Digital Research Data Sharing and Management

December 14, 2011

National Science Board

National Science Foundation

National Science Board

Est. 1950

Cover Design by Brandon J. Powell, Policy Branch
National Science Board Office, National Science Foundation

NSB-11-79
December 14, 2011

Digital Research Data Sharing and Management

December 2011

Task Force on Data Policies
Committee on Strategy and Budget
National Science Board

National Science Board

Ray M. Bowen, *Chairman*, President Emeritus, Texas A&M University, College Station, Texas, and Visiting Distinguished Professor, Rice University, Houston, Texas
Esin Gulari, *Vice Chairman*, Dean of Engineering and Science, Clemson University, Clemson, South Carolina

Mark R. Abbott, Dean and Professor, College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, Oregon
Dan E. Arvizu, Director and Chief Executive, National Renewable Energy Laboratory, Golden, Colorado
Bonnie Bassler*, Howard Hughes Medical Institute Investigator, Squibb Professor of Molecular Biology Princeton University, Princeton, New Jersey
Camilla P. Benbow, Patricia and Rodes Hart Dean of Education and Human Development, Peabody College of Education and Human Development, Vanderbilt University, Nashville, Tennessee
John T. Bruer, President, The James S. McDonnell Foundation, Saint Louis, Missouri
France A. Córdova, President, Purdue University, West Lafayette, Indiana
Kelvin K. Droegemeier, Vice President for Research, Regents' Professor of Meteorology and Weathernews Chair Emeritus, University of Oklahoma, Norman, Oklahoma
Patricia D. Galloway, Chief Executive Officer, Pegasus Global Holdings, Inc., Cle Elum, Washington
José-Marie Griffiths, Vice President for Academic Affairs and University Professor, Bryant University, Smithfield, Rhode Island
Louis J. Lanzerotti*, Distinguished Research Professor of Physics, Center for Solar Terrestrial Research, Department of Physics, New Jersey Institute of Technology, Newark, New Jersey
Alan I. Leshner, Chief Executive Officer, Executive Publisher, *Science*, American Association for the Advancement of Science, Washington, DC
W. Carl Lineberger, Fellow of JILA, E. U. Condon Distinguished Professor of Chemistry, University of Colorado, Boulder, Colorado
G.P. "Bud" Peterson, President, Georgia Institute of Technology, Atlanta, Georgia
Douglas D. Randall, Professor Emeritus and Thomas Jefferson Fellow and Director Emeritus Interdisciplinary Plant Group, University of Missouri-Columbia, Columbia, Missouri
Arthur K. Reilly, Retired Senior Director, Strategic Technology Policy, Cisco Systems, Inc., Ocean, New Jersey
Anneila I. Sargent, Benjamin M. Rosen Professor of Astronomy, Vice President for Student Affairs, California Institute of Technology, Pasadena, California
Diane L. Souvaine, Professor of Computer Science and Mathematics, Tufts University, Medford, Massachusetts
Arnold F. Stancell, Emeritus Professor and Turner Leadership Chair, School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia
Claude M. Steele, Dean, School of Education, Stanford University, Stanford, California
Thomas N. Taylor, Roy A. Roberts Distinguished Professor, Department of Ecology and Evolutionary Biology, Curator of Paleobotany in the Natural History Museum and Biodiversity Research Center, The University of Kansas, Lawrence, Kansas
Richard F. Thompson, Keck Professor of Psychology and Biological Sciences, University of Southern California, Los Angeles, California
Robert J. Zimmer, President, University of Chicago, Chicago, Illinois

Member *ex officio*: **Subra Suresh**, Director, National Science Foundation, Arlington, Virginia

Michael Van Woert, Executive Officer, National Science Board and National Science Board Office Director, Arlington, Virginia

Committee on Strategy and Budget Task Force on Data Policies

José-Marie Griffiths, *Chairman*

Mark R. Abbott	**John T. Bruer**	**G.P. "Bud" Peterson**
Camilla P. Benbow	**Arthur K. Reilly**	**Thomas N. Taylor**

National Science Foundation Members of the Task Force on Data Policies
Alan Blatecky Myron Gutmann Farnam Jahanian Edward Seidel

Philip Bogden and **Robert Pennington**, Task Force on Data Policies Executive Secretaries
Blane Dahl, Task Force on Data Policies Staff Lead

*Board Consultant

Contents

National Science Board

December 14, 2011

MEMORANDUM FROM THE CHAIRMAN OF THE NATIONAL SCIENCE BOARD

SUBJECT: Digital Research Data Sharing and Management

The progress of science and engineering has always been dependent on the collection of data. A core expectation of the scientific method is the documentation and sharing of results, underlying data, and methodologies. The increasing ease with which digital research data are gathered, processed, analyzed, and disseminated has expanded the scale, scope, and complexity of science and engineering data collections and highlights the need for improved research data policies. One of the functions of the National Science Foundation (NSF) is "to provide a central clearinghouse for the collection, interpretation, and analysis of data on scientific and engineering resources" ("National Science Foundation: Functions," Title 42 *U.S. Code*, Chpt. 16. Sec. 1862). Therefore, NSF is dedicated to improving and implementing policies that provide a strong and sustainable foundation for sharing and managing digital research data for the benefit of the science and engineering research community. This report of the National Science Board (Board) presents key challenges and recommendations related to the sharing and management of digital research data generated by NSF-funded activities.

In February 2010, the Board established the Task Force on Data Policies under the Committee on Strategy and Budget. The task force was charged with the further refinement of NSF data policies to address key challenges and outline possible options to more effectively use digital research data to meet the mission of NSF. This strategy builds on past and ongoing efforts by the Board, NSF, and other organizations. In addition to the National Science and Technology Council's report, *Harnessing the Power of Digital Data for Science and Society*, and the National Research Council's *Ensuring the Utility and Integrity of Research Data in a Digital Age*, the 2005 Board report, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (NSB-05-40), is especially relevant to the challenges and recommendations presented in this report.

The Board believes that timely attention to digital research data sharing and management is fundamental to supporting U.S. science and engineering in the twenty-first century. This report recognizes the evolving role of data in science and society and strong and sustainable data sharing and management policies as a critical national need. We exhort you to join the Board in encouraging digital research data sharing and management for the purpose of science and engineering progress.

Ray M. Bowen
Chairman

Acknowledgements

The National Science Board (Board) appreciates the numerous individuals who contributed to the work of the Board's Committee on Strategy and Budget Task Force on Data Policies, including the distinguished panelists and discussants who participated in the March 2011 expert panel discussion and provided significant input into the development of this report.

We are particularly indebted to Dr. Subra Suresh, Director of the National Science Foundation (NSF), and Dr. Cora Marrett, NSF's Deputy Director, for their support of the Task Force's efforts and providing experts from within NSF to assist and advise the Board on this endeavor. Especially deserving of thanks are the Task Force's Executive Secretaries, Dr. Philip Bogden and Dr. Robert Pennington.

The Board would also like to acknowledge the efforts of Dr. Gina Walejko, Dr. Jason Gallo, Ms. Allison Laskey, and Mr. Sam Thomas at the IDA Science and Technology Policy Institute for their tireless efforts during meetings of the task force, their expert analytical work supporting the project, their input in drafting major sections of the report, and their responses to questions during the development of the report.

The National Science Board Office (NSBO) provided essential support to the work of the Task Force on Data Policies. Especially deserving of recognition are Mr. Blane Dahl, NSBO staff lead for the Task Force on Data Policies, for his thoughtful and diligent work throughout the duration of this initiative; Ms. Jennie Moehlmann, for policy guidance and critical review of numerous drafts of the report; Ms. Ann Ferrante, for editorial assistance throughout the drafting process; and Ms. Betty Wong for providing administrative support for the expert panel discussion. Lastly, Dr. Michael L. Van Woert, Executive Officer of the Board and NSBO Director, provided essential guidance and support throughout the duration of the project.

Process for Producing the Report

The National Science Board's Committee on Strategy and Budget Task Force on Data Policies was established at the February 3-4, 2010, National Science Board (Board) meeting and charged with further defining identified data policy issues and outlining possible options to more effectively use digital research data to meet the mission of the National Science Foundation (NSF). Appendix A contains details of the Board's charge to the task force. The Board considered significant background material on current NSF data policies, data policies at other Federal agencies and at international counterparts to NSF, and the views of NSF awardees on the value of data policies and their administrative burden.

Digital Research Data Sharing and Management proceeds from four phases of work under the Board:

- The Task Force on Data Policies heard presentations from three invited speakers[i] in December 2010. Presentations were followed by a discussion of their content. The goal of these presentations was to gain a better understanding of open access publishing.

- An NSF-wide requirement, implemented in January 2011, requires researchers to submit a data management plan with each grant proposal submitted to NSF.[ii] This plan describes how the proposal will conform to NSF policy on the dissemination and sharing of research results, and it is considered as part of the merit review criterion.

- Appendix B is the Statement of Principles related to data sharing and management approved by the Task Force on Data Policies in February 2011. This statement was intended to guide the Board in framing and examining pertinent digital research data issues and developing relevant digital data policies.

- The Board sponsored an expert panel discussion in March 2011. The goal of this workshop was to obtain input from researchers, universities, research libraries, publishing companies, industry, scholarly societies, and public and private funding agencies in order to examine and frame issues associated with science and engineering digital research data. The expert panel discussion's charge to workshop invitees, participant agenda, workshop participants, and summary notes are available in Appendixes C, D, E, and F, respectively.

---

[i] The three invited speakers were John Vaughn, Ph.D., Executive Vice President, Association of American Universities, Washington, DC; David Lipman, M.D., Director of the National Center for Biotechnology Information at the U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland; and Bernard Schutz, Ph.D., Director of the Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Potsdam, Germany.

[ii] See National Science Foundation web site, Data Management & Sharing Frequently Asked Questions (FAQs), http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp.

In December 2011, the Board invited public comments on a pre-publication copy of *Digital Research Data Sharing and Management* via a *Federal Register* notice. The Board received and reviewed these public comments and incorporated relevant changes in this final version of the report.

The recommendations of the Board in this report are based on inputs of the three invited speakers, implementation of the data management plan in NSF grant proposals, the Statement of Principles laid out by the task force, and findings from the Board-sponsored expert panel, in addition to significant Board deliberations. The recommendations reflect the Board's firm commitment to ensuring broad, timely, and sustained access to digital research data; addressing the cost burdens associated with managing digital research data; and developing a qualified workforce in data-enabled science and engineering.

Introduction

Progress in science and engineering depends on the collection of data through observation, experimentation, and computation. A core expectation of the scientific method is the documentation and sharing of results, underlying data, and methodologies. This process enables other researchers to reproduce experiments and studies, verify and validate results, and build upon previous work to produce further scientific advances. The increasing ease with which research data can be gathered, processed, analyzed, and disseminated digitally has greatly expanded the scale, scope, and complexity of science and engineering data collections. Increased National Science Foundation (NSF) funding of data-intensive projects, which are often, but not exclusively, large-scale, collaborative efforts, has highlighted the need for improved data policies to maximize the use and value of digital research data. Enabling access to digital research data helps promote broad participation at all levels of scientific and engineering research and education.

Digital research data have proliferated due to advances in information and computational technologies that have made gathering and processing data from large-scale, collaborative projects easier.[1] In this report *digital research data* refers to all data generated in a digital format, analog data that have been subsequently digitized, and digital metadata that may be associated with digital research data, as a result of research funded by NSF. The National Science and Technology Council Interagency Working Group on Digital Data defines *Digital research data* as data

> *…appropriate for use or repurposing for scientific or technical research and educational applications when used under conditions of proper protection and authorization and in accordance with all applicable legal and regulatory requirements. It refers to the full range of data types and formats relevant to all aspects of science and engineering research and education in local, regional, national, and global contexts with the corresponding breadth of potential scientific applications and uses.*[2]

Recognizing that the proliferation of digital research data has significant policy implications, the National Science Board (Board) Committee on Strategy and Budget established the Task Force on Data Policies to lead a broad examination on how research data collected with NSF funding are shared and managed to ensure broad, timely, and long-term availability to the research community. Further developing NSF's current policies to promote effective management of, and broad access to, digital research data is in the national interest and warrants careful examination by NSF. Such policies should be informed by past and ongoing efforts by the Board, NSF, and other organizations in this area and provide the flexibility to effectively and efficiently accommodate future digital research data needs.

The Board is committed to the continuing development, implementation, and assessment of policies that promote efficient management of, and broad access to, digital research data that result from NSF-funded activities. This commitment includes sharing of results, data, physical collections, and other supporting materials created or gathered in the course of NSF-funded research. Policies that ensure efficient management and broad access are critically important to

NSF as it carries out its mission to promote the progress of science and engineering. The Board, in taking up this topic, strongly encourages NSF to seize the opportunity to exercise national and international leadership to promote sharing and management of digital research data for the benefit of the science and engineering community and society.

Key Challenges

Science and engineering increasingly depend on abundant and diverse digital data. The increasing scale, scope, and complexity of datasets pose significant challenges for the science and engineering research community because they fundamentally change the way that researchers share, store, and analyze data. Thus, the reliance of the science and engineering community on digital data marks a transition in the conduct of research.

New approaches to data sharing and analysis are transforming the conduct of research in fields that are proactive in engaging with large-scale datasets. Other research communities are just beginning to grapple with the implications of proliferating data. Regardless of their experience, research communities will benefit greatly from the leadership, guidance, and coordination of national and international funding agencies.

This section presents ten key challenges organized under six areas: commitment to sharing; reproducibility; education, training, and workforce development; cyberinfrastructure; longevity and sustainability; and ethical and legal implications. The key challenges are drawn from several prominent reports that are listed in the bibliography, from the expert panel discussion convened by the Task Force on Data Policies, and from the Task Force on Data Policies' Statement of Principles, highlighted in Appendixes B through F. Addressing these challenges supports the entire research enterprise and enables the verification, reproducibility, and extension of science in the information age. The National Science Board discusses the following key challenges in the broad context of data, in general. This report's recommendations, however, relate specifically to digital data associated with scientific and engineering research.

Commitment to Sharing

Key Challenge #1: Broad stakeholder involvement and commitment to clear and realistic goals and measures of progress are necessary to ensure sustainable data sharing and management.

Stakeholder communities, including active researchers from multiple disciplines, universities, research libraries, publishing companies, industry, scholarly societies, and public and private funding agencies, play critical roles in sharing and managing data that can benefit the entire science and engineering research community. To address the challenges associated with increasing scale, scope, and complexity of data, each science and engineering research community should take the responsibility for determining its own standards and conventions for data stewardship and for coordination across the research enterprise. Funding agencies and stakeholder communities must partner together during data policy development so that recommendations can be implemented by each science and engineering research community. Chapter 3 of the Board's *Long Lived Digital Data Collections*[3] elaborates on the roles and responsibilities of individuals and institutions, namely data authors, data managers, data scientists, data users, and funding agencies, and notes the importance of collective action of these groups to pursue many higher-level goals of data sharing and management. This collective action must result in goals supported by realistic plans, processes and outcomes.

Key Challenge #2: As data collections expand in scale, scope, and complexity, successful data sharing and management require a change in research and institutional cultures.

Data sharing needs to be fully accepted as a common, beneficial practice by all science and engineering research communities. Data sharing and data management policies must acknowledge and provide for disciplinary nuances, while simultaneously establishing a culture of interdisciplinary sharing across the research communities. Thus, data sharing and management policies must be flexible and driven by communities of practice.

Instituting a culture of data sharing may occur both in bottom-up fashion within research communities and through top-down funding agency policies and guidance. Research communities should develop standards that are accepted across fields of science and engineering. Funding agencies should promote and reward exemplary projects and the implementation of data management plans. Funding agencies should also create policies that promote data availability in a timely fashion in order for data sharing to produce maximum benefit and foster productive research collaborations. Sharing can also be encouraged through the establishment of professional incentives such as promoting the publication of data in a format that allows for citation and verification.

Key Challenge #3: Data sharing requires the coordination of goals and efforts through international collaborations and activities.

Research sponsored by the U.S. relies heavily on international collaborations that co-develop, acquire, manage, and share datasets. U.S. institutions and researchers work within a global data environment; thus, it will be crucial to address the key challenges presented here within the context of international science. Stakeholders around the world are engaging in analogous discussions regarding data sharing and management, and recommendations should be implemented in conjunction with them, where appropriate.

Reproducibility

Key Challenge #4: The reproducibility of scientific findings requires that digital research data be searchable and accessible through documented protocols or methods.

Data are heterogeneous, often classified and cited with disparate conventions, and housed in distributed and autonomous repositories. Metadata standards have been identified as a means to enable data sharing, data management, and federated search functions across datasets. The standardization of data definitions, data formats, code for data analysis, and citation practices can support comparison across heterogeneous and autonomous datasets, thereby enabling scientific reproducibility. Standardization of metadata, persistent identifiers, and interoperable hardware and software systems allow diverse research communities to access data outside of their fields. Standards for descriptive, administrative, and structural metadata help establish a common framework for understanding the semantic meaning of data and addressing the heterogeneity of datasets generated by scientists and engineers. Metadata can also house information about the provenance of data and history of use, enhancing data reliability, reproducibility, and attribution. Additional stakeholder engagement is needed to improve the harmonization of metadata

standards across disciplines and establish standards that account for the continual evolution of hardware and software. NSF should also consider the use of unique identifiers for researchers and human subjects so that data associated with individuals can be tracked across domains, enabling attribution.

Education, Training, and Workforce Development

Key Challenge #5: New jobs and areas of expertise are emerging in response to the evolving role of data in science and engineering, yet opportunities for education, training, and workforce development are not fully recognized and supported.

The proliferation of shared, interoperable data creates new computational and data-enabled science and engineering research opportunities that require the support of trained experts and researchers. Training and education will help to enable broad access and use of digital research data for researchers, as well as the general public, through exposure to large-scale, distributed datasets and the principles of effective data sharing and management. Training should be multidisciplinary and designed for users with varying needs and backgrounds.

Sustaining, managing, and analyzing the expanding data collections for science and engineering research may necessitate the establishment of new fields of study and professional career paths. As science dependent on large-scale datasets becomes more common, a field of content and computational expertise will emerge. Data scientists and curators should be supported by funding agencies and by their home institutions by providing pathways for advancement to tenure and other reward mechanisms.

Cyberinfrastructure

Key Challenge #6: Cyberinfrastructure advances need to be deployed rapidly and supported appropriately to account for the expanding scale, scope, and complexity of science and engineering data collections.

The global science and engineering research enterprise produces and relies on a large, growing volume of heterogeneous and multifaceted digital data, allowing researchers to investigate increasingly complex research questions. The scale of data produced is growing rapidly from terabytes to petabytes and is estimated to reach exabytes in the near future as advances in computing hardware and software enable more research founded on large-scale datasets. As the volume of data increases, so too does the diversity of datasets that are critical to addressing increasingly complex problems. Geographically distributed collaborative research teams, computing resources, and large-scale datasets require robust cyberinfrastructure, including supercomputing resources, cloud computing, fiber optic networks and highly trained personnel, to conduct research and manage, retrieve, analyze, and share results. Large datasets pose specific challenges for geographically distributed research, as the bandwidth and storage space required to access and download data are often unavailable to individual researchers.

Cyberinfrastructure,[4] including researchers and highly trained technical personnel, must be expanded in order to adequately support science and engineering research that relies on large-

scale datasets and massive computational power. Enhanced access to shared data and computing resources, coupled with the fiber optic networking, may help reduce local costs, increase access to valuable datasets and analytical resources, and provide data storage services. Furthermore, the development of sustainable and dynamic cyberinfrastructure, designed to support collaboration, will allow for interoperability and accessibility, further expanding the ability of researchers to utilize digital research data.

Longevity and Sustainability

Key Challenge #7: Data stewardship is critical to the longevity and sustainability of data sharing and management throughout the data lifecycle, but it is unclear where the responsibilities for this effort lie.

Effective data storage, preservation, and stewardship are necessary to ensure the longevity and sustainability of data for the benefit of the science and engineering research enterprise. First, relevant stakeholder communities must develop mechanisms to determine what data should be stored, how it should be inventoried, and the appropriate preservation policies to maintain the integrity of datasets. Second, mechanisms must be developed to ensure data quality, access, and interoperability. Third, data must be curated to allow for effective stewardship and efficient data discovery through the development of user interfaces, the taxonomic structuring of data and datasets, and keyword and associative search functions.

Strategic partnerships between key stakeholder communities should be developed to collectively support the development of effective data repositories and stewardship policies. Funding agencies, university-based research libraries, disciplinary societies, publishers, and research consortia should distribute responsibilities that address the establishment and maintenance of digital repositories. These roles and responsibilities should be articulated and elaborated to alleviate uncertainty about where responsibility lies and how to meet broad stewardship needs. NSF will play an important role in aligning the interest of many stakeholder groups to coordinate and harmonize their various approaches to data sharing and management.

Key Challenge #8: Data stewardship must allow for broad and timely access to data.

The effective, continued, and extended use of data relies on appropriate and timely access by the broader research community. Storing and indexing data that can be accessed by a wide range of people is central to a democratic vision of science and engineering. Diverse science and engineering communities of practice should develop and implement appropriate data policies that address their needs, while considering the potential use of domain-specific data across the broader research community. An embargo period, or the release of data after a specified amount of time, may be necessary for researchers who perform time-intensive data collection. Furthermore, the maintenance of digital data repositories should be considered in research plans in order to sustain access to data.

Key Challenge #9: Long-lived data require long-term business models that ensure data stewardship.

The volume, complexity, and heterogeneity of data have associated processing, storing, archiving, and maintenance costs that are currently not well understood. It is unclear if current levels of support for such costs will be adequate, and it is unclear which stakeholders should ultimately be responsible for specific aspects of funding. Also, there is concern that inadequate funding commitments to material support for data repositories may lead to the uneven stewardship and orphaned repositories. New long-term business models may be required to address these and other risks and to stimulate funding to meet these needs of storing, preserving, and curating data collections to support the current and future science and engineering research community.

Ethical and Legal Implications

Key Challenge #10: Access to confidential data poses ethical and legal challenges.

Ethical and legal implications pose particular challenges for research involving the collection of sensitive data. In such cases, a balance must be found between providing appropriate protective measures and maintaining confidentiality while minimizing the constraints for sharing and re-using data. Operating securely across servers also poses technical challenges. Furthermore, researchers must be confident when they share data that they will be properly attributed and the provenance of the data is assured.

Research and training to promote data access that preserves privacy can contribute to developing clear guidelines on confidential data. If data are restricted to some researchers, certification criteria must be established. In addition, new data licensing mechanisms can preserve intellectual property rights and provide researchers with incentives to make their data public.

Recommendations for the National Science Foundation

The Board's Committee on Strategy and Budget Task Force on Data Policies proposes five digital research data policy recommendations for NSF and its associated research communities. These recommendations proceed from the Task Force on Data Policies' Statement of Principles, a set of seven principles approved by the Board on February 16, 2011 (Appendix B). Recommendations are organized under four areas: commitment to sharing; reproducibility; education, training, and workforce development; and longevity and sustainability. Although these five recommendations do not cover all key challenges associated with digital research data sharing and management, they represent a further step in NSF's implementation of strong digital research data policies that benefit the science and engineering research community.

Commitment to Digital Research Data Sharing

Digital research data policy issues involve multiple stakeholders with varied responsibilities, and successful policy development and implementation necessitate broad stakeholder involvement. One-size-fits-all solutions cannot adequately address most digital research data policy issues because each research community is best suited to address the nuances of its own data.

As a funding agency that supports basic research in multiple areas of science and engineering, NSF is uniquely positioned to provide digital research data policy leadership and promote forward-thinking digital research data policies among diverse scientific communities. NSF intends to continue providing leadership on digital research data sharing and management to this broad array of stakeholders.

Recommendation 1: *Provide leadership to Federal agencies and other national and international stakeholders in the development and implementation of digital research data policies, including the promotion of individual scientific communities to establish data sharing and management practices that align with NSF data policies.*

Reproducibility of Digital Research Data

Openness and transparency are critical to continued scientific and engineering progress and to building public trust in the nation's scientific enterprise. This applies to all materials necessary for verification, replication, and interpretation of results and claims associated with scientific and engineering research.

Reproducibility is critical to the pursuit of modern science and is a central value for scientific communities enabling both the validation and extension of research.[5] However, reproducibility does not mean the exact replication of results; rather, it means enabling the replication of research (experiments, models, simulations, etc.) to the extent that other researchers can attempt to reproduce previous results with fidelity. For example, the journals *Nature* and *Science* require authors to provide data for the purposes of replicating and verifying conclusions. *Science* goes as far as to state that data must be shared to "extend the conclusions of the manuscript."[6]

Using the Data Management Plan to determine the timeline for initiating the data sharing process recognizes the rights and responsibilities of investigators. Investigators should have the opportunity to analyze their data and publish their results within a reasonable time.

Recommendation 2: *Consistent with the digital research data generated in research projects, require grantees to make both the data and the methods and techniques used in the creation and analysis of the data accessible for the purposes of building upon or verifying figures, tables, findings, and conclusions in peer reviewed publications coincident with publication.*[7] *Similar requirements are appropriate when data are requested for the purpose of extending the scientific conclusions through further research. Data should be shared using persistent electronic identifiers, which enable automatic attribution of authors and award funding. Research data related to human subjects, proprietary content, and national security represent cases requiring appropriate care consistent with national and international regulations.*

Education, Training, and Workforce Development

The proliferation of shared, interoperable data and the implementation of the previous four recommendations will help create new computational and data-enabled science and engineering research opportunities that will require the support of trained experts and researchers. The importance of computational science was clearly identified in a 2005 President's Information Technology Advisory Committee report, which adopted a broad definition "to underscore the reality that harnessing software, hardware, data, and connectivity to help solve complex problems necessarily draws on the multidisciplinary skills represented in the computing infrastructure as a whole."[8] NSF currently promotes and supports computational and data-enabled science and engineering, and continued support will foster career opportunities for researchers and specialists in the fields of computer science, information science, and science and engineering.

Recommendation 3: *Continue to expand the support of computational and data-enabled science and engineering researchers and cyberinfrastructure professionals to take advantage of shared, accessible data and to forward emerging science.*[9]

Longevity and Sustainability of Digital Research Data

Stakeholder roles, responsibilities, and resources must be clearly identified and proactively established to support sharing, management, preservation, and long-term digital research data accessibility.

Recommendation 4: *Convene a panel of stakeholders to explore and develop a range of viable long-term business models and issues related to maintaining digital data and provide a key set of recommendations for action.*

Digital research data are long-lived, necessitating long-term commitments from stakeholders committed to the preservation and stewardship of that data. Sufficient standards (e.g., interoperability protocols) and long-term business models currently do not exist to meet these responsibilities. However, some communities of researchers have developed structures to share

and maintain domain-specific data. For example, the Inter-university Consortium for Political and Social Research (ICPSR) archives over 500,000 data files of social science research and offers courses related to the design, analysis, and curation of such data.[10]

Successful digital research data sharing and management plans depend, in part, on adequate consideration of funding, resources, and structural issues that may either facilitate or impede acceptance and implementation. These plans are especially important for small research institutions and research grants that may not have the resources available to share and manage long-lived data. Thus, just as a single data sharing and management policy will not apply to all research communities, a one-size-fits-all business model will not apply to all institutions and awards.

Recommendation 5: *Further the expansion of sustainable data management, including preservation and curation of pre-existing and newly generated long-lived data, by encouraging development and implementation of data sharing infrastructure and long-term business models that encompass the range of research communities, research institutions, and research grants, as outlined in recommendations of the panel formed to explore these issues in Recommendation 2.*

Conclusion

The increasing ease with which digital research data are gathered, processed, analyzed, and disseminated shapes the conduct and progress of science and engineering research. NSF must be prepared to meet the accessibility and management challenges that the proliferation of digital research data poses. Additionally, NSF should implement the Board's guidance to ensure that stakeholders, policies, infrastructure, and expertise are best positioned to support current and future science and engineering research. NSF's leadership in digital research data sharing and management will promote the U.S. science and engineering enterprise and support U.S. scientific progress.

Selected Bibliography

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, 2010.

Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, and National Academy of Sciences. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington, D.C.: National Academies Press, 2009.

High Level Expert Group on Scientific Data. Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data. European Commission, 2010.

Holdren, John P. "Scientific Integrity." Washington, D.C.: Office of Science and Technology Policy, 2010.

National Science and Technology Council, Committee on Science, and Interagency Working Group on Scientific Collections. *Scientific Collections: Mission-Critical Infrastructure for Federal Science Agencies*. Washington, D.C.: Office of Science and Technology Policy, 2009.

National Science Board. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century National Science.* Arlington, VA: National Science Foundation, 2005 (NSB-05-40).

National Science Foundation. *Changing the Conduct of Science in the Information Age: Summary Report of Workshop Held on November 12, 2010*. Arlington, VA: National Science Foundation, June 28, 2011.

National Science Foundation. *Proposal and Award Policies and Procedures Guide*. Arlington, VA: National Science Foundation, 2010 (NSF-11-1).

National Science Foundation. *Advisory Committee for Cyberinfrastructure, and Task Force on Data and Visualization*. Final Report. Arlington, VA: National Science Foundation, 2011.

Organization for Co-operation and Development. *OECD Principles and Guidelines for Access to Research Data from Public Funding*, 2007.

Panel on Communicating National Science Foundation Science and Engineering Information to Data Users, and National Research Council. *Communicating National Science Foundation Science and Engineering Information to Data Users: Letter Report*. Washington, D.C.: National Academies Press, 2011.

President's Information Technology Advisory Committee. *Computational Science: Ensuring America's Competitiveness.* Report to the President. Washington D.C.: National Coordination Office for Information Technology Research and Development, 2005.

National Science and Technology Council, Interagency Working Group on Digital Data. *Harnessing the Power of Digital Data for Science and Society*, 2009.

Appendix A: Task Force Charge

NSB -10-60
August 26, 2010

Committee on Strategy and Budget (CSB)
Task Force on Data Policies (DP)

CHARGE

Background

The increasing ease of gathering large amounts of varied data – including digital data, research specimens, artifacts, etc. – and funding of large-scale collaborative projects, have caused the broad policy issues surrounding the management of scientific and engineering research data to become critically important. How data collected with National Science Foundation (NSF) funding are shared and managed to ensure broad, timely, and long-term availability and accessibility to the entire research community is an important issue. A determination of what, if any, NSF policies related to data sharing and management would be in the best interests of the Nation's scientific and engineering enterprise warrants careful examination by the National Science Board (NSB).

Significant policy debate on this broad set of issues is ongoing at both national and international levels, with many stakeholders and organizations involved. Past and ongoing efforts by the Board, NSF as a whole, and other organizations could inform the current effort. In addition to reports from the National Science and Technology Council (NSTC) and the National Research Council (NRC), [1] especially relevant to this effort is the NSB Report *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, September 2005 (NSB-05-40).

Given that sharing and managing research data are problematic for the entire international research community, the NSB, in taking up this topic, has a real opportunity to contribute productively to a significant and ongoing policy discussion. The policy issues surrounding data are critically important at both national and international levels and for NSF as we carry out our mission to promote the progress of science.

The issues surrounding data sharing and management - of which there are many - are complex and include broad and timely access to data, sustainability of data (particularly of digital data), the cost burdens associated with data management, and openness of data generated with taxpayer dollars, to name a few.

Charge to the NSB CSB Task Force on Data Policies

The NSB CSB Task Force on Data Policies was established at the February 3-4, 2010 NSB meeting with the charge of further defining the issues and outlining possible options to make the use of data more effective in meeting NSF's mission.

Membership on the NSB CSB Task Force on Data Policies: Dr. José-Marie Griffiths, chairman, and Drs. Mark Abbott, Camilla Benbow, John Bruer, Bud Peterson, Diane Souvaine, Thomas Taylor, and Mr. Arthur Reilly, members, with Executive Secretary Dr. Philip Bogden, NSF. NSF Liaison members on the Task Force are Drs. Myron Gutmann (Assistant Director, SBE) and Ed Seidel (Assistant Director, MPS).

Process and Strategies

This work plan describes the process and strategies for gaining input from stakeholders regarding their understanding of NSF data policies along with current data sharing and management practices. The stakeholder groups are both internal and external to NSF and mainly include research communities and their institutions (external) and NSF program officers (internal). The input gained from this study will inform the task force on how best to proceed with follow-up action, which includes detailing the findings, deliberating recommendations, discussing recommendations with NSF leadership, and working together to find the best solutions.

The first step for the Task Force is to hear from the NSF Data Working Group. Then it will work with the Board and NSF senior staff to further define the issues and outline possible options to make the use of data more effective in meeting NSF's mission. During this period, the Task Force will solicit input widely from the research and stakeholder communities and may solicit special studies as appropriate.

The Task Force's strategy on developing Data Policies is multi-phased:

- NSF updated implementation of long-standing data policy – the Data Management Plan requirement – should go into effect in January 2011 and will become a starting point for the Task Force. The Task Force will monitor the impact of this implementation change in order to inform a review of NSF policy.

- Considering issues of data policy, Open Data movements, and related issues, the Task Force will then develop a "Statement of Principles."

- Provide guidance to subsequent Board efforts to develop specific actionable policy recommendations focused, initially, on NSF, but that could potentially promulgate through other Federal agencies in a national and international context.

This effort requires significant background material on current NSF data policies; data policies at other Federal agencies; data policies at international counterparts to NSF; and the views of NSF awardees on the value of data policies and the impact on the administrative burden. A survey of researchers/PIs may also need to be considered.

The steps in the process are as follows:

1. Receive update from Dr. Edward Seidel on NSF's plans to enhance the enforcement of existing data policy.

2. Determine the way the current data policies, and their instructions, are interpreted and utilized by both proposers and NSF program staff. Solicit input of Program Directors.

3. Interviews with key stakeholders conducted by Task Force leads.

4. Prepare a Statement of Principles.

5. Assess further need for NSB study.

Attached are a Proposed Timeline and an appendix of possible Data Policy Issues.

Data Policies Task Force Timeline

Date	Task
April – May 2010	Task Force members consider the questions they want answered; the information necessary to attain the answers; and the means by which to gather the information
May 4-5, 2010	Task Force meeting at Board meeting to discuss next steps in proceeding with internal and external research
May – August 2010	Develop a Statement of Principles
August 25-26, 2010	Task Force meeting at Board meeting to approve charge, review and revise plan, review draft Statement of Principles, discuss plans for workshop of key stakeholders to be held in winter
August – Sept. 2010	Review and compile findings
September 2010	Offsite Board meeting/Informal discussion of progress
Sept. – Dec. 2010	Proceed with internal and external research and begin to formulate recommendations
Dec. 1-2, 2010	Task Force meeting at Board meeting to review and discuss results of research
Dec. – Feb. 2011	1- or 2-day Workshop of key stakeholders
Feb. – May 2011	Draft final report with findings and recommendations for data policies

[1] NSTC Interagency Working Group on Digital Data, *Harnessing the Power of Digital Data for Science and Society* (January 2009); and NRC's *Ensuring the Utility and Integrity of Research Data in a Digital Age* (2009).

Appendix: Possible Data Policy Issues

1. Internal policies that could be addressed include:

    a. Defining what constitutes the release of "complete" data. Would complete data release include the original, "raw" data; cleaned-up, publication-ready data, along with the methods for clean-up; publication-ready data with the meta-data necessary to reproduce any interpretations of the data; raw data with software to make it usable to others; data organized in a way that is inter-operable to some standard; etc.?

    b. Defining what types of "data" are to be shared - should we add specimens, samples, etc.?

    c. Defining what "sharing" entails - what is expected of principal investigators and awardee institutions? Who is responsible for ensuring persistent access?

    d. Defining good data management/curation practices.

    e. Timeline for release of data (e.g., a certain time period after collection, after publication of results, etc.).

    f. Timeframe for continued availability of data - forever?

    g. Balance between acknowledging variations in the expectations of different disciplines and research communities regarding the proprietary nature of data and setting agency-wide data policies.

    h. Potential NSF guidelines to awardees relating to management of data that could, for example, require awardees to develop a data management plan with certain components that is peer-reviewed and considered part of the terms and conditions of the award.

    i. Particularly significant impact of the data policies of NSF-funded large facilities and centers on whole research communities. Merit, if any, of including data policies as part of the site-visits and design reviews of large centers and facilities.

    j. NSF role, if any, in setting standards for meta-data requirements. If processed data is made available, determining what the requirements should be for making available the work processes performed on the data so that its provenance can be established.

    k. NSF role, if any, in setting standards for data formats for sharing and exchange, as well as for long-term curation.

    l. NSF role, if any, in setting requirements for data "publishing" or deposit.

m. NSF role, if any, in off-setting or funding the administrative burden placed on awardee institutions and principal investigators by any required data management policies.

2. Technical considerations in archiving and ensuring the accessibility of many types of data that are becoming more and more complex. Just as "publications" are often no longer exclusively a printed piece of paper and often involve supplemental material provided in a variety of electronic media, "data" may not be simply original data or measurements, but raw data in the context of its associated meta-data.

3. What proprietary rights, if any, are appropriate for a principal investigator relating to data retention and usage?

4. Accessibility of data for evidence-based policy development.

5. Identification of the appropriate party or parties who should be responsible for ensuring the long-term archiving and curation of data, both for the cost burden and implementation. Possibilities include NSF, awardee institutions, principal investigators, a combination of the above, etc.

6. Merit, if any, of a national repository (or multiple repositories) for data and the appropriateness of NSF's assisting in funding such repositories, helping set standards for such an effort, and/or requiring awardees to deposit data in such repositories.

7. Impact of the NSF DataNet program on data management.

8. International complexities, particularly for large facilities with international partnerships.

9. Legal complexities.

10. Potential overlap of policy issues between the curatorship of physical specimens and the management of large, and often digital, datasets.

Appendix B: Statement of Principles

NSB -11-20
February 16, 2011

National Science Board
Committee on Strategy and Budget
Task Force on Data Policies

Statement of Principles

The progress of science and engineering has always been dependent on the collection of data through observation, experimentation and, more recently, computation. A core expectation of the scientific process is the documentation and sharing of results along with the underlying data and methodology, thereby allowing others to verify data, reproduce results, validate interpretations, and build upon previous work. The processes of peer review and formal publication have been pillars of scientific openness for centuries.

Recently, the increasing ease with which data can be gathered, processed, analyzed, and disseminated and funding of large-scale collaborative projects have greatly expanded the scale, scope and complexity of science and engineering data collections and highlighted the need for improved data policies. Furthermore, NSF has a commitment to broadening the participation of those involved in scientific and engineering research and education and access to data is intricately linked to this commitment. The accessibility of data created with NSF funds represents an opportunity to maximize the size and diversity of the user community for data.

The NSB is committed to the development, implementation and assessment of data sharing and data management policies for NSF-funded activities. This includes the sharing of results, data, physical collections and other supporting materials created or gathered in the course of NSF-funded work. The current policy appears in Chapter VI, Section D, of the NSF Proposal and Award Policies and Procedures Guide (pages VI-8 and VI-9 of NSF Document 10-1):

> ***4. Dissemination and Sharing of Research Results***
>
> *a. Investigators are expected to promptly prepare and submit for publication, with authorship that accurately reflects the contributions of those involved, all significant findings from work conducted under NSF grants. Grantees are expected to permit and encourage such publication by those actually performing that work, unless a grantee intends to publish or disseminate such findings itself.*
>
> *b. Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Privileged or confidential information should be released only in a form that protects the privacy of individuals and subjects involved. General adjustments and, where essential, exceptions to this sharing expectation may be specified by the funding NSF Program or Division/Office for a*

particular field or discipline to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate the legitimate interest of investigators. A grantee or investigator also may request a particular adjustment or exception from the cognizant NSF Program Officer.

c. Investigators and grantees are encouraged to share software and inventions created under the grant or otherwise make them or their products widely available and usable.

d. NSF normally allows grantees to retain principal legal rights to intellectual property developed under NSF grants to provide incentives for development and dissemination of inventions, software and publications that can enhance their usefulness, accessibility and upkeep. Such incentives do not, however, reduce the responsibility that investigators and organizations have as members of the scientific and engineering community, to make results, data and collections available to other researchers.

e. NSF program management will implement these policies for dissemination and sharing of research results, in way appropriate to field and circumstances, through the proposal review process; through award negotiations and conditions; and through appropriate support and incentives for data cleanup, documentation, dissemination, storage and the like.

The Board is working with NSF leadership and other science and engineering stakeholders to frame and examine current and emerging issues associated with science and engineering data and develop relevant policies. This preliminary statement of principles will guide these efforts.

Note:

1. *Openness and transparency are critical to continued scientific and engineering progress and to building public trust in the nation's scientific enterprise. This applies to all materials necessary for verification, replication and interpretation of results and claims, associated with scientific and engineering research.*

A strong statement about openness and transparency is an important first step.

2. *Open Data[1] sharing is closely linked to Open Access[2] publishing and they should be considered in concert.*

This principle is included because there need to be bidirectional pointers between peer-reviewed and other published literature and the available supporting materials. All

---

[1] *Open Data* refers to the concept and practice that certain data be made freely available, without restrictions, for no more than the cost of reproduction and distribution.

[2] *Open Access publishing* refers to the free availability of publications (either immediately upon publication or within a specified time period) on the public internet, permitting users to perform a variety of functions – read, download, copy, distribute, print, search, link, etc.

these materials need be made discoverable and the discoverability will require relevant metadata, ontologies, standards, etc., to be applied.

3. *The nation's science and engineering research enterprise consists of a broad array of stakeholders[3], all of which should participate in the development and adoption of policies and guidelines.*

It is important to recognize the many different stakeholders and their respective roles and current/potential responsibilities. Their involvement in the development and implementation of policies is crucial to successful implementation.

4. *It is recognized that standards and norms vary considerably across scientific and engineering fields and such variation needs to be accommodated in the development and implementation of policies.*

The statement will be important to signal that we do not anticipate a "one size fits all" solution.

5. *Policies and guidelines are needed for open data sharing which in turn requires active data management.*

Our primary goal is the sharing of data and other supporting materials. Once available for sharing, there is a need for proactive management and preservation for long-term accessibility. The policies, roles and responsibilities vary across these different but related functions.

6. *All data and data management policies must include clear identification of roles, responsibilities and resourcing.*

These 3 R's are often omitted from consideration while the more technical aspects of policies are developed. However, in the increasingly complex scientific and engineering research enterprise, the likelihood of success will improve with consideration of the socio-economic issues that can impede or facilitate acceptance and implementation.

7. *The rights and responsibilities of investigators are recognized. Investigators should have the opportunity to analyze their data and publish their results within a reasonable time.*

---

[3] Stakeholders include researchers, research institutions, research funders, various government agencies, professional societies, publishers, data repositories, data and metadata libraries and archives, and public advocacy groups.

Appendix C: Charge to Workshop Invitees

In February 2010, The National Science Board (NSB) established a Task Force on Data Policies under the Committee on Strategy and Budget (CSB) with the charge of further defining the issues and outlining possible options to make the use of data more effective in meeting the National Science Foundation (NSF) mission. The NSB website defines the charge, membership and goals of the Task Force that is convening the workshop.

As an invitee to the March 27-29, 2011 workshop, the Task Force requests a 1-3 page white paper that describes your perspective on the issues presented below. You are also encouraged to use this opportunity to raise other questions if you have them. Your comments will be shared with other invitees in advance of the workshop and will become part of the record. Please send your white paper to the NSB Office staff (bdahl@nsf.gov) by Friday, March 18.

Background

The issues surrounding data sharing and management - of which there are many - are complex and include broad and timely access to data, sustainability of data (particularly of digital data), the cost burdens associated with data management, and openness of data generated with taxpayer dollars, to name a few. Significant policy debate on this broad set of issues is ongoing at both national and international levels, with many stakeholders and organizations involved. Past and ongoing efforts by the Board, NSF, and other organizations will inform the current effort. In addition to reports from the National Science and Technology Council (NSTC)[1] and the National Research Council (NRC)[2], especially relevant is the 2005 NSB report on *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century.*[3]

The Task Force's strategy on developing Data Policies is multi-phased:

Phase 1. As a starting point, the Task Force is monitoring the impact of a recently updated NSF-wide implementation of long-standing data policy – the Data Management Plan requirement – that went into effect in January 2011.

Phase 2. The NSB recently approved the Task Force's "Statement of Principles" that considers various data policies, Open Data movements, and related issues.

---

1 NSTC Interagency Working Group on Digital Data, *Harnessing the Power of Digital Data for Science and Society* (January 2009).

2 NRC's *Ensuring the Utility and Integrity of Research Data in a Digital Age* (2009).

3 NSB-05-40, September 2005.

Phase 3. The Task Force will provide guidance to subsequent Board efforts to develop specific actionable policy recommendations focused, initially, on NSF, but that could potentially promulgate through other Federal agencies in a national and international context.

This workshop provides the basis for Phase 3. The goal of the workshop is to provide input to help the Task Force develop a report with recommendations for the NSF and the nation, anticipating wide variation across many science domains and stakeholder interests.

Organization of Workshop and White Papers
The workshop will be organized into the following sessions/themes. Each will involve discussion of the related guiding questions. Please organize your own white paper accordingly.

Session I: The Vision of Data-Intensive Science

Guiding questions: What are some of the defining characteristics of data-intensive science? What are the goals for enabling re-use and re-purposing of data? What new opportunities and new types of science have yet to be realized? These questions build upon the vision for a new NSF-wide program in computational and data-intensive science.

Session II: Reproducibility -- First Step & Guiding Principle

Guiding questions: Reproducibility starts to scope the problem and drives all sorts of related issues (curation, cost, etc.). What does this mean for types of discovery that need data sharing (e.g., medical research, such as work on Alzheimer's disease)? What are the implications for data publishing and data citation? What are the implications for simulation and software? What constitutes the release of "complete" data? Would complete data release include the original, "raw" data; cleaned-up, publication-ready data, along with the methods for clean-up; publication-ready data with the meta-data necessary to reproduce any interpretations of the data; raw data with software to make it usable to others; data organized in a way that is interoperable to some standard; etc.?

Session III: Exemplars, Lessons Learned

Guiding questions: What has been your experience? What types of incentives can be created? How has data publication impacted innovation? Examples include the Virtual Observatory, Interuniversity Consortium for Political and Social Research (ICPSR), Protein Data Bank, etc.

Session IV: Impacts

Guiding questions: What are the measurable impacts? What is the early experience with the NSF-wide requirement for Data Management Plans? What are the impacts on research universities? What are the international complexities, particularly for large facilities with international partnerships? What are the legal complexities? What is the potential for overlap of policy when comparing the curatorship of physical specimens and the management of large, and often digital, datasets?

Session V: Policy Issues

Guiding questions: Frame the issues for institutions, government agencies, publishers and any other stakeholders. What are the relative merits of various types of repositories for data? How should the various repositories be funded? To what extent should NSF assist in development and adoption of standards for such efforts? To what extent should deposit in repositories be required of awardees?

Appendix D: Participant Agenda

NSB/CSB/DP-11-2
March 25, 2011

National Science Board
Expert Panel Discussion on Data Policies
March 27-29, 2011
Arlington, Virginia

PARTICIPANT AGENDA

Sunday, March 27
F. Scott Fitzgerald Ballroom A
The Westin Arlington Gateway Hotel
Participant Orientation Reception and Dinner

5:30 p.m.Reception and Cash Bar

6:30 p.m.Dinner

7:00 p.m.Keynote Address
Making Open Science Real
Adam Bly, Seed Media Group

Monday, March 28
The NSB Board Room
National Science Foundation, Room 1235

8:00 Welcome, Board Processes, and Participant Introductions

Welcome from **Dr. Ray M. Bowen,** Chairman of the National Science Board

Welcome from **Dr. Diane L. Souvaine,** Chairman of the Committee on Strategy and Budget, NSB

Participant Introductions, Workshop Overview and Goals, **Dr. José-Marie Griffiths,** Chairman, Task Force on Data Policies, Committee on Strategy and Budget, National Science Board

8:20 – 10:00 Session I: The Vision of Data-Intensive Science

Guiding questions: What are some of the defining characteristics of data-intensive science? What are the goals for enabling re-use and re-purposing of data? What new opportunities and new

types of science have yet to be realized? These questions build upon the vision for a new NSF-wide program in computational and data-intensive science.

Moderator: Dr. Diane L. Souvaine

Panelists:

- **Roberta Balstad,** Columbia University
- **Francine Berman,** Rensselaer Polytechnic Institute
- **Michael Lesk,** Rutgers, The State University of New Jersey
- **Maryann Martone,** University of California San Diego

10:00 – 10:15 Break

10:15 – 12:00 Session II: Reproducibility, First Steps and Guiding Principles

Guiding questions: Reproducibility starts to scope the problem and drives all sorts of related issues (curation, cost, etc.). What does this mean for types of discovery that need data sharing (e.g., medical research, such as work on Alzheimer's disease)? What are the implications for data publishing and data citation? What are the implications for simulation and software? What constitutes the release of "complete" data? Would complete data release include the original, "raw" data; cleaned-up, publication-ready data, along with the methods for clean-up; publication-ready data with the meta-data necessary to reproduce any interpretations of the data; raw data with software to make it usable to others; data organized in a way that is interoperable to some standard; etc.?

Moderator: Mr. Arthur K. Reilly

Panelists:

- **Timo Hannay,** Digital Science
- **Brooks Hanson,** Science Magazine
- **Randall LeVeque,** University of Washington
- **Victoria Stodden,** Columbia University

12:00 – 1:00 Lunch

Lunch will be delivered for participants who order in advance.

12:30 Lunch Presentation: High Performance Cyberinfrastructure is Needed to Enable Data-Intensive Science and Engineering

Dr. Larry Smarr, Harry E. Gruber Professor, Department of Computer Science and Engineering, University Of California, San Diego Jacobs School of Engineering; and Director, California Institute for Telecommunications and Information Technology

1:00 – 3:00 Session III: Exemplars, Lessons Learned

Guiding questions: What has been your experience? What types of incentives can be created? How has data publication impacted innovation? Examples include the Virtual Observatory, Interuniversity Consortium for Political and Social Research, Protein Data Bank, etc.

Moderator: **Dr. Camilla P. Benbow**

Panelists:

- **George Alter,** Interuniversity Consortium for Political and Social Research
- **David Lynn,** Wellcome Trust
- **Reagan Moore,** University of North Carolina at Chapel Hill
- **Alex Szalay,** The Johns Hopkins University

3:00 – 3:15 Break

3:15 – 5:30 Session IV: Impacts

Guiding questions: What are the measurable impacts? What is the early experience with the NSF-wide requirement for Data Management Plans? What are the impacts on research universities? What are the international complexities, particularly for large facilities with international partnerships? What are the legal complexities? What is the potential for overlap of policy when comparing the curatorship of physical specimens and the management of large, and often digital, datasets?

Moderator: **Dr. Mark R. Abbott**

Panelists:

- Ravi Bellamkonda, Georgia Tech
- Tony Hey, Microsoft Research
- Michael Huerta, National Institute of Mental Health
- Michael Mabe, International Association of Science, Technical & Medical Publishers

5:30 Dinner on your own

Tuesday, March 29
The NSB Board Room
National Science Foundation, Room 1235

8:30 National Science Foundation Perspective

Remarks from officials from the National Science Foundation

8:45 – 10:30 Session V: Policy Issues

Guiding questions: Frame the issues for institutions, government agencies, publishers and any other stakeholders. What are the relative merits of various types of repositories for data? How should the various repositories be funded? To what extent should NSF assist in development and adoption of standards for such efforts? To what extent should deposit in repositories be required of awardees?

Moderator: Dr. José-Marie Griffiths

Panelists:

- **Daniel Atkins,** University of Michigan
- **Mike Keller,** Stanford University
- **Celeste Rohlfing,** White House Office of Science and Technology Policy
- **Ann Wolpert,** Massachusetts Institute of Technology

10:30 – 10:45 Break

10:45 – 11:00 Public Comment Period

Dr. José-Marie Griffiths will take a few comments and questions from the audience present at the workshop

11:00 – 12:30 Session V: Policy Issues (continued)

Discussion by Task Force Members and Stakeholders

12:30 Adjourn

Appendix E: Workshop Participants

George Alter, Interuniversity Consortium for Political and Social Research
Daniel Atkins, University of Michigan
Roberta Balstad, Columbia University
Ravi Bellamkonda, Georgia Tech
Francine Berman, Rensselaer Polytechnic Institute
Adam Bly, Seed Media Group
Steve Breckler, American Psychological Association
Joe Bredekamp, NASA
Steve Goff, iPlant Collaborative
Chris Greer, White House Office of Science and Technology Policy
Timo Hannay, Digital Science
Brooks Hanson, Science Magazine
Fred Heath, The University of Texas at Austin
Tony Hey, Microsoft Research
Michael Huerta, National Institute of Mental Health
Mike Keller, Stanford University
Michael Lesk, Rutgers, The State University of New Jersey
Randall LeVeque, University of Washington
David Lynn, Wellcome Trust
Michael Mabe, International Association of Science, Technical & Medical Publishers
Maryann Martone, University of California San Diego
Kevin Marvel, American Astronomical Society
Reagan Moore, University of North Carolina at Chapel Hill
Sethuraman Panchanathan, Arizona State University
Celeste Rohlfing, White House Office of Science and Technology Policy
Bernard Schutz, Max Planck Society
Larry Smarr, University Of California, San Diego
Victoria Stodden, Columbia University
Alex Szalay, The Johns Hopkins University
Crispin Taylor, American Society of Plant Biologists
John Vaughn, Association of American Universities
John Wilbanks, Creative Commons
Ann Wolpert, Massachusetts Institute of Technology

Mark Abbott, National Science Board
Camilla Benbow, National Science Board
Ray Bowen, Chairman, National Science Board
José-Marie Griffiths, National Science Board, Chairman of the Task Force on Data Policies
Louis Lanzerotti, National Science Board
Art Reilly, National Science Board
Diane Souvaine, National Science Board

Alan Blatecky, Office of Cyberinfrastructure, NSF
Myron Gutmann, Directorate for Social, Behavioral and Economic Sciences, NSF
Cora Marrett, Office of the Director, National Science Foundation
Edward Seidel, Directorate for Mathematical and Physical Sciences, NSF

NSB/CSB/DP-11-5
April 25, 2011

Summary Notes on Expert Panel Discussion on Data Policies

The National Science Board (NSB) hosted an Expert Panel Discussion on Data Policies in Arlington, Virginia on March 28-29, 2011 to focus on policy issues surrounding data-intensive science, data sharing, data access, and data stewardship, and to develop key policy recommendations for the NSB to consider. The stakeholder communities represented at the expert panel discussion included active researchers from multiple disciplines, universities, research libraries, publishing companies, industry, scholarly societies, and public and private funding agencies. Six major themes emerged from the expert panel discussion and are summarized below. The bulleted lists summarize suggestions for future actions posed by workshop participants.

1. Standards and interoperability enable data-intensive science.

The primary goal for enabling re-use and re-purposing of data must be to facilitate the application of data to advance relevant scientific research. Data are heterogeneous, often classified and cited with disparate schema, and housed in distributed and autonomous databases and repositories. Standards for descriptive and structural metadata will help establish a common framework for understanding data and data structures to address the heterogeneity of datasets. Standards and conventions for persistent identifiers, unique identifiers for researchers and human subjects, and interoperable hardware and software systems will enable diverse research communities to access data from other fields of research.

- Funding agencies should reinforce expectations for sharing by supporting new norms and practices for citation and attribution to facilitate discovery of datasets and so that data producers, software and tool developers, and data curators are credited for their contributions.

- Funding agencies should work with stakeholders and research communities to support the establishment of standards that enable sharing and interoperability internationally and across disciplines through award requirements and data management plans.

- Funding agencies should work with stakeholders and research communities to support the development of persistent identifiers that enable the tracking of provenance, ensure data integrity, and facilitate citation and attribution.

2. Data sharing is an identified priority.

Data sharing supports partnerships between geographically distributed research teams. Data sharing also enables the reproducibility of scientific experimentation. Currently, data sharing is

viewed as a compliance issue in some parts of the research community, rather than as a beneficial practice.

- Ethical and legal implications pose particular challenges for research involving the collection of sensitive data from human subjects, such as medical and sociological research. A balance must be found that provides appropriate protective measures while minimizing the constraint of data for sharing and re-use.

- Data sharing policies and practices must acknowledge and work within specific disciplinary cultures, while simultaneously establishing a culture of sharing between all disciplines in the broader research community.

- Funding agencies and institutions must promote and reward exemplary projects and the implementation of data management plans. Incentives could include support for pilot projects that explore data sharing architectures.

- Data availability must be timely. Stakeholders disagreed about the utility of data embargoes and the ideal duration of restricted use. The benefits of transparency and reproducibility to the scientific enterprise may help to resolve areas of dispute.

3. Recognize and support computational and data-intensive science as a discipline.

New methods that rely on computational and curatorial expertise are being created to analyze and process the increasing volume of complex data and datasets developed across the international science and engineering enterprise. The increasing data-intensity of research is leading to new fields of study and professional careers in the computational and data-intensive sciences.

- Funding agencies and research institutions should recognize and reward computational and data scientists and data curators by providing research funding and supporting advancement to tenure.

- Funding agencies and research institutions should support the training of current researchers and staff at stakeholder institutions as well as undergraduate and graduate students entering this field.

- Funding agencies should support and reward international collaborations, as these partnerships are crucial to develop cyberinfrastructure, promote data stewardship, interoperability, and a culture of sharing on an international scale.

- New funding and economic models are needed to support data-intensive science that take into account the cost of processing, storing, archiving, and maintaining datasets. Stakeholders shared considerable uncertainty regarding which stakeholders should be responsible for specific aspects of funding.

- Stakeholders questioned which aspects of the transition toward computational and data-intensive research should be driven by funding agencies and publishers and which should be emerge from research communities.

4. Storage, preservation, and curation of data are critical to data sharing and management (data stewardship).

The expert panel identified university-based research libraries, disciplinary societies, publishers, research consortia, and funding agencies as critical stakeholder communities needed to support digital data repositories. Workshop attendees emphasized the importance of federation and data stewardship so that data can be accessed widely. Several attendees noted the importance of promoting training in data curation and informatics to support these activities.

- Stakeholders proposed that funding agencies commit to ongoing financial support for new and existing repositories. They expressed concern that inadequate support may lead to existing repositories being orphaned.

- Stakeholders suggested that funding agencies support the development of multiple as well as overlapping digital repository architectures.

- Standardized curatorial mechanisms must be created to determine what data should be stored, how data should be inventoried, and to ensure accessibility and enable efficient data discovery.

- Establish strategic partnerships between stakeholder communities to support the development of data repositories and stewardship policies. Funding agencies could support this through award requirements and data management plans.

- Stakeholders outlined several competing visions of the appropriate distribution of responsibilities for digital repositories, debating which types of organizations should house repositories, whether public repositories should be established, and the proper level of government financial support and influence.

- One participant suggested that data repositories should be independently audited in order to ensure data quality, access, and interoperability.

5. Cyberinfrastructure is necessary to support data-intensive science.

The science and engineering research enterprise produces and relies upon a large volume of heterogeneous digital data that will continue to grow as advances in computing enable more data-intensive research. The geographic distribution of collaborative research teams, computing resources, and datasets requires robust cyberinfrastructure (including shared resources and services for supercomputing and cloud computing resources, and fiber optic networking) to conduct research and manage, retrieve, analyze, and share results. Cyberinfrastructure needs include shared applications and services for analysis, visualization and simulation. Investments could encourage standardization of infrastructure to allow for interoperability and accessibility.

- Stakeholders agreed that funding agencies and research institutions should make capital investments in cyberinfrastructure. They proposed that funding agencies include additional funds to support cyberinfrastructure development in research awards; however there was no consensus on the appropriate ratio of infrastructure to research funding.

Endnotes

[1] The term *research data* is formally defined by the U.S. Office of Management and Budget in Circular A-110 as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues." This definition includes both analyzed data and the metadata that describe how those data were generated. In this instance *analyzed data* is limited to digital information that describes the outcomes of NSF-funded research, including digital images, published tables, and tables of the numbers used to create charts and graphs. Necessary *metadata* includes, but is not limited to, descriptions or suitable citations of experiments, apparatuses, raw materials, computational codes, model parameters, and input conditions. *Digital research data* is any digital data, as well as the methods and techniques used in the creation and analysis of that data, that a researcher needs to verify results or extend scientific conclusions**,** including digital data associated with non-digital information, such as the metadata associated with physical samples.

[2] NSTC Interagency Working Group on Digital Data, *Harnessing the Power of Digital Data for Science and Society* (January 2009), http://www.nitrd.gov/About/Harnessing_Power_Web.pdf.

[3] National Science Board. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century National Science.* Arlington, VA: National Science Foundation, 2005 (NSB-05-40).

[4] National Science Foundation, Cyberinfrastructure Council. *Cyberinfrastructure Vision for 21st Century Discovery*. Arlington, VA. National Science Foundation. 2007. Page 6. This report notes, "The comprehensive infrastructure needed to capitalize on dramatic advances in information technology has been termed cyberinfrastructure (CI). Cyberinfrastructure integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools." Cyberinfrastructure may also include, "professionals with expertise in algorithm development, system operations, and applications development."

[5] In 1942 renowned sociologist Robert Merton argued that a key tenet of good scientific research included accepting that scientific results are common property of the entire scientific community. See R. K. Merton, "The Normative Structure of Science," in *The Sociology of Science: Theoretical and Empirical Investigations,* ed. R. K. Merton (Chicago, IL: University of Chicago Press, 1973).

[6] See *Nature* website, Availability of data and materials, http://www.nature.com/authors/policies/availability.html, and *Science* website, General Information for Authors, Submission requirements and conditions of acceptance, and Data and materials availability, http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail.

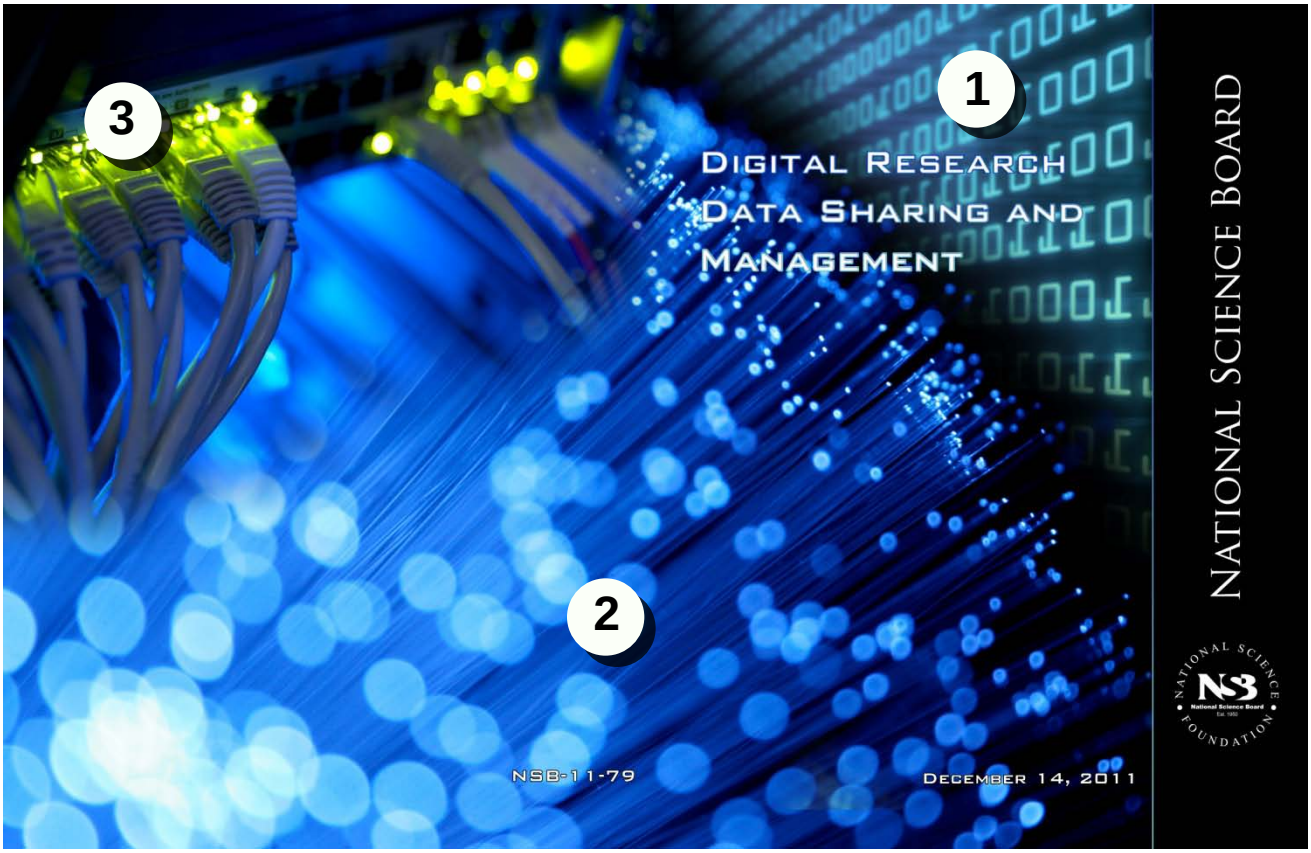
[7] As noted previously, requiring data sharing does not detract from the legal rights of grantees, nor does it require open source access or violate the Bayh-Dole Act (Patent and Trademark Law Amendments Act) or copyright laws. Furthermore, according to the NSF *Proposal and Award Policies and Procedures Guide*, data-sharing provisions can be made "to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate the legitimate interest of investigators" (p. VI-8).

[8] President's Information Technology Advisory Committee, *Computational Science: Ensuring America's Competitiveness*, Report to the President (June 2005), http://www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf.

[9] This recommendation corresponds to language from a document signed by Arden L. Bement, Jr., Director, NSF, on 22 May 2010.

[10] See ICPSR's website, http://www.icpsr.umich.edu/icpsrweb/ICPSR/.

COVER IMAGE CREDITS



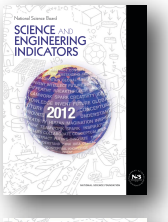
1. This is a JPEG of some binary code. The source of the picture is unknown by the designer, but it is widely used by random blogs on the internet. This particular image copy was taken from the British Broadcasting Corporation's College of Journalism Blog - via Google image search. -- A binary code is a way of representing text or computer processor instructions by the use of the binary number system's two-binary digits 0 and 1. This is accomplished by assigning a bit string to each particular symbol or instruction. For example, a binary string of eight binary digits (bits) can represent any of 256 possible values and can therefore correspond to a variety of different symbols, letters or instructions. - Wikipedia

2. This is a picture of a bunch of optical fiber (or optical fibre). Optical fiber is a flexible, transparent fiber made of a pure glass (silica) not much thicker than a human hair. It functions as a waveguide, or "light pipe", to transmit light between the two ends of the fiber. - K. Thyagarajan, Ajoy K. Ghatak, Fiber Optic Essentials, page 34

3. This is a picture of a network switch. A network switch or switching hub is a computer networking device that connects network segments or network devices. The term commonly refers to a multi-port network bridge that processes and routes data at the data link layer. - Wikipedia

   Credit: LeaseWeb

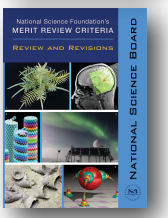
National Science Board Recent Publications



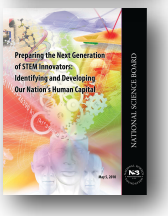
Science and Engineering Indicators 2012 (NSB-12-01)

Digest of Key Science and Engineering Indicators 2012 (NSB-12-02)

Research & Development, Innovation, and the Science and Engineering Workforce (NSB-12-03)



National Science Foundation's Merit Review Criteria: Review and Revisions (NSB-11-86)



Preparing the Next Generation of STEM Innovators: Identifying and Developing Our Nation's Human Capital (NSB-10-33)

Recommended Citation:

National Science Board. 2012. *Digital Research Data Sharing and Management.* Arlington VA: National Science Foundation (NSB-11-79).

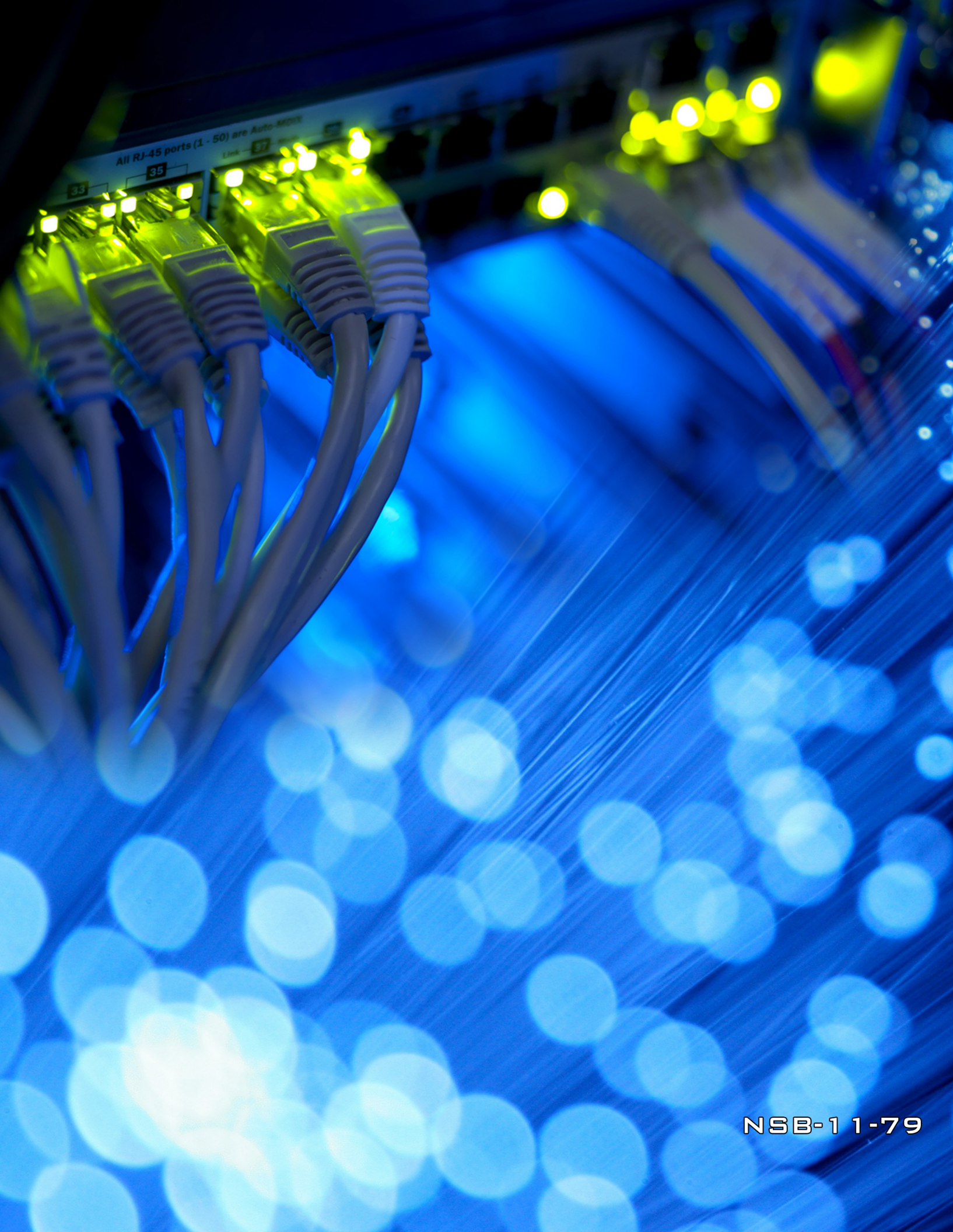
Obtaining the Board Report:

The report is available electronically at: *http://www.nsf.gov/nsb/publications/2011/nsb1179/*

Paper copies of the report can be ordered by submitting a web-based order form at: *http://www.nsf.gov/publications/orderpub.jsp* or contacting NSF Publications at: 703-292-7827

Other options for obtaining the document: TTY: 800-281-8749; FIRS: 800-877-8339

For special orders or additional information, contact the National Science Board Office: *NationalScienceBrd@nsf.gov* or 703-292-7000.


All RJ-45 ports (1 - 50) are Auto-MDIX
33
35
37
NSB-11-79