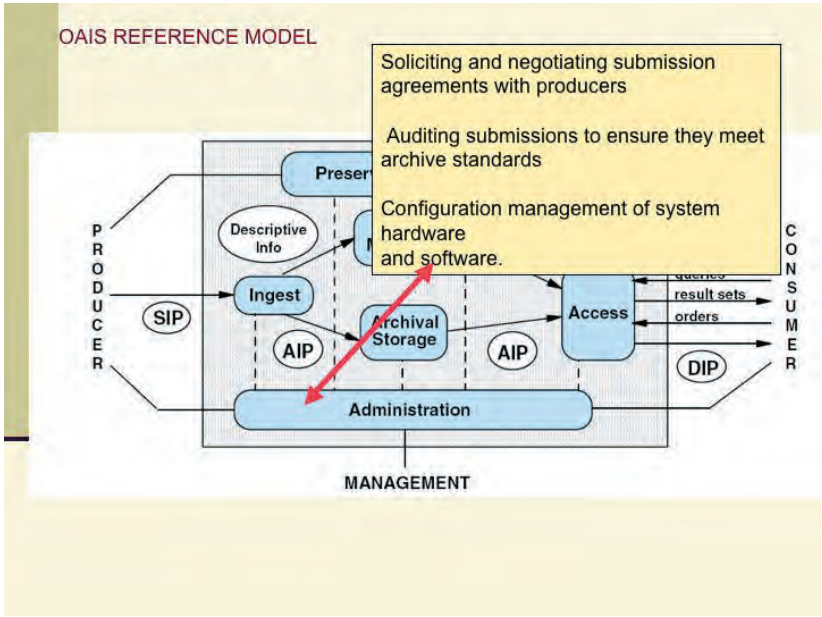


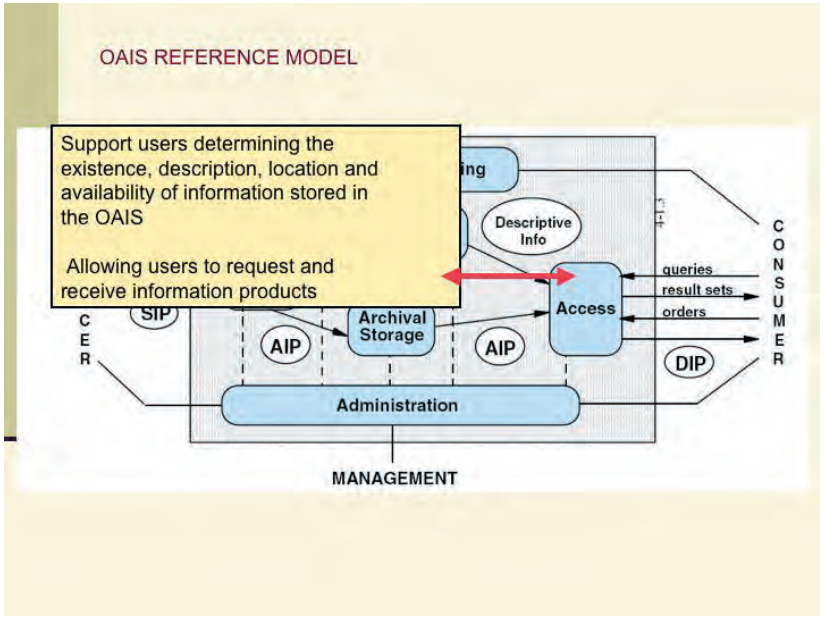
http://public.ccsds.org/publications/archive/650x0b1.pdf




Preservation Planning
Descriptive Info
Data
Ingest
SIP
AIP
PRODUCER
Monitoring the environment of the OAIS
Providing recommendations to ensure that
information stored in the OAIS remains accessible
Data migration
Monitoring technology environment


OAIS REFERENCE MODEL
Preservation Planning
Descriptive Info
Data Management
Descriptive Info
Ingest
SIP
AIP
Archival
Access
queries
result sets
orders
PRODUCER
CONSUM
Populating, maintaining & accessing descriptive
information which identifies archive holdings &
administrative data used to manage the archive
Database updates
Reports


OAIS REFERENCE MODEL
Soliciting and negotiating submission agreements with producers
Auditing submissions to ensure they meet archive standards
Configuration management of system hardware
and software.
Descriptive Info
Ingest
SIP
AIP
Archival Storage
Access
result sets
orders
DIP
Administration
MANAGEMENT
PRODUCER
CONSUMER


OAIS REFERENCE MODEL
Support users determining the existence, description, location and availability of information stored in the OAIS
Allowing users to request and receive information products
Descriptive Info
SIP
AIP
Archival Storage
Access
queries
result sets
orders
DIP
Administration
MANAGEMENT

Capacity and Models

Support creation of *collaboratories* that model essential infrastructure for data-enabled research.

- Scaled proposals: small, medium, large
- Collaborations among stakeholders
- Instantiate data models, technical and organization architectures
- Discipline-based or cross-disciplinary
- Incorporate sustainability plan

Policy

Create policy to ensure contribution of research data as shared research asset, enabling reuse in new research contexts - shared public goods concept

- Incentives for researchers to contribute data to collaborative environments (deposit)
- Support structures and training
  - Encourage archive-ready data and objects with appropriate metadata and formats

Research

Invest in problem-based research that will fuel collaborative data environments.

- Interoperable data models
- Specification of IP and access rights
- Security and trust
- Collaboration tools
- Cross-disciplinary discovery
- Automatic generation of metadata

Education

Stimulate the development of expert data curators and informed scientific community.

- Partner with IMLS to support new programs in data science/curation.
- Develop scientist capacity and culture to contribute to data collaboratives.
- Next generation scientist and information specialist

The Role of Academic Libraries in the Digital Data Universe

Break-Out Session:
New Partnership Models

Bob Hanisch and Brian Schottlaender
Co-Leaders

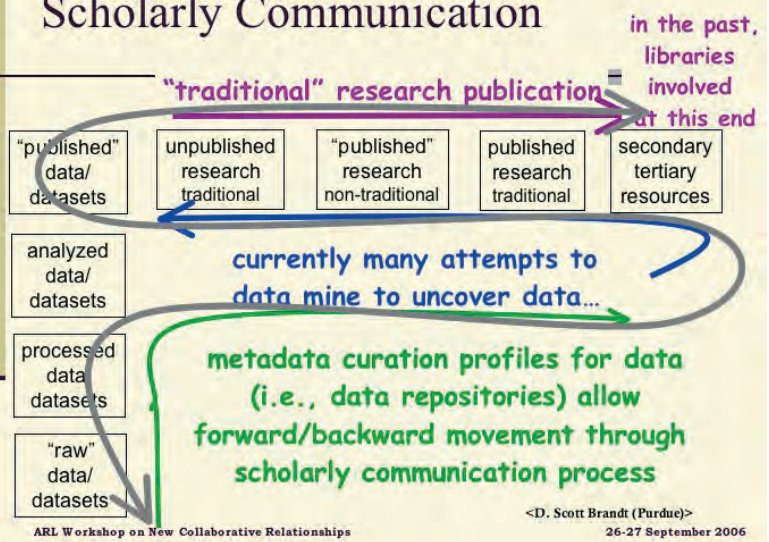
Objective

- Develop framework for collective action
- What should we do?
- When?
- Who should do it?
- Where?

Challenges

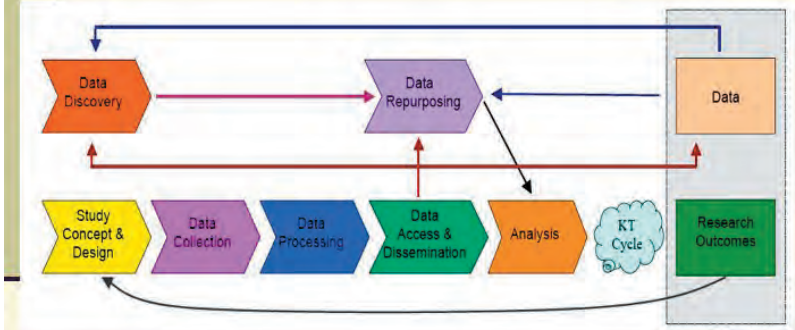
- **Overarching**: Crafting partnerships in which the focus is on long-term data stewardship
  - Current partnerships tend to focus on interoperability and integrated access, but lack a long-term component
  - Requires a different kind of institutional commitment and different funding strategies
- **Conundrum**: Short-term, and pressing, need
- Overlapping spheres of responsibility along the research process chain
- Credibility building
- Capacity building
- Changing roles of libraries, and changing perceptions of the roles of libraries (both in- and outside libraries)
  - Current emphasis in libraries is on information discovery, rather than information management (including storage)
  - Libraries need to re-think the partnerships into which they enter (including partnerships with other libraries)
- Definition of the functions comprehended by the word "curation
  - Curation and preservation are not the same thing
  - Preservation is a necessary condition for curation, but not a sufficient one
  - A lot gets preserved (and should) that is not immediately curated
- **Overarching**: Identifying where in the research process chain—or, where in the life-cycle of data—curatorial/preservation activities need to take place
  - Where do partnerships come into play
  - Where are the hand-offs?
  - How do we lower the barriers to participation?

Scholarly Communication



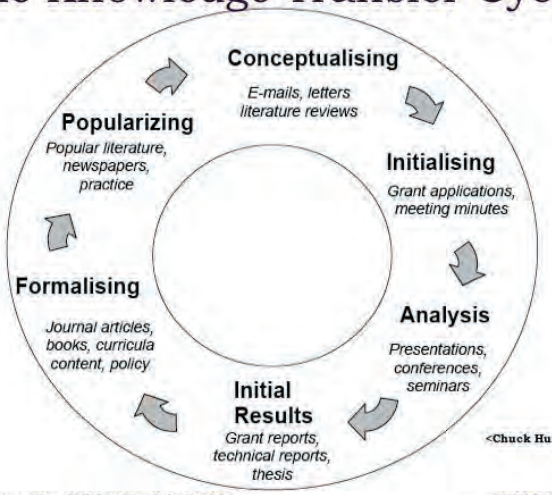
<D. Scott Brandt (Purdue)>

The Life Cycle of Research



<Chuck Humphrey (Alberta)>

The Knowledge Transfer Cycle



<Chuck Humphrey (Alberta)>

Stipulations

- Just as the problem space is a distributed one, so too will the solution space be distributed.
- Long-term stewardship is not about saving bytes; it's about creating, building, and evolving expertise in the community.
- There are multiple players/responsible parties in the problem (and solution) space, who have varying levels of understanding of and interest in the issues:
  - Universities
  - Libraries and librarians (lower-case "l")
  - Domain specialists
  - Computer scientists
  - Standards-setting bodies
  - Editors
  - Professional societies
  - Publishers
  - Commercial and not-for-profit vendors
  - Funding agencies
- "It takes a research community to preserve its data." [emphasis supplied]

Steps to Address Challenges

- Raise awareness and create demand in the research community
- Understand and define the requirements for repositories
  - Granularity
  - Metadata
  - Etc.
- Distribute curation responsibilities across a body of responsible parties roughly equivalent in magnitude (i.e., size, capacity) to the magnitude of the collective data store in need of curation
- Ensure that the work environments of those responsible parties are well supplied with curatorial tools that facilitate their carrying out their responsibilities
- Prototype and test
- Deploy and measure
- Develop business models

Recommendations

- **Overarching: NSF should facilitate the establishment of a sustainable institutional framework for long-term data stewardship involving the players enumerated above. This framework must:**
  - Encourage the articulation what, exactly, constitutes "curation" in various disciplines.
  - Encourage a diversity of designs and approaches that are sympathetic to the needs, practices, and relationships within affected research communities. One size does not fit all.
  - Encourage the development of distributed partnerships between libraries and research institutions.
- NSF should fund pilot projects/case studies that demonstrate the intersections between libraries, a limited number of scientific/research domains, and extant technologies bases.
- NSF should fund projects in which university research libraries develop deep archives of irreplaceable data, assuring descriptions of these data at a minimal level (floor, not ceiling) and facilitating discovery and access to these data, according to prevailing community standards.

[N.B. in re: 2 and 3 above: It will be important/valuable to find the right balance between prototypes and longer-term commitments.]

ARL Workshop on New Collaborative Relationships 26-27 September 2006

Recommendations

- NSF should require that data management plans submitted as part of the application process identify the players involved in the custodial care of data for the whole of its life cycle, and should support training initiatives to ensure that the research community can fulfill this requirement.
- NSF should foster the training and development of a new workforce in data science
  - Promote new curricula
  - Develop new programs
  - Link to training of domain scientists and information/library scientists
- NSF should partner with IMLS to train information and library professionals (extant and future) to work more credibly and knowledgably on data curation as members of research teams

ARL Workshop on New Collaborative Relationships 26-27 September 2006

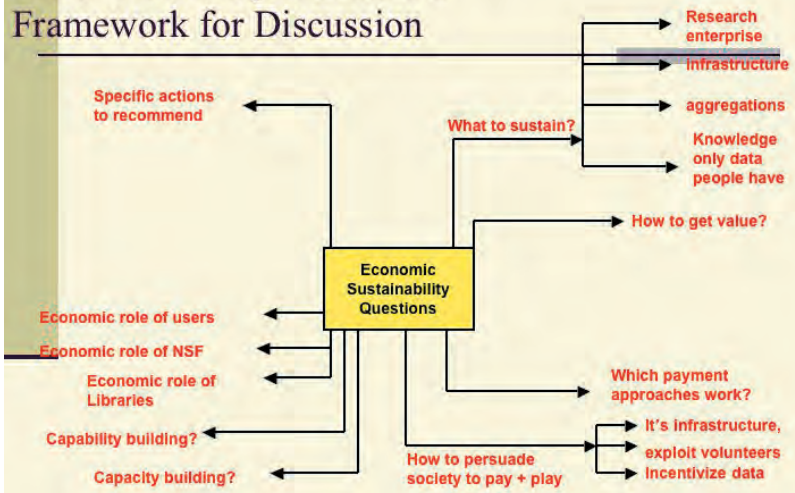
Summary of Economic Sustainability Models Breakout Session

Chris R. and Fran, Co-leads
Julia and Heather, Scribes

Focus: Economic Sustainability

What models are required to sustain data management and preservation efforts over the long term?

Chris' Mindmap Provided a Framework for Discussion



What economic models are relevant?

- The group began the discussion by describing the economic models which support their current activities.
- During the discussion, we discussed a spectrum of traditional and non-traditional related economic models including
  - **ICPSR** (subscription, user fees, federal, private funding)
  - **The Mormon Church** (tithing, user fees, sales)
  - **PBS** (donations, federal, state?, volunteers [donated time, expertise], sales)
  - Volunteer activity (archiving @ home)
  - **Markets** (DRI, data "futures", shares, etc.)
  - **Hybrid** (federal+state, public+private, etc.)

A thought experiment: Abstracting ICPSR

- **What has made ICPSR successful as a model?**
  - Robust environment with low barrier to access
  - Content which is of great value
  - Business model and structure which reflect the culture of the domain and constituent users
  - Useful tools associated with data
  - Trusted repository

Key to start from state-of-the-art rather than to reinvent the wheel

- **Economic sustainability models should utilize existing theory and practice as a foundation** – critical to have economists and sustainable infrastructure expertise in the discussion.
  - This is **symptomatic of a more general problem** – we shouldn't reinvent the wheel in economics, business, archiving, etc. Rather we should use the existing knowledge and experience base as a starting off point
  - This will mean the need for venues for more in-depth cross-cultural discussion and projects to help educate communities
- **Preservation will require both research into new viable models, and experimentation with new ideas**
  - Five years is short for an experiment, 5+5 is better
  - Risk taking: failure is an option!

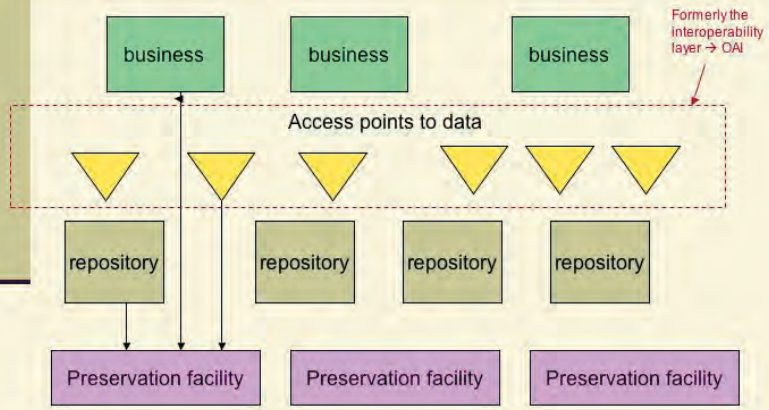
Many "building block" issues

- How do researchers and librarians sort through the **legal and policy issues** regarding ownership, use, confidentiality, privacy, liability etc.?
- What is the **minimal level of service** that makes data preservation worthwhile?
- What is the **cost of not keeping data**? When is it productive to re-compute, replicate experiments, re-do?
- What is the **data version of the "Earth Simulator"**? (i.e. what is the newsworthy item that will get U.S. competitive juices flowing and help generate new funding for data management and preservation)

Interesting Issues

- Large projects doing a reasonable job of putting data on the radar. **Small projects are the most at risk**.
- **Most libraries do not currently host substantive research data** – both library and research community need more experience with one another's cultures. Is there a way NSF can help foster greater engagement?
- **Good infrastructure must have a plan for "the end"** – how do we reappraise if necessary, how do we hand-off, how do we become self-sufficient?

Eric's Updated Version of the Cliff Lynch model



"Actionable" Recommendations 1

"We don't get anywhere if we don't start somewhere."

1. **Involve economics and social science experts** in developing economic models for sustainable data preservation – research should ultimately generate models which could be tested in practice.
2. **Set up multiple repositories and treat them as experiments**
   - Require that repository experiments develop plans to address key issues such as transition between media/formats/institutions, self-sustainability, exit strategy, etc.
3. **Develop usable and useful tools for automated services and standards** which make it easier to understand and manipulate data. Develop incentives to encourage community use. Invisible metadata creation!

"Actionable" Recommendations 2

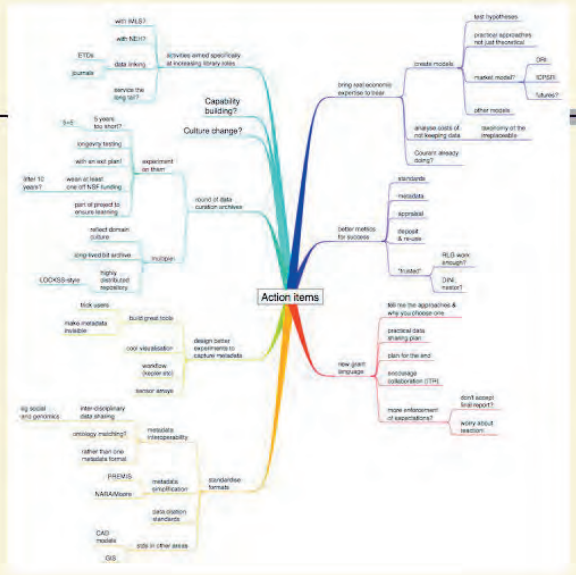
4. **Require data sharing plan in proposals** that has practical value (and appropriate support). Plans for resource and reference data should contribute to community data stewardship
5. **Create and enforce data sharing policies among NSF awardees** (e.g. final report not accepted unless awardee is compliant with stated data management plan)
6. **Use NSF program process to help the library community take more responsibility** for the stewardship of research data (with other funders?)

"Actionable" Recommendations 3

7. **Use NSF program process to change culture** in research community
8. Undertake capacity & capability building activities

A Bolder Vision? Remember Dli!

- US Digital Curation Initiative
  - A major, inter-disciplinary, cross-directorate, inter-agency program, with options built in for international collaboration (UK, EU, Australia at least), that will both experiment on models and build sustainable curation services!



Bring real economic expertise to bear

- Create models
  - test hypotheses
  - practical approaches not just theoretical
  - market model?
    - DRI, ICPSR, futures?
  - other models
- Analyse costs of not keeping data
  - taxonomy of the irreplaceable
- Courant already doing?

Design better metrics for success

- Standards
- Metadata
- Appraisal
- Deposit & re-use
- "trusted"
  - RLG work enough?
  - DINI, nestor?

New grant language

- Practical data sharing plan
  - tell me the approaches & why you choose one
- Plan for the end
- Encourage collaboration (cf ITR)
- More enforcement of expectations?
  - don't accept final report?
  - worry about reaction!

Work to standardise formats

- Metadata interoperability
  - inter-disciplinary data sharing
    - eg social and genomics
- Preservation metadata simplification
  - PREMIS
  - NARA/Moore
- data citation standards & promotion
  - To change culture
- Support standards in other areas
  - CAD models, GIS, etc

Design better tools & experiments

- Sensor arrays & other experimental engineering to capture metadata
- Build great tools
  - Robust, reliable, useful, usable
  - "trick users", make metadata generation invisible
    - Eg cool visualisation, workflow (Kepler etc)

New round of data curation archives

- Multiple!
  - reflect domain cultures
- Experiment on them (take risks)
  - 5 years too short- 5+5?
  - with an exit plan!
  - longevity testing
  - part of project to ensure learning
  - wean at least one off NSF funding
    - after 10 years?

New round of data curation archives

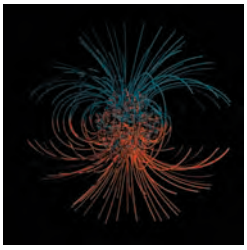
- Specific possibilities
  - long-lived bit archive
  - Build on IR work
  - highly distributed repositories
    - LOCKSS-style
    - Archiving @ home

Activities aimed specifically at increasing library roles

- Libraries have significant opportunities to extend their roles in info discovery, archiving etc in research data
- Partnerships with domain researchers
- Forum for outreach & scientific communication
- Data linking from
  - ETDs
  - journals
- Capacity building opportunities in library educaiton etc with NEH and/or IMLS?

Other things

- Capability building
  - Education: librarians, data scientists, researchers
- Capacity building
- Culture change?



Appendix E. Position Papers

1. Henry E. Brady
2. Suzanne Carbotte
3. Robert S. Chen
4. Sayeed Choudhury, Robert Hanisch, and Alex Szalay
5. Paul Constantine
6. Peter Cornillon
7. Bernard Dumouchel and Richard Akerman
8. Stephanie Hampton, M. Jones, and M. Schildhauer
9. Margaret Hedstrom
10. Charles Humphrey
11. John Leslie King
12. Rick Luce
13. Barbara Lust and Janey McCue
14. James L. Mullins
15. James D. Myers
16. Frank Rack
17. Mark Sandler
18. MacKenzie Smith
19. Eric F. Van de Velde
20. Todd Vision
21. Tyler O. Walters

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Henry E. Brady, Professor of Political Science and Public Policy, Director of the Survey Research Center and UC DATA, University of California, Berkeley

There are many problems confronting efforts to preserve and manage scientific data in digital form including making decisions about what to keep, developing plans for discarding data when it is no longer useful, providing adequate meta-data, ensuring long term preservation given frequent changes in media and software, and finding and training staff to do these tasks. I will focus, however, on three problems that are especially pertinent to the social sciences.

Linkage of Data Sets. The social sciences are benefiting enormously from the easy availability of large-scale, computerized datasets such as vital statistics, census data, employment records, educational data, welfare and social security records, voting data, medical records, commuting and transportation data, and many other kinds of information. These datasets are even more useful when they can be easily linked together to study events or transitions such as the transition from welfare to work, from illness to a job, from school to citizen, from prison to everyday life, or from home to work. Coding of geographic, contextual, genetic, environmental, and other information can make these data even more valuable for understanding the impact of neighborhoods, institutions, physical distance, individual characteristics, and other factors. Yet these data come in many different forms (different types of databases, different units of observation, and various levels of reliability), and linking them poses significant challenges. If data libraries are to be truly useful for social sciences, they must provide users with the software tools to link these very-large and unwieldy data-sets easily, reproducibly, and reliably.

Confidentiality. Although the availability and the linkage of social science data provide tremendous opportunities for answering important social science questions, they also exponentially increase the dangers of disclosing personal information through the possibility of identifying individuals – even though social science researchers are almost always interested in general statements about behavior and almost never interested in individuals. The problem of confidentiality has increased the requirements for Human Subject Reviews, decreased the availability of many kinds of data, and made linkage especially suspect. A number of technical and institutional methods are being developed to deal with these problems, but we are still far from having generally accepted approaches to them. Moreover, although confidentiality has been an especially difficult problem for the social sciences, it is increasingly a problem for the medical, environmental, and even the geo-sciences.

Institutional Models. One answer to the problems of linkage and confidentiality is to develop better institutional models that provide ways that researchers can have access to data in ways that protect individuals while allowing for extensive data linkage. One example is the Census Research Data Centers which allow researchers to access non-public Census data under rigidly controlled conditions. This model, however, only allows for access on a case-by-case basis,

and it does not currently allow for long-term access to data. Institutions are also important for a larger reason: At the moment, we have nothing comparable to the "University Library" which has historically made rational acquisitions through "collection specialists" working with researchers, developed meta-data through classification and indexing, and paid for the development of documentation and the preservation of information. Thus, researchers with data typically do not have any place to go on the University campus, and even if there is an institution concerned with digital social science data, it is typically woefully under-funded and unable to help the researcher archive and preserve data. Some libraries and some computer centers have begun to take up these challenges, but each has other responsibilities and agendas which impede their efforts.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Suzanne Carbotte, Program Director, Marine Geoscience Data System, Lamont-Doherty Earth Observatory, Columbia University

How do we ensure the engagement of data producers at various stages in the data life cycle. To develop comprehensive digital data resources of maximum use to the research and education community requires the active involvement and engagement of individual scientists who are essential data producers throughout the data life cycle. A diverse range of individual scientists may be involved at various stages throughout the data lifecycle as field data are processed, reprocessed for new applications, integrated into data syntheses, and higher level derived data products are developed. It must be easy (transparent?) for scientists to document and contribute their data products, scientists need to have their contributions adequately protected and acknowledged, and new rewards for contribution are needed.

Inadequate enforcement of data policies. New governance structures for enforcement of data policies are needed. In some realms of scientific research, existing NSF data policies have been difficult to enforce, partly because appropriate digital data repositories have not existed, but also because mechanisms to document compliance are not in place. Another aspect of the problem is that the scientist may be the only one who knows of the existence of data and compliance of the individual scientist with a data policy must be based on trust and commitment to data preservation as part of the scientific process.

Inadequate incentives for scientist participation in data preservation. New incentives for scientists to contribute to data collections go hand in hand with the need to fully engage data producers in the data preservation process and the need for new structures for enforcing data policies. We need to change how we reward and credit scientists for data contribution. Contribution to databases needs to be part of the publication process, and we need a new system of professional recognition that acknowledges the value of data production and contribution to data collections. New partnerships with the academic journals will be needed to develop policies for publication, which include linking publications to digital data resources.

How do we ensure the long-term security of digital data collections in an uncertain funding climate? New scenarios and partnerships to ensure long-term funding are essential for both the development and security of digital data collections. To adequately manage and preserve the complex heterogeneous data that are produced in an increasingly multidisciplinary research environment requires data managers with a high level of expertise in both the domain sciences as well as information technology. Such people are difficult to find and keep in an uncertain and short term funding climate.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Format

Comments by Dr. Robert S. Chen, Interim Director, CIESIN at Columbia University; Manager of the NASA Socioeconomic Data and Applications Center (SEDAC); and Secretary General, Committee on Data for Science and Technology (CODATA) of the International Council for Science (ICSU)

Digital Rights Management. Clarification of rights to archive, use, and disseminate data and any applicable restrictions is essential to long-term data curation, and would be greatly facilitated by digital standards and technologies that permit data sources and users to quickly and easily specify and understand rights and restrictions in ways that meet their needs and concerns. This has been a key element in the success of the Creative Commons and needs active support in the realm of scientific data and information. Key issues that need to be addressed include protection of confidentiality, limitations on liability of data sources, use of data for humanitarian purposes, and definitions of appropriate uses (e.g., private sector vs. public sector research).

New Institutional Partnerships. A range of new partnerships is needed across disciplines, within and between universities, between sponsors and data managers, across the public and private sectors, and with the broader scientific community to establish appropriate and sustainable long-term data management structures covering all or most of science. Is a mix of disciplinary, cross-disciplinary, and institutional repositories going to evolve that can provide sustainable data curation and management in most fields? Can we identify gaps and find ways to fill them? Are there contingency plans when a particular institutional arrangement for a particular field encounters problems with sustainability? Are there ways to involve the private sector and/or the open source community in these arrangements to help infuse interoperable technologies and reduce costs without risking long-term sustainability or access? Should existing consortia, e.g., of universities, of libraries, or disciplinary data centers, be asked to take on long-term curation responsibilities or are new ones needed?

Science Education and Community Outreach. Many scientists continue to use traditional approaches to data, i.e., developing custom datasets for their own use with little attention to long-term reuse, dissemination, and curation. Although there has been considerable progress in data stewardship for "big science" projects, even modest collaborative projects are inconsistent in their attention to data management and few individual scientists think beyond posting selected results and data on the Internet or submitting a final data product to a data archive if required to do so. Changing this sort of behavior will require a range of efforts, including investment in approaches to make data documentation, sharing, and preservation easier, establishment of an infrastructure to accept and assume responsibility for data (e.g., a local university depository or a disciplinary data center), and, perhaps most important of all, concerted efforts to educate current and future scientists to adopt better practices.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Sayeed Choudhury, Associate Director for Library Digital Programs, Johns Hopkins University; Robert Hanisch, Senior Scientist, Space Telescope Science Institute; and Alex Szalay, Alumni Centennial Professor, Department of Physics and Astronomy, Johns Hopkins University

The National Science Foundation's Cyberinfrastructure Vision for 21st Century Discovery describes a broad-based effort that is transforming the manner in which scientists, social scientists and engineers (and perhaps even the humanists) conduct research, teach and learn, and disseminate their research findings and publications. Projects such as the Virtual Observatory (VO) provide ample evidence of data-driven scholarship, which offers both challenges and opportunities for academic research libraries, especially in the realm of data curation. Even with ambitious new efforts to create large corpora of digitized text such as Google Book Search or the Open Content Alliance, libraries still represent a core element of the preservation picture.

Given the scale and complexity of even a single cyberinfrastructure-based project such as the VO, it is not reasonable to assume that a single library or organization can manage the entire range of data curation needs. Rather, libraries must find ways to work together with an array of organizations such as other libraries, supercomputing centers, museums, archives, publishers, and corporations. For different stages and applications of data, various organizations will need to identify appropriate roles and develop systems that interface—technically and organizationally—with a range of partner institutions. Such a complex array of relationships and technological infrastructure may benefit from examination and leadership from the highest levels of the academic and corporate community. However, at this stage, there are major research and development questions that remain unaddressed. In this current environment, it's essential to develop prototype systems that demonstrate both technical and organizational infrastructure to support data curation. These prototype development efforts will help us better understand appropriate technologies, potential costs, and organizational relationships that will be necessary to support cyberinfrastructure-based projects and programs.

At Johns Hopkins University, we are working with a network of libraries, publishers, scholarly societies, and corporate partners to develop a repository-based system that will support an end-to-end process for capturing, curating, preserving, and providing access for the long term to derived data that is cited in electronic publications. We are prototyping such a process and system. Our goals include assessing the scientific impact of this new approach to astronomical data, as well as working out sustainable business models for increasing the value of data in this way. Our prototype phase focuses on astronomy because of the technological maturity of electronic publications and data management in this discipline, and because of the wide access to digital data archives, and the unique, established relationship between the astronomers and libraries at Johns Hopkins University.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Paul Constantine, Associate Dean for Research and Instructional Services, University of Washington Libraries

Within the context of New Partnerships here are what I see as key issues facing us.

Long-term Preservation

Appraisal

- There is simply too much data to preserve everything over the long term; we need to devise ways to determine exactly what data is worthy of long-term preservation.
- Can we develop an understanding of how data might be re-purposed in ways totally unpredicted by its original creator/gatherer/researcher?
- What legal/ethical constraints encourage or discourage the long-term preservation of particular datasets?

Management

Who is best positioned to manage long term preservation?

- Scientists are often ready to move onto to their next project and don't necessarily see the value/need or have the resources to preserve their data over time.
- Data sitting on a scholar's computer are not easily accessible to other researchers wishing to use and/or repurpose it.
- Campus Computer Centers are not always funded or equipped to preserve data for the long-term.
- Neither researchers nor computer centers are especially used to creating metadata schema and assigning metadata in ways to make datasets more easily discoverable.
- Libraries, while skilled at creating metadata schema and assigning metadata in ways to make datasets more easily discoverable, are generally not funded or equipped to preserve digital data for the long-term.

Curation of Scientific Data in Digital Form

I see curation in many ways as the overall "thing." Long-term preservation and managements are components of data curation.

Funding

- Many libraries and computer centers are not funded to provide data curation

Participation

- Some researchers need to be convinced of the importance of data curation and the long-term preservation of their datasets.
- Convincing researchers to share their datasets.

Intellectual Property Issues

- Have the data been copyrights or their use patented?
- Are the data licensable?
- Who owns the data?
- What restrictions have they imposed?

Preservation and Access

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Academic Libraries in the Digital Data Universe: The Reference Desk and Technical Memoranda

Comments by Peter Cornillon, Professor of Oceanography, Graduate School of Oceanography, University of Rhode Island

There are many issues that relate to the role of libraries in a digital age. The two that I have been concerned about for a long time are:

The Reference Desk. Finding data of interest[1] and then accessing these data can be extremely difficult. It is often difficult to know just where to start and there appears to be a widening gap in the expertise of the reference librarian in this regard and the state of the technology. For example, there exist today a number of high level directories that will help one find data sets of potential interest—the Global Change Master Directory (GCMD) for the Earth sciences, the National Space Science Data Center (NSSDC) Master Catalog for the space sciences, etc.—but few reference librarians are aware of these or how to use them. In fact, the expertise in data discovery has moved away from the library to the researcher. This is both a problem and an opportunity. It's a problem for the new researcher or student who is looking for data; they have to go through the same, often painful, discovery process as all of their colleagues. It's an opportunity in that expertise exists at many institutions to help train new librarians in these areas and to retrain librarians already on the job. Unfortunately, this is an opportunity that is not been exploited. NSF might investigate the funding of courses at library schools that draw on the data discovery talents of the local research community. This course could be offered both as a recertification opportunity for reference librarians as well as basic instruction for students in library science programs. The course could also address data access methodologies.

In addition to the research community benefitting from more expertise in the library with regard to data discovery, there would also be a direct benefit to those developing data discovery and access methods from more input from the library community. Bottom line: there are a number of benefits that would derive from a tighter coupling of the research and library communities as relates to data discovery and access.

Technical Memoranda—Gray Literature. Although universities have taken the lead in the development of end-to-end data systems in a highly distributed environment, there is one area in which they have taken a significant step backward. In the past researchers often "published" their data in paper form as *technica*l *memoranda* or some equivalent and these reports were (and still are) archived in the university's library. With the advent of the Web such technical reports have all but disappeared with researchers "publishing" their data on personal Web sites; i.e., the institutional commitment to a long-term archive of the data has all but disappeared. This is a trend that must be reversed or much of these data will be lost forever – universities must provide a mechanism for researchers to "publish" their data electronically for permanent archival in the university library.

Acronyms
GCMD Global Change Master Directory
GSO Graduate School of Oceanography
NSF National Science Foundation
NSSDC National Space Science Data Center
URI University of Rhode Island

[1]In this position paper, my references to finding and accessing data refer to finding and accessing digital data on the Web; i.e., finding and accessing remote repositories of data.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Bernard Dumouchel, Director General, Canada Institute for Scientific and Technical Information (CISTI), and Richard Akerman, Technology Architect-IT, Architecture, CISTI

We feel that there is a role for CISTI in promoting the use and stewardship of Canadian research data, in order to maintain the position of Canadian science. As well, we foresee a potential role in coordinating open data activities, similar to the UK Digital Curation Centre (DCC) model. We would like to explore new partnerships in this area. We feel that there are many possibilities for enhanced use of and access to data, such as wider and richer linking of data to publications.

The Internet is enabling much greater openness in several areas:

Open access is partly about publication funding models (which will not be discussed herein), but also importantly about providing free, public access to information.

Open data is generally less contentious that open access, as the funding model aspect is less important. There is general agreement even amongst publishers that scientific data should be open to all (with appropriate privacy and other constraints).

Open discourse is about broadening the scientific discussion beyond the confines of traditional venues. Without the constraints of a printed page or a conference session, rich discussion is possible both amongst scientists and between scientists and the general public.

Libraries are well positioned for these trends, which could be considered part of the development of **Open Science**. Research libraries can play a role in the promotion and understanding of all of these areas, as well as potentially providing infrastructure or coordination.

In terms of the infrastructure aspects, to some extent these are already being addressed by existing e-Science or cyberinfrastructure programs, although they have a focus more on computing and storage resources for researchers.

The Canada Institute for Scientific and Technical Information (CISTI) has a long history of participation in the realms of scientific computation and scientific data. In particular, we have a longstanding role in cataloguing data and promoting its use. CISTI hosts the Canadian secretariat of CODATA, participates in ICSTI, and was involved with the production of a report on Canadian access to scientific research data (the NCASRD report). We provide a Depository of Unpublished Data and our Research Press journals support the concept of supplementary data.

There are a number of issues that need to be discussed: Can the concept of trusted digital

repository be extended to data repositories? What additional management elements and criteria will be needed? How can we deal with existing scientific data, which may be well managed but not necessarily compliant with any repository or access standards? How shall data be catalogued and identified, particularly across scientific fields? How should data sets be cited, how can versions be handled as data sets grow and are updated?

We look forward to discussing these and other issues as we explore the digital data universe together.

References

Digital Curation Centre
http://www.dcc.ac.uk/

Canadian National Committee for CODATA
http://www.codata.org/canada/tofr.shtml

Report on Data Activities in Canada
http://dac.cisti.nrc.ca/datact_e.cfm

Canadian National Consultation on Access to Scientific Research Data (NCASRD)
http://ncasrd-cnadrs.scitech.gc.ca/

Depository of Unpublished Data
http://cisti-icist.nrc-cnrc.gc.ca/irm/unpub_e.html

Research Press – Supplementary Data
http://pubs.nrc-cnrc.gc.ca/rp/rptemp/rp2_news2_e.html

Networks: recipe for success in the knowledge age
http://cisti-icist.nrc-cnrc.gc.ca/media/news/cn20n3_e.html#a0

open discourse + access + data equals open science?
http://scilib.typepad.com/science_library_pad/2006/09/open_discourse_.html

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Managing Collections of Highly Dispersed, Hererogeneous Data

Comments by S. Hampton, M. Jones, and M. Schildhauer, National Center for Ecological Analysis and Synthesis (NCEAS), University of California, Santa Barbara

Libraries have historically played a critical role in the long-term preservation of scholarly works, especially books and other artifacts written in natural language. Curation of scientific data has challenged the capabilities of existing systems because of the unique characteristics of scientific data and the special services which must accompany archiving of scientific data. Unlike books and similar publications, scientific data are generally intended for analysis, modeling, and visualization rather than reading and browsing. Analysis and modeling activities often require the quantitative integration of multiple data sets from dispersed locations that vary tremendously in their structure and semantics, which creates a need for much more detailed metadata than is available in traditional library usage. Depending on the discipline, scientific data can be small and complex, requiring substantial documentation to accurately interpret the data, or large but uniformly structured, requiring less documentation but presenting system scalability issues. These and other fundamental differences in the way one uses scientific data lead to the need for new partnerships that can effectively provide for simultaneous preservation, discovery, access, integration, and analysis of data.

Heterogeneity and Dispersion. Dealing with heterogeneity and dispersion in the context of integration, analysis, and modeling is the major key to successfully building data collections. Disciplines that are strongly bounded along lines of similar information (such as genetics and proteomics) can have highly integrated solutions like GenBank or the Protein Data Bank, yet other fields (like Ecology) can vary widely in the types of information that are necessary within the context of a single study (e.g., a single ecological study may require data from population biology, genetics, hydrology, and meteorology). Consequently, one archival solution might not fit all disciplines, unless that solution provides interfaces that enable both breadth of coverage and depth of resolution within any given discipline. For example, traditional metadata systems used in libraries provide metadata that assists with discovery at a coarse-grained level, but understanding heterogeneous data requires detailed metadata that describes the structure, content, and semantics of data and the protocols used to generate the data. Although metadata standards overlap tremendously, discipline-specific extensions must be created to fully understand and utilize data. Data dispersion can play an important role in structuring collections. Local institutions are typically the best curators of scientific data because they best understand the data collection and quality assurance processes. Although specific versions of scientific data sets are static and can be preserved as-is, data users find errors and omissions that are fixed in subsequent versions, requiring an active curatorial system that dynamically links actions of local scientists to regional, national, and global archives. For libraries to provide an effective archival system for science data, they must build semantically rich data infrastructure

that allows direct access to heterogeneous data directly from within analysis and modeling systems used by scientists and that allows for curatorial linkages among data systems from local, regional, national, and global scales.

Long-term Preservation. Many disciplines, including ecology, lack a mechanism for assuring the long-term preservation of scientific data. Although some nationally scoped data archives exist (e.g., the NASA DAAC's, the NODC, etc), many of these are federally funded and are subject to the vagaries and cycles associated with public federal funding. These archive centers tend to focus on archiving data without fully dealing with the difficult and expensive aspects of long-term curation, including the creation and maintenance of new data versions, media migration, and software obsolescence. A partnership of libraries and data centers that each contains replicas of scientific data linked to local mechanisms for data curation and update would be far more durable over the long-term than single, centralized data archive systems.

Data Sharing. Despite widespread agreement that sharing data is paramount to the scientific method and essential to synthetic advances that span scales and disciplines, institutional and individual sociological barriers to intellectual rights of data use remain a serious problem. Diverse approaches to preserving data that gradually migrate disciplines into more open and unquestioned sharing of data will benefit science but require new partnerships among scientists, data centers, libraries, scholarly societies, and universities. One approach is to provide incentives to data sharing directly to scientists, e.g., as the Kepler scientific workflow system has done by directly linking analysis and modeling capabilities to data archives and sensor networks. New partnerships that promote generic data access interfaces allow us to build integrated systems that scientists can use to access data archives during the course of analysis and modeling, thereby providing an incentive to share data.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Sustainable Economic Models

Comments by Margaret Hedstrom, Associate Professor, School of Information, University of Michigan

The Intensive Care Unit Analogy. As a society, we are alarmed at the rising cost of health care. Think of what health care would cost if all of the patients were in the Intensive Care Unit or the Emergency Room. What is the analogy to long-term preservation? All of our patients (that is all of the data that needs to be diagnosed and treated for the disease of decay) today are in the ICU or the ER. [This is a bit polemical, but I hope you get my point.] An affordable (e.g., economically viable) system for long-term preservation requires preventative medicine, a system of diagnosis of similar problems, treatment protocols and good practices, criteria for triage, processes and tools that support healthy data, and an infrastructure oriented to health of data rather than illness. Getting to this point will involve research (what are the good practices, which types of diseases affect which types of data, how do we motivate data producers to be "health conscious" about their data, etc. etc. especially in the absence of a known or quantifiable future demand.

The Value Proposition. If we take as a given that not all data are created equal and that we will not be able to afford to keep everything, how do we decide where to invest in preserving data. This is fundamentally an information problem. How do we make effective economic decisions in the face of uncertainty about the supply of data and the future demand? Are there any economic models or research issues that provide insights into comparable problems? What happens when the future value of a particular set of data is contingent upon its relationship to other data that have been preserved and can therefore be aggregated? At what level of granularity do we make selection (e.g. investment) decisions, given that deciding what to preserve is a very labor-intensive and expensive process.

Public Goods with an Unknown Future Value. I assume that there are numerous similar cases of public goods with an unknown future value, but how can we learn from these and make a similar case for long-term preservation of [the right] digital data? I think we could also leverage present value, but we need some good examples.

Disclaimer: I am not an economist, not do I play one on TV.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Charles Humphrey, Academic Director, Research Data Centre, and Head, Data Library, University of Alberta

Preserving scientific data is a process best characterized by a life cycle model that differentiates the various stages through which research is conducted.[1] The life cycle perspective helps visualize a global representation of this process and helps identify the digital objects produced throughout the full research cycle.[2] Within each stage are outputs in digital format that record or summarize research activities. For example, literature reviews, research prospectuses and grant applications are typical products of the Study Concept and Design stage that later become important sources when documenting data. Some of these products are specific to a particular stage while others are passed between stages. For example, a data file from the Data Processing stage will be passed to a subsequent Analysis stage. The life cycle model helps monitor both the digital objects bound within a stage and those objects that flow across stages.

This type of model also depicts the wide mix of implicit and explicit partnerships that occurs during research, including relationships among researchers, grant agencies, universities, data producers, scientific publishers, libraries, data repositories and others. New scientific research is stimulated by the intellectual capital, resources and infrastructure brought together through such partnerships and much of today's research is shaped by these interdependencies. The big picture from the life cycle model ensures that the combination of relationships within a project is recognized and well described.

Given a life cycle perspective, what are the key issues of long-term preservation, data management and the curation of data in the context of new partnerships, new organizational models and sustainable economic models?

New Partnerships. The preservation of scientific data is dependent on the custodial care of the digital objects produced throughout the research process. The traditional practice of gathering a paper trail of research outputs long after a scientific investigation has been concluded and depositing them with an archive is inapplicable in the digital era. Too much valuable research data are either at high risk of being lost or have been destroyed because of inappropriate practices that were carried over from a time when paper was the dominant medium. In the digital era, the challenge is to coordinate among partners the care of research data throughout the life cycle. Digital custodianship requires clearly articulated roles for the care of the digital objects, including which partner has primary responsibility in each stage of the life cycle.

New possibilities exist for librarians to serve as partners in the life cycle of research. Today, data librarians are on staff in many academic research libraries where collections of data are made available through library data services.[3] For the most part, data librarians are not engaged in primary research being conducted on their local campuses. Instead, they mainly

support researchers undertaking secondary data analysis. While this is an important service, the potential for data librarians to be much more involved in activities across the life cycle of research remains untapped. For example, data librarians could contribute significantly to the management of metadata throughout a research project.

Research data require high quality metadata for proper preservation and to be of value for re-use. New metadata standards are emerging that facilitate the discovery and repurposing of data.[4] The enhancement of such standards requires participation by all of the partners in the life cycle of research. With the production of comprehensive metadata based on open standards, new partnerships will be needed to develop open-source tools for mining this rich metadata. Research libraries should provide access both to the body of scientific literature and to the data upon which this literature is based. The publishers of scientific literature and the providers of library data services need to agree upon standard metadata elements that will facilitate the dynamic linking of data with scientific literature. With such metadata in place, new partnerships can be forged to develop the tools for integrating data with literature.

New Organizational Models. Short-term access to data is often best facilitated by keeping the data in close proximity to its origins. However, long-term access is completely dependent upon thorough preservation practices and standards. One challenge we face is to establish a network of organizations with varying levels of responsibility to span the life cycle of research. For example, a local digital repository may take initial responsibility for providing access to research data but the long-term preservation and access becomes the responsibility of a topical (discipline-specific) or general national repository. Coordinating the division of responsibilities across multiple digital repositories is a major organizational task.[5] A model based on a federation of data repositories[6] is one approach that would address the need for strong organizational coordination.

New data repositories of national prominence need to be launched that take on the long-term responsibilities of preserving data and that work closely in coordination with local repositories responsible for short-term access to data.[7] An open consultation is needed to determine how many of these repositories are required and whether their focus should be general or topical.

The emergence of local digital repositories with recognized responsibilities for the care of research products requires a certification process to ensure best practices and to build a level of trust between researchers and the providers of repository services. The work by the joint digital repository certification task force between the Research Libraries Group and the U.S. National Archives and Records Administration[8] has provided a framework for such a system. One or more organizational homes will be needed, however, to implement a certification process.

Sustainable Economic Models. One of today's most serious threats to science is the commodification of research data,[9] which includes the acts of selling research data at a cost in excess of the Bromley guideline and of inappropriately hoarding data under the pretext of intellectual ownership. Science flourishes in an environment of openness where ideas are

exchanged, challenged and tested. This principle of openness also applies to the data upon which research findings are based. The replication of research depends on an open exchange of data. The challenge we face as a scientific community is to find ways of preserving and exchanging research data that are not based on the commodification of research data. If we accept the premise that scientific research data must be a public good, how will the services to preserve and provide access to the data be financed?

[1] A position statement that I wrote for the ARL E-Science Task Force presents an example of a research life cycle model. See "e-Science and the Life Cycle of Research" (2006) available online at http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc.

[2] Life cycle does not refer to life span, which is the time from birth to death. Rather, life cycle is used to describe the processes within an environment under which resources are formed, transformed and re-used. For a brief summary of life cycle models and references to examples in addition to the one previously cited, see the item by Ann Green, "Conceptualizing the Digital Life Cycle" on the **IASSIST Communiqué** at http://iassistblog.org/?p=26.

[3] Original data collection is a defining aspect of primary research.

[4] The Data Documentation Initiative (DDI) is an example of such a metadata standard for survey, aggregate and time series data. Version 3 of DDI introduces a metadata model based on the life cycle of data. For further information, see http://www.icpsr.umich.edu/DDI/.

[5] For a recent, comprehensive discussion of the issues of digital repositories and their applications with research data, see Ann Green and Myron Guttmann, "Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives," 2006. Pre-print deposited in: http://deepblue.lib.umich.edu/.

[6] An example of a federation of repositories in the social sciences is the Data Preservation Alliance for the Social Sciences (Data-PASS), which is supported by the Library of Congress National Digital Information Infrastructure and Preservation Program. For more information, see http://www.icpsr.umich.edu/DATAPASS/.

[7] The argument for new national data archives of prominence has been made by James Jacobs and myself in "Preserving Research Data," **Communications of the ACM**, Vol. 47 (9), pp. 27–29.

[8] For further information about the RLG-NARA digital repository certification task force, see http://www.rlg.org/en/page.php?Page_ID=20769.

[9] The concept of pricing data at the marginal cost of reproducing a copy of the data is one of the Bromley Principles, which was published by the Committee on Earth and Environmental Sciences, National Science Foundation in "Data Management for global change research policy statements July 1991" **U.S. Global Change Data and Information Management Program Plan**, Washington, DC. 1992, pp. 42–48.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by John Leslie King, Vice Provost for Academic Information, and Professor, School of Information, University of Michigan

Access and Agency. The digital data universe brings significant changes in fungibility and interactivity in the origination, movement and use of data, information, and presumably, knowledge. The concept of "source" must be broadened to include not only human-mediated works of the kind we are used to, but direct sensory data from machine sensor networks and machine-constructed works at all levels from simple tabulation to syntheses that can take many forms (text, visuals, audio). These can be moved around at virtually zero marginal cost, and mashed-up into new creations by machines or humans for uses not even foreseeable at this time. Unlike the current "broadcast" model of most data/information transfer, in which a producer "sends" content to a consumer, the new model will embrace a growing population of producer/consumers whose roles are more difficult to distinguish. And, to top it off, if historical trends hold, this access and agency will eventually extend to the population of humans with access to telephony—at present about 2.8 billion people and growing rapidly (figure half the population of the Earth by 2031, a mere 25 years away).

Productivity Implications. Traditional labor productivity is simply the output produced as a function of labor input. Productivity is not very well understood in the realm of knowledge work, but everyone seems to think knowledge work is the important work for the 21st century. Assuming the traditional case – that dramatic changes in factor costs (e.g., the cost of moving information around) alter factor ratios (e.g., the amount of information that one person can provide to the population) – we can expect astonishing improvements in knowledge work productivity. We have been stuck trying to answer the question of what value academic libraries really provide for the academy and for the society at large. Other than the usual shibboleths about public goods and conservation of human knowledge, our stories are pretty anemic. This problem might soon change to one of explaining why we are not moving more quickly to provide the enormous benefits available to the world in the digital data universe. Time to think outside the stale old box.

Who's to Say? We've gotten used to knowing what's "real" and what's not in the realm of information and knowledge because we've built a huge credentialing infrastructure to answer such questions. We are moving to a world where producers and consumers are increasingly indistinguishable from one another, and the traditional production pathways can be ignored, along with their credentialing mechanisms. It will become more difficult to challenge the veracity and reliability of particular "entries" in the digital data universe, but more troubling, it will be increasingly difficult for anyone to claim and hold the authority to decide the answers to such questions. This should be of particular concern to academic libraries, which are residual claimants on such authority in many societies.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Rick Luce, Vice Provost and Director of the University Libraries, Emory University

Many challenges confront the preservation and access to digital scholarly information, as well the development of new capabilities which extend beyond the analog print paradigm. At the heart of these challenges lie infrastructure issues that surround the requirements for interoperability.

Data Model Interoperability. Beyond simple Web portal access and harvesting static repositories, raises issues related to defining (and implementing) the data model to utilize for digital objects, which must be commonly represented across heterogeneous, non-static repositories. The data model is important as a higher level of abstraction that must persist over time and support the following:

- Abstraction for digital objects, required so digital objects can be seen as an instance of the class defined by the data model, and provision for a level of abstraction which persists over time regardless of evolution of changing technologies and formats;
- Definitions for roles and quality assurance pertaining to the creation and maintenance of metadata, both man and machine generated;
- Quality assurance pertaining to the curatorial role of automated datasets;
- Rights, from confidentiality to DRM; and
- Sustainability (economic, social, organizational, technological).

Repository Interoperability. Enabling new value chains initiated in repositories that are:

- Cross-repository interoperable and federated. Note: repositories may organized by domain, discipline-orientation, institution or organization, type (e.g., dataset, learning object, format), etc., however, they should not be considered static nodes in a communication system merely tasked with archiving digital objects, and making them accessible through discovery interfaces. Rather, these repositories are be part of a loose, global federation of repositories, and scholarly communication itself is regarded to be a global workflow (or value chain) across dynamic repositories;
- Support a set of core services utilized via both machine and human user interfaces; and
- Facilitate emergence of richer cross-repository services.

Ecological Interoperable. Which enables:

- Persistent communication infrastructure, independent of changing technology, which records and expresses the origin and authority of the unit of scholarly communication;
- Global and automatically executed workflows and grid-enabled workflows which support

use and reuse across scholarly repositories;
- Data provenance in a heterogeneous networked environment;
- Distributed interoperable instruments and sensor-based registries;
- Information filtering which automatically pushes information to the user(s); and
- Emergent forms of social software, collaboration environments and networked based user profiles and network traversal log activity analysis.

Increasingly, value resides in the relationships between papers, their associations, and the supporting data sets and materials. To manage and utilize the potentially rich and complex nodes and connections in a large knowledge system such as the distributed Web, system-aided reasoning methods would be useful to suggest relevant knowledge intelligently to the user. As our systems grow more sophisticated, we will see applications that support not just links between authors and papers but relationships between users and information repositories, and users and communities. What is required is a mechanism to enable communication between these relationships that leads to information exchange, adaptation and recombination. A new generation of information-retrieval tools and applications are being designed that will support self-organizing knowledge on distributed networks driven by human interaction to support trans-disciplinary science. Through the use of these new tools, we will derive a shared knowledge structure that is based on users and usage in addition to that provided by author citations. Thus, the aggregated connections that readers make between papers and concepts will provide an alternative conceptualization of a given knowledge space. Such techniques will be coupled with classical search and retrieval methods, and these capabilities have an obvious utility for discovering and supporting evolving knowledge from these networks. The same concepts can be applied to data sets and rich media sources.

This emerging adaptive Web will analyse and use the collective behaviour of communities of users, utilizing concepts such as adaptive linking, which facilitates the evolution of knowledge structures based on collective user behaviour over time, and spreading activation, which uses a memory-recall process model from cognitive psychology. For example, using known keywords to search across distributed open archives, a user would receive recommendations of other conceptually related keywords, relevant articles, data sets and so on, based on semantic proximities linked across a multitude of distributed information resources. At the same time, the knowledge system the user has interacted with can begin to reorganize itself by incorporating feedback from the interaction into its knowledge structure. From the user perspective, such systems can use adaptive webs as a communication fabric to manage and co-evolve the knowledge traded with communities of members and users. Correspondingly, these new tools and systems will influence the adaptation of the structure and semantics of scientific discourse. Many questions remain unresolved, such as how we evaluate the knowledge structures and representations of such size and complexity.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Basic theme: Organizational Models

Comments by Barbara Lust, Professor, Human Development, and Janet McCue, Associate University Librarian for Life Sciences, Cornell University

Engaging the Research Community. Research scientists now face the necessity of a major shift in "zeitgeist" of how they must think about their data, and their labs in general, if they are to take advantage of the power and promise of the cyber-infrastructure-based digital environment in which we now exist. If research data sets created by individual research labs are to become part of a national or international digital data framework which is "open, extensible, and evolvable," they can no longer function in isolation. For example, they must adopt revised methods of data management which insure data preservation and data sharing. Unless the research community can be brought to understand the significance and the usefulness of such changes, they will not adapt.

Partnerships Between Research Labs and University Library and Across Libraries. Establishing partnerships between research labs and the university library provides a possible new infrastructure which is essential to the overall goal (a new "system of science" in the digital framework). This partnership fosters an environment where library and research lab do not function in isolation from each other. However, this infrastructure requires: (i) personal investment from both research lab and library personnel; (ii) interpersonal collaboration at an infrastructure level; (iii) university level support; (iv) new middleware for collaboration and coordination; (v) reciprocal adaptation by both the research lab and the library to the information structure of the particular data and materials involved. These needs arise at a repeated but higher order level when inter library exchange is developed.

Establishing Knowledge Networks and Related Ontologies Which Bridge Numerous Research Centers. New intermediary infrastructures can potentially help bridge the divides that now exist between individual research labs, between institutions housing these labs, and between lab and university libraries. One such structure is the "Virtual Center" in a knowledge area.

Basic theme: Sustainable Economic Models

Barbara Lust (Professor, Human Development) and I (Associate University Librarian for Life Sciences) are co-PIs on a project at Cornell that relates to research data and library/laboratory collaboration. The purpose of our NSF Small Grant for Exploratory Research is to test the feasibility of extending the role of large research libraries in supporting value-added services for research data, including access, metadata, outreach, training, and archiving. The exploratory grant was focused on one laboratory (Language Acquisitions Lab); a supplemental award targeted a second research group (Agricultural Ecology Program). In the supplement, we are

evaluating whether the conceptual model developed for language acquisition data is applicable to other disciplines. For an overview of the project's goals and accomplishments, see the project Web site. http://metadata.mannlib.cornell.edu/lilac/

Based on our experience in this planning grant, there are many significant issues to address in considering sustainable economic models, including capacity building, scaling-up, and determining future costs.

Staffing. Although the Teragrid is a reality, it will only reach its full potential when it is heavily traveled by a broader spectrum of the research community. It is a significant challenge to build the human capacity to deliver data and associated services in ways that support the research community. We will need skilled programmers to develop the tools for data-driven research and facilitate discovery and access; agile librarians with strong academic backgrounds to curate the collections and support end users; and sympathetic researchers who understand the value of good metadata, best practices, and archival decisions. For example, in our work with Lust's lab, both the metadata librarian and the programmer have linguistics backgrounds; in our AEP grant, our Research Data/Environmental Sciences Librarian, who has a graduate degree in Ecology & Evolutionary Biology, works closely with the 12 co-PIs in the project. Having these specialized backgrounds allows the library to more easily translate the needs of the research lab into services and to understand the curatorial and preservation aspects-of the data.

Scaling-up. We are working with two small projects in two labs within a single university, and some close collaborators. How do we scale-up to deal with oceans of data in diverse disciplines bridging multiple institutions? Can we leverage what we learn with one project and apply it to another? Can we mainstream some activities so that specialized staff consult and support staff process? Can we do a better job of capturing data at the point of creation, in formats that can be made accessible, re-purposed, archived, and mined?

Estimating Future Costs. It is difficult to determine long-term costs and long-term commitments when the models are still evolving. How do we determine long-term costs when the issues related to long-term availability/preservation of data are still puzzling us? If we develop collaborative repositories, how invested will the individual institutions and individual researchers be in sustaining of a cross-institutional entity? Can the costs be generalized for other institutions and other disciplines? Who is likely to bear the costs associated with research data discovery and preservation—research institutions? Granting agencies? Will STM vendors or universities or new entities offer subscriptions to institutions for services related to research data and will institutions/researchers be willing to pay for those services?

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

A Statement for the Management of Massive Datasets: New Partnerships

Comments by James L. Mullins, Dean of Libraries, Purdue University

In the future, a researcher will read an article multi-dimensionally. Not only will there be text, with hyperlinks to related literature or citations within the article, there will be links to the data reported within the article, through graphs, tables, illustrations, that will link to related datasets.

These datasets will be organized and retrieved based upon accepted, documented and well understood taxonomies and ontologies within the discipline, crosswalks will link, automatically, between one taxonomy and another to lead a researcher from one field of enquiry to another, thereby making connections from the findings and results in one field to another.

Disciplinary scientists, computer scientists, computer technologists, and librarians as data scientists within university, professional societies and other research entity will work as teams to collaborate, from creation of the research project to the dissemination of the findings to the curation of the data for the present and for the future.

Creating a "community proxy" will be the work of the disciplines in collaboration with librarians to determine the logical description and structure in which data would be organized and accessed. Massive data set repositories will be distributed around the world, in locations adjacent to related research centers, providing access to the international research community. Universities, research laboratories and governmental units will share in this undertaking, each picking off a "piece" of the massive undertaking, as exemplified by the recent agreement between the San Diego Super Computer (SDSC) and the National Archives and Records Administration (NARA).

This brave new world will be time consuming, challenging and expensive to create. It will be imperative that research-funding organizations such as the National Science Foundation step up to help facilitate and cause this vision to materialize.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

The Coming Metadata Deluge

Comments by James D. Myers, Associate Director, Cyberenvironments, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

With the rise of computing, our ability to produce content – from raw data to summarizing documents – has exploded. People speak of "information overload" and "data deluge" to describe the problems caused by this explosion for those trying to find, analyze, comprehend, and store the growing body of material available. To date, the automation of processes to capture the history and context of data has not kept pace, making the development of systems for data preservation, curation, and discovery extremely labor intensive. However, it is not clear that this situation is permanent, and, in fact, there are many reasons to think that there will soon be a 'metadata deluge' as our ability to capture and share metadata catches up with our capability to produce data. If such a metadata deluge occurs, it would profoundly affect the role of libraries and the design of curation and preservation infrastructure. An analogous change occurred with the data deluge – as our ability to create, store, and share content increased, our ability to organize information became a bottleneck, and infrastructure such as the World Wide Web arose. The Web, which directly supported the ability for experts and non-experts alike to organize information, created a market for third-parties to re-organize existing material and for 'competing' entities to offer alternate organizations. The Web also enabled those without the means or expertise to maintain content to none-the-less develop collections. More recent innovations such as blogs, wikis, and community spaces (e.g. MySpace and virtual 3-D worlds such as Second Life) go even further in enabling content creation and organization without technical expertise or owned infrastructure.

While this stack of Web technologies does not support the requirements for creating, managing, curating collections and for long-term information preservation, there are emerging extensions in this area that have the potential to spark a transformation in these areas analogous to the Web transformation of information publishing and organization. For example, global identifier schemes such as Handles and Digital Object Identifiers and Life Science Identifiers now provide Web gateway mechanisms to create persistent URLs. More recent schemes such as the Archival Resource Key (ARK) bring a more Web-centric view and additionally provide a means of decoupling the roles of the initial information provider and subsequent curator(s) in the way identifiers are generated and in how metadata is attributed. Extensions to HTTP such as WebDAV and URIQA provide means of managing versions and metadata in XML and RDF formats. Specifications such as the Java Content Repository API (JSR 170 and JSR 283) are standardizing the same types of functionality at the programming interface level. These technologies are making their way into large data grid and repository software, but they are also being used directly in scientific applications and environments to enable up-front capture of metadata and data provenance information. For example, the CombeChem project has

developed an experiment planning and execution environment that captures the experimental design, the provenance of data in specific experiments, and electronic notes related to the experiments as a single web of RDF information that can subsequently be searched, viewed, and potentially harvested. The Collaboratory for Multiscale Chemical Science (CMCS) with which I have been involved provides a similar service for applications to record any and all information related to data and experimental procedures that has focused more on connecting information across scientific disciplines and related to dynamic community assembly and evaluation of reference data and associated computational tools. Many other examples could be cited – from ones like these that are primarily driven by the goal of directly increasing researcher and community productivity to those that are more specifically focused on the long-term curation of data.

Working from the Web analogy, emerging semantic and content management technologies, and the exploratory projects using them, one can anticipate a period of rapid change in the curation and preservation of digital data and in the role of libraries. Very rich information – with all the detail captured and/or used by all instruments and applications in scientific experiments – will be available via standard protocols in self-descriptive schema and directly available, given authorization, for inclusion in institutional repositories, community databases, reference collections maintained by scientific associations, etc. The information collected and any additional annotations generated by third parties will be transferable (thanks to unique identifiers) and the data/metadata could be migrated, cached, replicated as needed by the organizations interested in it. Questions about what to collect may become much more graded – should the content be indexed only, should it be cached for performance, should it be copied to reduce risk or to extend the retention period, what metadata and ancillary data should be indexed, cached, copied along with the primary artifacts of interest? It is possible to imagine that different institutions may make very different choices in these areas to customize their solutions and provide added value for specific user bases, with or without global coordination or concepts as master copies or tiered collections.

In planning for the next-generation of digital data curation and preservation capabilities, it is important to question our assumptions. While the expertise gained over centuries in curation and preservation will be central to robust solutions, it will be necessary to disentangle principles of information management from practices that actually represent compromise based on the current limits of technologies and organizational structures. Conversely, while technological progress will play a driving role, complex socio-technical issues will be faced in defining practical solutions that align with cultural and economic realities and are 'just complex enough' to serve society's needs. If the web analogy is broadly valid, we are about to enter a period of rapid progress, new ideas, and new partnerships that will dramatically change and improve our ability to understand the world's information.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Frank Rack, formerly Director, Ocean Drilling Programs, Joint Oceanographic Institutions; currently, Executive Director, ANDRILL (Antarctic geological Drilling) Program

My views on the topic of the role of academic libraries in the preservation of digital data collections are primarily derived from a geosciences perspective, based on experiences from participating in and directing scientific drilling programs. These programs, which have enjoyed a long history of sustained support from NSF and international partners, have produced a substantial volume of physical samples (e.g., over 330 kilometers of sediment and rock cores) and data (both analog and digital) that are organized into structured collections using relational data base management systems as well as archives of unstructured data. These data holdings are expanding rapidly with the increasing acquisition of visual imagery and specialty data sets (e.g., volumetric imaging using X-ray CT and NMR/MRI) that are common to biomedical domains. Scientific drilling projects and programs cross NSF organizational boundaries spanning Earth, Ocean, and Polar Programs. These activities are now expanding to include the collection of observational time series across time and space with requirements for data streaming and remote user control of sensors and other resources. New models of sharing, storing, analyzing, and archiving data are required. Academic libraries have an important role to play in this new data universe.

Academic Libraries as Collaboration Centers for Research, Education, and Public Outreach. Academic libraries are naturally the center of the campus knowledge management and information exchange process and can be enhanced with physical resource investments to become "amplified collaboration environments" serving as a focal point for cyberinfrastructure access on academic campuses. The past NSF investment in connecting academic campuses to the high-end research networks, such asInternet-2 and National Lambda Rail can be enhanced with campus-wide dark fiber networks that support broadband connectivity to campus buildings and facilities, like libraries, that may be remote from the IT-2 or NLR node, but are centers of learning. These dark fiber networks are owned and operated by the academic IT infrastructure and provide capabilities to link researchers, educators, and students in a distributed collaborative environment dedicated to knowledge management and information sharing supported by data discovery, analysis, and visualization tools that can be accessed through the campus library system. The University of California at San Diego has already taken this step and is currently operating a dark fiber network on their campus.

These nodes could provide connectivity to the ACCESS GRID for collaborations with virtual research groups across campus or at other institutions and provide links with the DATA GRID for accessing computing resources and visualizations that are either pushed to the site or pulled from the site for research, educational or outreach activities. Investing in library infrastructure (hardware and software plus skilled staff) creates synergies with both local and remote research groups and encourages partnerships among researchers, academic staff and students through

opportunities for training and participation in focused demonstrations of the research outcomes using visualizations that can also be shared with the broader academic community and the general public. Academic libraries should be encouraged to establish collaborative relationships with local research groups and existing centers of excellence who provide digital content in exchange for data aggregation, public access, archiving and preservation services (either locally or remotely through networked data centers that provide tools and Web services that can be accessed easily by trained users), including publication of data.

Academic Libraries as Partners in Preservation of Analog Reference Collections and Legacy Data. Academic libraries are uniquely positioned to play an important role in the preservation of analog reference collections and legacy data for projects that are aligned with the research and education mission of a particular academic institution or the mission of local centers of excellence on each campus. The network of academic libraries should coordinate with each other to collaborate on content preservation efforts that build on their strengths while minimizing duplication of effort on a national (or better yet, and international) scale. Domain specific groups of researchers, working as community agents, could work with designated academic libraries to digitize analog collections for long-term preservation with appropriate metadata. This type of partnership would combine the traditional strengths of library professionals with the opportunities provided by collocated teams of researchers in a coordinated way to support a large-scale, interoperable, networked architecture for data discovery, information sharing and knowledge creation. The academic libraries would play a fundamental role in supporting research and educational goals within the context of a pervasive cyberinfrastructure that would require partnerships between federal, state and local (academic institution and local community) for investments that leverage technology to provide access to knowledge resources and training and outreach to stakeholders at all levels.

An example from scientific ocean drilling is the need to transform 40 years of analog "Proceedings" of the Deep Sea Drilling Project (DSDP) and Ocean Drilling Program(ODP) into electronic format through document scanning and OCR. These volumes weigh approximately 2000 pounds as a set and contain both data and metadata that could serve a wide community of users if they were readily available across the network. Plans to undertake this scanning/digitization project has been made and pilot studies have begun to transform this pile of paper into digital content through a partnership between the Texas A&M University Digital Library and the ODP Science Services group located at Texas A&M. Similar collections of key reference materials should be identified by specific domain science communities and prioritized for digital access in partnership with academic libraries.

The evolution of a distributed, networked, partnership among academic libraries, technology centers and research groups would require strategic planning and phased investments that leverage existing programs and initiatives to create new opportunities and enhance synergies among all parties. The prior NSF investments in establishing point-source academic infrastructures (e.g., Internet2 and NLR connections/nodes) should be leveraged by establishing a process to encourage the construction of dark fiber campus-wide networks

connecting libraries to research centers to support data sharing, open access, and preservation/archiving of data that can be provided to researchers, educators and communities of learners. The infrastructure investments should be combined with opportunities for the development of Web service architectures to support data discovery, analysis and visualization to create a transformational environment of innovation that would enhance knowledge creation and dissemination and stimulate learning.

Observational data in the future will be streaming from thousands to millions of field sensors that will require scalable visualization resources to allow humans to readily understand and comprehend the significance of these data. Academic libraries have a unique opportunity to establish a strategic role for themselves as centers for data integration, analysis and visualization, combined with a strong education and outreach mission. In order to realize this dream, academic libraries will have to form innovative partnership with a variety of research groups and broad-based communities of educators and technologists who can translate these challenges into coordinated action plans that capitalize on the opportunities promised by this transformation. Academic libraries will become next generation centers for learning, education, and public outreach, and will need to provide training to a broad range of users to articulate the significance of the new world view that accompanies this change in the data universe.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Some Key Issues for Consideration

Comments by Mark Sandler, Director, Center for Library Initiatives, CIC, Champaign, IL

Archiving. Archivists in academic settings often comment on the lack of interest that scholars show in the notes and working files of their colleagues. Likewise, in the social sciences as well as other disciplinary clusters, there seems to be more reward in gathering new data than building on top of the sources of others. Nonetheless, we know this to be mixed, and ICPSR and other data centers probably have a sophisticated understanding of the kinds of data that does and doesn't support new and important research. We should be tapping into that knowledge to establish guidelines for archiving datasets, both at the level of which datasets should or shouldn't be saved, as well as which should be saved as bit streams with minimal investment as opposed to those worthy of being refreshed, migrated forward and kept readily accessible to subsequent generations of researchers.

Aggregation. Aggregating data is efficient in terms of storage and management, and efficient as well in terms of retrieval. Perhaps disciplinary data farms should be developed, or perhaps this needs to be approached by funding agency or by academic institutions or collectives of academic institutions and research centers. A key to successful aggregation of data will be the emergence of standards around defining variables and achieving a degree of consensus about data gathering techniques that will permit greater comparability across studies. I understand that some differences in survey design and research methodology represent advances in a discipline, but I also understand that far too often these shifts are less about "progress" than idiosyncratic deviations masquerading as a research advance.

In the world of text archives, there is increasing emphasis on standards, keyboarding and scanning guidelines, DTDs, and substantive metadata, all of which further content integration and system interoperability. The underlying theory here is that data gathered for a particular study or purpose might be more valuable as part of a larger whole than if self-contained and self-referential.

Transparency. However it's done, data (especially data gathered with the assistance of public funding) should be more broadly available for public scrutiny and further analysis. I think we all respect the right of a researcher to have the time required for careful analysis of data he or she has gathered before it is opened up to the world. On the other hand, closing down access to useful data for many years on the off chance that the researcher will someday be inclined to return to the dataset seems selfish and not in the best interest of advancing scholarship. As with so many other issues in the underlying social relations of academe, we need to become clearer as a community about the social responsibility of scholars to engage in dialogue with the larger society.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by MacKenzie Smith, Associate Director for Technology, MIT Libraries

There are two aspects of scientific and engineering data that relate to academic libraries:

- Data as *primary source material* available for further research and experimentation, using particular datasets or groups of datasets; and
- Data as part of *enhanced publications* that form the basis of modern, digital scholarly communication.

Academic research libraries and archives are closely involved with both of these as a part of their mission and expertise. However, broadening the scope of libraries and archives to include digital scientific research data brings big challenges. There are unanswered questions about the:

- Technical infrastructure, and who will develop and manage it;
- Collection practices involving decisions about what data will be kept, when, in what form, with what tools, what description;
- Digital preservation practices (of unknown difficulty and expense); and
- Legal framework that is necessary to allow this to happen at all.

Libraries and archives will probably not be the primary providers of the large-scale storage infrastructure required. Nor will they provide the specialized tools to work with the data (sometime at the level of individual datasets). They will also not provide detailed information about the data (which falls to researchers, or specialists from their societies and publishers). Nor will they provide the legal framework to enable open science. However to achieve economies of scale *across all scientific research domains* and not just create data silos within particular scientific sub-disciplines, there is value in library practices around:

- Collection policies and practices (appraisal, selection, weeding, destruction, etc.);
- Data clean-up, normalization, description, and submission to archives; and
- Collaboration with researchers around scholarly communication practices of the disciplines (e.g., educating students about these practices, or helping researchers find appropriate archives or publications).

It's unclear whether libraries will provide the technical solutions to long-term digital data preservation. It is certainly within the mission of research libraries and archives to preserve the scholarly record, but the technical challenges and costs involved are large, and libraries will need to invest seriously in this area if they wish to help find solutions.

Finally, for "enhanced publications" that include scientific data as a useful part of networked documents, there are missing standards that academic libraries are well positioned to help

define, including:

- Ontologies (for complex publications that include data);
- Identifiers for publication parts that work across disciplines;
- Consistent description practices for enhanced publications and their parts;
- Data structuring conventions; and
- Interoperability protocols for searching and retrieving data.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Eric F. Van de Velde, Director, Library Information Technology, California Institute of Technology

Both large and small research projects produce data of historical value worthy of preservation. Large projects must incorporate data preservation as an essential part of their project. Small projects need institutional support to help them implement high-quality data-preservation policies. Funding agencies should create Centers of Excellence in Preservation and encourage peer review of data sets and associated services and software. Computer-and library-science curricula should include data preservation.

The size and specialized nature of the data of large research projects (Caltech-MIT's LIGO, Human Genome Project…) require that data preservation is considered as an essential component of the project. These projects have the responsibility to manage the data, to make it available with appropriate services, and to preserve the data and the associated software and documentation. Often, the desired outcome is a data set that continues to grow indefinitely, supplemented with data from newer, more accurate, observations and experiments. Terminations of projects like these are relatively rare events that should be handled on a case-by-case basis as part of the closing-down process, which should be supervised and audited by the funding agency.

Small research projects need institutional help with the work required to comply with preservation mandates. Managing scientific data requires scientific know how at an expert level, and the local research library cannot be expected to handle preservation of data of all disciplines. It might be feasible, however, for a library to specialize in the preservation of data in one or two disciplines. For the rest, the library is the agent between local researchers and specialized data archives.

Funding agencies should fund (distributed) Centers of Excellence in Data Preservation, each specializing in a particular discipline. As part of the competitive funding process, interested institutions would develop collaborative organizational networks capable to implement effective preservation of specific data. This approach allows for organic growth, proportional to the actual needs. This approach also builds on the strengths of existing institutions (universities, research laboratories, and their libraries).

Funding agencies should also provide incentives to accomplish more than "just archiving." Data obtained at great effort and expense should be made available as widely as feasible together with supporting services and software. Peer review of data sets and associated services and software would make it easier to consider this kind of work in tenure and promotion. Under suitable conditions, for-profit organizations could provide services using publicly available data, ensuring the use of this data for society's benefit.

Finally, we must ensure that the talent to preserve scientific data will be available. The preferred approach is to provide incentives for computer-science and library science departments to include suitable disciplines in their curricula.

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Todd Vision, Associate Director of Informatics, The National Evolutionary Synthesis Center, University of North Carolina at Chapel Hill

The Respective Roles of Scientific Publishers and Institutional Libraries. Currently, the responsibility for management of data accompanying scientific publications falls to publishers in the form of supplemental data collections. Academic journals have little incentive to invest in the establishment and maintenance of digital data repositories that can be used for anything beyond minimal documentation of published reports. If there were a uniform and reliable system for digital data management that was hosted by researcher's home institution, this could potentially supplant the current system whereby unstructured supplemental data is deposited at publication. However, such a system would need to be available widely, and not just at elite institutions, in order to be a viable alternative for publishers.

The Untapped Value of Raw Data to the Researcher. In many scientific fields, difficult-to obtain and essentially irreproducible datasets (such as long-term field observations in ecology) are analyzed and reanalyzed by the same research group over a period of many years, resulting in multiple publications, each containing only such statistical summaries of the data as are necessary to support the claims in the publication. Such datasets may continue to grow and accrete value over time. Understandably, the researcher feels entitled to an indefinite term of exclusive use, and there may be no single moment at which he or she would be comfortable providing even limited access to the full dataset. How common is this situation, and how worthwhile is it to invest in a system of scientific data preservation that would exclude such unique and valuable data collections? Or is there a way to manage such data while protecting, or alleviating, the researcher's concerns of exclusivity.

The Burden of Metadata Curation. Digital data is nearly useless without extensive and high-quality metadata, both for resource discovery and for interpretability of the data itself. The researcher knows the data, but doesn't necessarily have sufficient expertise in information science to provide quality metadata curation, while librarians have the opposite problem. What incentives can be provided to researchers to undertake the burden of careful metadata curation, how can this task be made more manageable to nonexperts, and what incentive can be provided to institutions for QC of metadata produced by researchers?

Top Three Issues in the Long-Term Preservation, Management, and Curation of Scientific Data in Digital Form

Comments by Tyler O. Walters, Associate Director for Technology and Resource Services, Georgia Institute of Technology, Library and Information Center

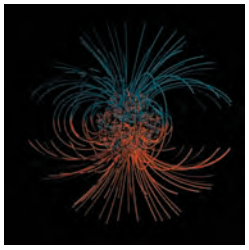
Issues in Digital Data Curation Leading Us to the Need for New Partnerships
Among the top three issues effecting the growth and future of digital data curation involves: 1) Culture and policy frameworks; 2) Technology integration and data curation tools; and 3) Economic sustainability of our partnership and consortial models.

The expense and enormity of the challenge to ensure the preservation, curation, and overall management of scientific digital data lends itself to a multi-institutional solution. However, in order to build effective partnerships, we must first begin with the single organization. Raising awareness of the need and benefits to managing digital data will be paramount. Universities will need to create programs and incentives that **embed digital data curation in university culture**, becoming an integral part of research projects. The initial **change driver can be policy**. Granting agencies such as NSF and NIH (DOE could be another source for a data curation policy) are considering data access and management policies that may drive universities with agency-funded research projects in this direction. In response, universities will need to make a commitment to the use, re-use, and maintenance of digital data. However, pragmatics also will dictate that not all data can be sustained. This situation requires designing criteria for selecting which sets of digital data will and will not become a long-term university responsibility. Libraries can contribute to the selection process by adapting archival appraisal theory as well as other parameters to judge which research materials are worthy of long-term accessibility. This combination of policy setting, awareness raising, cultural engineering, and selection criteria building, will become essential components in the rise of digital data curation programs. Libraries can assist with developing and implementing this agenda and be equal partners with scientists, advanced technologists, and policy makers. New partnerships and services akin to Genbank, operated by the NIH / NLM's National Center for Biotechnology Information, may need to be established. They may take on a discipline-based alignment, such as several universities with deep interests in astrophysics forming a consortium to provide data curation services for their home institutions' related data sets.

Much work remains in areas such as **improving technology integration and building reliable data curation tools**. Data does not reside in just one information system; therefore, integration between systems is critical. Universities have research data residing in many applications such as databases (commercial and open), digital asset management systems, content management systems, and repositories. Curating this data consequently may include moving it from one system to another, linking it between systems, and migrating it to a central system. The exact future architecture is undetermined and many universities and consortia may take divergent paths. There will be common format and metadata portability issues. Ongoing development projects such as the Global Digital Format Registry (GDFR) can play an essential role in

matching formats (including database programs and protocols) and providing information on software that can read certain databases and data formats, and recommend migration paths. Data curators will need tools for data and metadata extraction, database emulation, data provenance tracking, and to document the origin, use, and re-use of data. Partnerships and consortia can play a key role. They can promote the further development of new data curation technologies as well as standards and technical protocols to ensure interoperability and data migration. They can be vital to maximizing present resources and strategies in generating these new technologies through their synergistic activities.

If the work described above progresses, then the **economic sustainability of these types of partnerships and consortia** will become mandatory. The bottom line – funding and revenue streams need to be established. I mixture of funding sources will best guarantee the success of these new entities and help them to not become too reliant on any one source of funds. Partner dues, seed and project monies from grant agencies and private foundations, revenues from a variety of service and consulting fees, and several other creatively produced sources of funding are examples of cooperative ways to sustain the new partnerships. Those interested and vested in digital data curation should explore deeply new and dynamic models of organizational and economic sustainability. This is an opportunity to reinvent ourselves for the better as we face inherently new challenges in managing complex research objects such as data sets.



Appendix F. Examples of Scientific Community Archives

Examples of Scientific Community Archives

Rick Luce, Emory University

A number of disciplines maintain archives with submissions from their communities. These are hosted under a variety of rubrics, including some publishers. Some of them include:

arXiv (1991–), Physics, mathematics, computer science; main administration site at Cornell University, multiple mirrors worldwide, manages access to over 230,000 papers, abstracts include links to citation analysis for the paper by SLAC Spires and Citebase. http://www.arXiv.org/.

Citeseer (1998–, aka ResearchIndex), developed at NEC Research Institute, NJ, USA, caches openly accessible full-text research papers on computer science found on the Web in Postscript and PDF formats for autonomous citation indexing. http://citeseer.nj.nec.com/cs.

ebizSearch (2001–), administered by the eBusiness Research Center at Pennsylvania State University, based on Citeseer software, academic articles, working papers, white papers, consulting reports, magazine articles, published statistics, and facts. http://gunther.smeal.psu.edu/.

K-theory Preprint Archives*\ (papers from 1995), managed at Mathematics Department, University of Illinois at Urbana-Champaign. http://www.math.uiuc.edu/K-theory/.

Topology Atlas preprint server (papers from 1995), was most active in 1996 and 1997 and still accepts submissions but suggests using the Mathematics Archive (arXiv.org or its Front) for distributing and finding preprints, hosted at York University, North York, Ontario. http://at.yorku.ca/topology/preprint.htm.

Algebraic Number Theory Archives (papers from 1996, frozen since Jan. 2003), hosted at the Mathematics Department, University of Illinois at Urbana-Champaign. http://www.math.uiuc.edu/Algebraic-Number-Theory/.

Mathematical Physics Preprint Archive, mp_arc (papers from 1991), hosted by Mathematics Department, University of Texas at Austin. http://rene.ma.utexas.edu/mp_arc/index.html.

Hopf Topology Archive (papers from 1997), hosted by the Department of Mathematics, Purdue University. http://hopf.math.purdue.edu/.

Preprints on Conservation Laws (papers from 1996), administered at Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim. http://www.math.ntnu.no/conservation/.

MCMC (Markov Chain Monte Carlo) methodology Preprint Service (papers from 1993), administered at the Statistical Laboratory, University of Cambridge. http://www.statslab.cam.ac.uk/~mcmc/index.html.

Jordan Theory Preprint Archives (papers from 1996), hosted at Institut fur Mathematik, Universitat Innsbruck. http://mathematik.uibk.ac.at/mathematik/jordan/index.html.

Groups, Representations. and Cohomology Preprint Archive (papers from 1995), managed at Department of Mathematics, University of Georgia, USA. http://www.math.uga.edu/archive.html.

Field Arithmetic Archive, located at Ben Gurion University in Be'er-Sheva, Israel, stores electronic preprints on the arithmetic of fields, Galois theory, model theory of fields, and related topics. http://www.cs.bgu.ac.il/research/Fields/.

MGNet preprints (papers from 1991, last paper deposited 2001), Department of Computer Science, Yale University, repository for information related to multigrid, multilevel, multiscale, aggregation, defect correction, and domain decomposition methods. http://casper.cs.yale.edu/mgnet/www/mgnet-papers.html.

Cogprints (1997–), an electronic archive for self-archived papers in any area of Psychology, Neuroscience, and Linguistics, and many areas of Computer Science, Philosophy, Biology,

Medicine, Anthropology, as well as any other areas pertinent to the study of cognition, initially a project in the JISC Electronic Libraries (eLib) Programme, administered by the IAM Group, University of Southampton. http://cogprints.ecs.soton.ac.uk/.

E-LIS, E-Prints in Library and Information Science. http://eprints.rclis.org/.

DList, Digital Library of Information Science and Technology (October 2002–), managed by School of Information Resources and Library Science and Arizona Health Sciences Library, University of Arizona. http://dlist.sir.arizona.edu/.

NetPrints Clinical Medicine and Health Research (December 1999–), BMJ Publishing Group and HighWire Press, a repository of non-peer reviewed original research. http://clinmed.netprints.org/home.dtl.

Chemistry Preprint Server (August 2000–), ChemWeb.com, Elsevier. http://www.chemweb.com/preprint.

Computer Science Preprint Server (November 2001–), Elsevier. http://www.compscipreprints.com/comp/Preprint/show/index.htt.

Mathematics Preprint Server (May 2001–), Elsevier, a registered data provider, supporting OAIv2.0. http://www.mathpreprints.com/math/Preprint/show/. (842 papers at 25th March 2003)

HTP Prints, the History & Theory of Psychology Eprint Archive (September 2001–), administered at York University, Toronto. http://htpprints.yorku.ca/.

Education-line (1997–), a freely accessible database of the full text of conference papers, working papers, and electronic literature which supports educational research, policy, and practice, initially a project in the JISC Electronic Libraries (eLib) Programme, administered by the Brotherton Library, University of Leeds. http://www.leeds.ac.uk/educol/.

Social Science Research Network (SSRN), Social Science Electronic Publishing, Inc., working papers and abstracts are provided by journals, publishers, and institutions for distribution through SSRN's eLibrary, which consists of two parts: a database containing abstracts on over 49,200 scholarly working papers and forthcoming papers, and an Electronic Paper Collection containing over 30,800 (27 March 2003) downloadable full-text documents. SSRN is composed of specialized research networks/journals in the social sciences: Accounting, Economics, Financial Economics, Legal Scholarship, Management, Negotiations.

Electronic Colloquium on Computational Complexity (papers from 1994), led by the chair of theoretical computer science and new applications at the University of Trier.

Research reports, surveys, and books in computational complexity. http://www.eccc.uni-trier.de/eccc/.

Cryptology ePrint Archive (2000–), maintained by the International Association for Cryptologic Research (IACR), incorporates contents of the Theory of Cryptology Library 1996–1999. http://eprint.iacr.org/.

The Digital Library of the Commons (DLC), Indiana University, contains a Working Paper Archive of author-submitted papers, as well as full-text conference papers, dissertations, working papers, and pre-prints. (The commons is a general term for shared resources in which each stakeholder has an equal interest. Studies on the commons include the information commons with issues about public knowledge, the public domain, open science, and the free exchange of ideas.) http://dlc.dlib.indiana.edu/.

Organic Eprints (September 2002–), established by the Danish Research Centre for Organic Farming (DARCOF), open access archive for papers related to research in organic agriculture. http://orgprints.org/.

University of California International and Area Studies (UCIAS) Digital Collection (October 2002–), partnership of the University of California Press, the California Digital Library (CDL), and internationally oriented research units on eight UC campuses, publishes articles, monographs, and edited volumes that are peer-reviewed according to standards set by an interdisciplinary UCIAS Editorial Board and approved by the University of California Press. http://repositories.cdlib.org/uciaspubs/.

Formations, Faculty of Arts, University of Ulster, hosts eprints in Media Studies and participative "eLearning Forums" based on short discussion papers. Initially a project in the JISC Electronic Libraries (eLib) Programme. http://formations2.ulst.ac.uk/.

Ecology Preprint Registry (papers from July 2001), hosted at the National Center for Ecological Analysis and Synthesis, dissemination of new research results destined for publication (i.e., not white papers or gray literature), only preprints with a theoretical basis can be submitted, the scope may be expanded to include submissions from the entire discipline of ecology. http://www.nceas.ucsb.edu:8504/esa/ppr/ppr.Query.

Note: This is a selection of the types of resources that are available and is offered for purposes of illustration. Inclusion in the list does not represent an endorsement nor should any meaning be inferred about resources that have not been identified here.

ASSOCIATION OF RESEARCH LIBRARIES