
Communicating Science and Engineering Data in the Information Age

Communicating Science and Engineering Data in the Information Age

Panel on Communicating National Science Foundation Science and Engineering Information to Data Users

Committee on National Statistics
Division of Behavioral and Social Sciences and Education

Computer Science and Telecommunications Board
Division of Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by the National Science Foundation under a grant to the National Academy of Sciences. Support of the work of the Committee on National Statistics is provided by a consortium of federal agencies through a grant from the National Science Foundation (award number SBE-0453930). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

Library of Congress Cataloging-in-Publication Data
or
International Standard Book Number 0-309-0XXXX-X 978-0-309-22209-9
Library of Congress Catalog Card Number 97-XXXXX

Additional copies of this report are available from National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu

Suggested citation: National Research Council. (2011). *Communicating Science and Engineering Data in the Information Age*. Panel on Communicating National Science Foundation Science and Engineering Information to Data Users, Committee on National Statistics and Computer Science and Telecommunications Board, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

Panel on Communicating National Science Foundation Science and Engineering Information to Data Users

Kevin Novak (*Chair*), Integrated Web Strategy and Technology, The American Institute of Architects
Micah Altman, Institute for Quantitative Social Science, Harvard University
Elana Broch, Population Research Library, Princeton University
John M. Carroll, Department of Information Sciences and Technology, Pennsylvania State University
Patrick J. Clemins, R&D Budget and Policy Program, American Association for the Advancement of Science, Washington, DC
Diane Fournier, Communications Division, Statistics Canada, Ottawa, Canada
Christiaan Laevaert, Eurostat, Statistical Office of the European Union, Luxembourg
Andrew Reamer, George Washington Institute of Public Policy, George Washington University

Emily Ann Meyer, *Co-Study Director*
Thomas Plewes, *Co-Study Director*
Michael J. Siri, *Program Associate*

Committee on National Statistics

Lawrence D. Brown (*Chair*), Department of Statistics, The Wharton School, University of Pennsylvania
John M. Abowd, School of Industrial and Labor Relations, Cornell University
Alicia Carriquiry, Department of Statistics, Iowa State University
William DuMouchel, Oracle Health Sciences, Waltham, Massachusetts
V. Joseph Hotz, Department of Economics, Duke University
Michael Hout, Department of Sociology, University of California, Berkeley
Karen Kafadar, Department of Statistics, Indiana University
Sallie Keller, Science and Technology Policy Institute, Institute for Defense Analyses
Lisa Lynch, The Heller School for Social Policy and Management, Brandeis University
Sally Morton, Statistics and Epidemiology, RTI International, Research Triangle Park, North Carolina
Joseph Newhouse, Division of Health Policy Research and Education, Harvard University
Ruth D. Peterson, Criminal Justice Research Center, Ohio State University
Hal Stern, Donald Bren School of Information and Computer Sciences, University of California, Irvine
John H. Thompson, National Opinion Research Center, Chicago
Roger Tourangeau, Survey Research Center, University of Michigan, and Joint Program in Survey Methodology, University of Maryland
Alan Zaslavsky, Department of Health Care Policy, Harvard Medical School

Constance F. Citro, *Director*

Computer Science and Telecommunications Board

Robert F. Sproull (*Chair*), Sun Microsystems (retired), Burlington, MA
Prithviraj Banerjee, Hewlett Packard, Palo Alto, CA
Steven M. Bellovin, Columbia University, New York
Jack L. Goldsmith III, Harvard Law School
Seymour E. Goodman, Sam Nunn School of International Affairs and College of Computing, Georgia Institute of Technology, Atlanta
Jon Kleinberg, Department of Computer Science, Cornell University
Robert Kraut, Department of Human-Computer Interaction, Carnegie Mellon University
Susan Landau, Radcliffe Institute for Advanced Study
Peter Lee, Microsoft Corporation, Redmond, WA
David Liddle, US Venture Partners, Menlo Park, CA
Prabhakar Raghavan, Yahoo! Research, Sunnyvale, CA
David E. Shaw, D.E. Shaw Research, New York
Alfred Z. Spector, Google, Inc., New York
John Stankovic, Computer Science Department, University of Virginia
John Swainson, Silver Lake Partnership, Islandia, NY
Peter Szolovits, Computer Science & Artificial Intelligence Lab, Massachusetts Institute of Technology
Peter Weinberger, Google Inc., New York
Ernest J. Wilson, Annenberg School for Communication, University of Southern California
Katherine Yelick, Computer Science Division, University of California, Berkeley

Jon Eisenberg, *Director*

Contents

Preface

The National Center for Science and Engineering Statistics (NCSES), as a means of fulfilling its mandate to collect and distribute information about the science and engineering enterprise for the National Science Foundation (NSF), conducts a program of data dissemination that includes provision of data in hard copy and, increasingly, electronic-only publication and tabulation formats; hosts a website that provides access to NCSES reports and methods by topic; and maintains two web-based tools for retrieving data from the NCSES database: the Integrated Science and Engineering Resource Data System (WebCASPAR) and the Scientists and Engineers Statistical Data System (SESTAT). These products and tools serve a community of information users with wide-ranging data needs and diversity in statistical savvy, access preferences, and technical abilities.

In 2010, in view of an expanded scope of responsibilities recognized in the America COMPETES Reauthorization Act of 2010, NCSES requested that the Committee on National Statistics and the Computer Science and Telecommunications Board of the National Research Council form a panel to review the NCSES program of collection and distribution of information on science and engineering and to recommend future directions for the program.

In accomplishing this review, the Panel on Communicating National Science Foundation Science and Engineering Information to Data Users has conducted two workshops. Their purpose was to gather information from data users and experts on various aspects of data storage, retrieval, dissemination, and archiving. At the request of NCSES, the panel issued an interim report (National Research Council, 2011), which summarized the first workshop and recommended action by NCSES on four key issues: data content and presentation, meeting changing storage and retrieval standards, understanding data users and their emerging needs, and data accessibility. The interim report pointed out that the recommended actions should be considered as preliminary steps that would assist NCSES in preparing for a transition from current practices and approaches to an improved program of data dissemination. The analysis and recommendations from the interim report are carried into this final report, along with the findings of a second workshop and the results of subsequent analysis by the panel.

The panel is grateful for the active participation of Lynda Carlson, director of NCSES, and her senior staff and for their informative and frank discussion of the status of the dissemination programs in the meetings and workshops conducted by the panel. Special thanks go to John Gawalt, who was program director for the Information and Technology Services Program of NCSES at the beginning of this study and later was named deputy director of

NCSES. He went out of his way on many occasions to respond to questions posed by the panel and to provide helpful materials as the review progressed. His replacement, Jeri Mulrow, continued this willing cooperation as she fulfilled the many requests for information to assist in framing the issues and arriving at recommendations.

A large group of experts from government agencies, the academic community, and various other user organizations freely gave of their time to prepare presentations for the workshops and enter into a dialogue with the panel as it gathered information for this report. The users were represented by Paula Stephan, Georgia State University; Jeffrey Alexander, SRI International; Kei Koizumi, Office of Science and Technology Policy, Executive Office of the President; and Bhavya Lal and Asha Balakrishnan of the Institute for Defense Analysis's Science and Technology Policy Institute.

Several experts gave presentations on various aspects of dissemination technology developments focusing on government-wide or statistical agency approaches. Alan Vander Mallie, program manager, Data.Gov, briefed the panel on the Data.gov initiatives; George Thomas, Office of Enterprise Architecture, U.S. Department of Health and Human Services, provided perspective on Data.gov and similar government initiatives to take advantage of the Internet. Suzanne Acar, senior information architect, U.S. Department of the Interior, and co-chair, Federal Data Architecture Subcommittee, gave a presentation on the work of the World Wide Web Consortium (W3C) group, which is making great headway in developing government-wide solutions to Internet issues. Judy Brewer, director of the Web Accessibility Initiative of W3C, gave a forceful presentation on the importance of ensuring that data products on the web are accessible to persons with disabilities and other limitations.

The panel benefited from the observations of Ronald Bianchi, director of the Information Services Division of the Economic Research Service of the U.S. Department of Agriculture, and chair of the Statistical Community of Practice and Engagement (SCOPE) working group, which is seeking to develop a collaborative structure for federal statistical agencies to develop and share best practices—including, for example, several areas of importance for dissemination, such as information quality, metadata, and common definitions. Jeffrey Sisson, program manager, American FactFinder, and Cavan Capps, chief, DataWeb Applications of the U.S. Census Bureau, gave presentations on these powerful dissemination tools.

The important area of archiving data was discussed by Margaret Adams, manager of the Archival Records Program, and Theodore Hull, senior archivist of the National Archives and Records Administration. Jeffrey Turner, director of sales and marketing of the U.S. Government Printing Office, and Donald Hagan, associate director, Office of Program Development of the National Technical Information Service of the U.S. Department of Commerce, discussed powerful new initiatives and tools that permit agencies to move away from dissemination of information in hard-copy formats.

The private sector is playing a growing role in the dissemination of public data sets, such as those produced by NCSES. The Google Public Data Explorer initiative was explained by Benjamin Yolken, product manager, and Jürgen Schwärzler, statistician on the Public Data Team of Google. Steve McDougall, product manager, and Stephan Jou, technical architect for IBM, described the lessons that have been learned concerning the Many Eyes website, wherein users can experiment with, download, and create visualizations of data sets.

The panel is grateful for the excellent work of the staff of the Committee on National Statistics and the Computer Science and Telecommunications Board for their support in developing and organizing the workshop and this report. Tom Plewes and Emily Ann Meyer,

co-study directors for the panel, ably supported our work. Michael Siri provided administrative support to the panel. We are especially thankful for the personal participation of Constance F. Citro, director of the Committee on National Statistics, and Jon Eisenberg, director of the Computer Science and Telecommunications Board, in the conduct of the workshops and in the preparation of this report. Their sage advice benefited the report in numerous ways.

The interim report and this final report have been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Research Council. The purpose of this independent review is to provide candid and critical comments that assist the institution in making its reports as sound as possible, and to ensure that the reports meet institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

The panel thanks the following individuals for their review of the interim report: John Bertot, College of Information Studies, University of Maryland; Margaret Hedstrom, School of Information, University of Michigan; Shirley M. Malcom, Education and Human Resources, American Association for the Advancement of Science; Gary Marchionini, School of Information and Library Science, University of North Carolina; Kathryn Pettit, National Data Repository, The Urban Institute; and Daryl Pregibon, Research Scientist, Google, Inc.

A similar note of appreciation is extended to the following individuals for their review of this final report: Andrew A. Beveridge, Department of Sociology, Queens College and Graduate Center, CUNY; Martin Grueber, Research Leader, Battelle, Cleveland, OH; James Hendler, Tetherless World Constellation Chair and Director, IT and Web Science Program, Computer and Cognitive Science Departments, Rensselaer Polytechnic Institute, Troy, NY; Joan K. Lippincott, Associate Executive Director, Coalition for Networked Information, Washington, DC; Kathryn Pettit, Senior Research Associate, National Data Repository, The Urban Institute, Washington, DC; Juana Sanchez, Department of Statistics, University of California, Los Angeles; and Julie Steele, Editor, O'Reilly Media, New York, NY.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of the interim report was overseen by Robert F. Sproull, Sun Labs, Oracle, Burlington, MA; he also oversaw the review of the final report. Appointed by the National Research Council, he was responsible for making certain that the independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of the report rests entirely with the authoring committee and the National Research Council.

Kevin Novak, *Chair*
Panel on Communicating National Science
Foundation Science and Engineering
Information to Data Users

Summary

The National Center for Science and Engineering Statistics (NCSES) of the National Science Foundation (NSF) communicates its science and engineering (S&E) information to data users in a very fluid environment that is undergoing modernization at a pace at which data producer dissemination practices, protocols, and technologies, on one hand, and user demands and capabilities, on the other, are changing faster than the agency has been able to accommodate.

NCSES asked the Committee on National Statistics and the Computer Science and Telecommunications Board of the National Research Council to form a panel to review the NCSES communication and dissemination program that is concerned with the collection and distribution of information on science and engineering and to recommend future directions for the program according to its statement of task (Box S-1):

Box S-1

Statement of Task

An ad hoc panel will review the communication and dissemination program of the National Science Foundation (NSF) National Center for Science and Engineering Statistics (NCSES) that is concerned with the collection and distribution of information on science and engineering and recommend future directions for the program. Specifically, the panel will:

(1) Review NCSES's existing approaches to communicating and disseminating statistical information, including the division's information products, web site, and database systems. [This review will be conducted in the context of both current "best practices" and new and emerging techniques and approaches.]

(2) Examine existing NCSES data on websites, information gathered by and from NCSES staff, volunteered comments of users, and input solicited by the panel from key user groups, and assess the varied needs of different types of users within NCSES's user community.

(3) Consider the impact current federal and NSF web site guidance and policies have on the design and management of the NCSES's online (Internet) communication and dissemination program.

(4) Consider current research and practice in collecting, storing, and utilizing metadata, with particular focus on specifications for social science metadata developed under the Data Documentation Initiative (DDI).

(5) Consider the impact of government-wide activities and initiatives (such as FedStats, Data.gov) and the emerging user capability for online retrieval of government statistics.

The panel will facilitate its review by conducting a two-day public workshop that will feature invited presentations and discussions. The panel will subsequently prepare an interim letter report that will focus on issues regarding transition from current approaches and a final report with specific recommendations, including a discussion of related technical, staffing, and funding issues.

The Panel on Communicating National Science Foundation Science and Engineering Information to Data Users reviewed NCSES's existing approaches to communicating and disseminating statistical information, including the division's information products, website, and database systems; examined existing NCSES data on websites, information gathered by and from NCSES staff, volunteered comments of users, and input solicited by the panel from key user groups; assessed the varied needs of different types of users in the NCSES user community; considered the impact that current federal and NSF website guidance and policies have on the design and management of the NCSES online (Internet) communication and dissemination program; considered current research and practice in collecting, storing, and utilizing metadata, with particular focus on specifications for social science metadata; and considered the impact of government-wide activities and initiatives (such as FedStats, Data.gov) and the emerging user capability for online retrieval of government statistics.

In accomplishing this review, the panel conducted two workshops to gather information from data users and experts on various aspects of data storage, retrieval, dissemination, and archiving. An interim report issued early in 2011 addressed data content and presentation, meeting changing storage and retrieval standards, understanding data users and their emerging needs, and data accessibility. The analysis and recommendations from the interim report are carried into this final report, along with the results of subsequent analysis by the panel.

These are exciting and challenging times for federal government statistical agencies responsible for disseminating their data products to their user communities, and the times are especially challenging for the National Center for Science and Engineering Statistics, which is finding the importance of its data magnified many fold by the growing recognition of the role that science and engineering investment is playing as a source of economic growth. The vision of a data dissemination program for NCSES is also in a time of flux. The agency is confronting new roles and missions, as directed in the America COMPETES Act, which changed more than its name. Technology is also opening the door to significant leaps in the ability of NCSES to communicate data and analytical products to data users. The promise of such services as Data.gov and the emergence of such private-sector solutions as the Google Public Data Explorer are just becoming recognized. The semantic web (Web 3.0) holds promise of communicating data to users in entirely new ways, much to the advantage of users and the federal agencies themselves. These technological advances open the way to new opportunities, but they are also

problematic in that they are rapidly promulgated and, many times, they rapidly become obsolete. The panel suggests that NCSES adopt an approach to modernization that stresses the basics of data provision (common formats with appropriate metadata) and partnerships with the private sector as opportunities become available, so that NCSES will avoid the issue of rapid obsolescence associated with rapid change in the particular tools and systems offered by the private sector.

In the face of these environmental and technological forces, we make a number of recommendations to the National Center for Science and Engineering Statistics to improve its dissemination program. The first set of recommendations has to do with how the survey-based data are received and input into the NCSES database, managed once there, and preserved for posterity. (The recommendations are numbered as they appear in the body of the report.)

Recommendation 3-1. The National Center for Science and Engineering Statistics should incorporate provisions in contracts with data providers for the receipt of versioned microdata at the level of detail originally collected, in open machine-actionable formats.

Recommendation 3-2. The National Center for Science and Engineering Statistics should transition to a dissemination framework that emphasizes database management rather than data presentation and strive to use auditable machine-actionable means, such as version control, to ensure integrity of the data and make the provenance of the data used in publications verifiable and transparent.

Recommendation 3-3. The National Center for Science and Engineering Statistics should require that data received from contractors be accompanied by machine-actionable metadata so as to allow for automated production of NCSES publications, comparability with previous analysis, and efficient access for third-party visualization, integration, and analysis tools.

Recommendation 3-4. The National Center for Science and Engineering Statistics should proceed to make its data available through open interfaces and in open formats compatible with efficient access for third-party visualization, integration, and analysis tools.

Recommendation 3-5. The National Center for Science and Engineering Statistics should develop a plan for redesign of its retrieval tools utilizing the emerging, sustainable capabilities of other government and private-sector resources.

Recommendation 3-6. The National Center for Science and Engineering Statistics should work with the National Archives and Records Administration to ensure long-term access and preservation of all of its publications and all data necessary to replicate these publications. As a necessary step, the National Center for Science and Engineering Statistics should review and update the request for disposition authority that is filed with the National Archives and Records Administration to ensure prompt and complete disposition of records and should regularly review the status of compliance with the records retention directive.

Engaging with its data users is an essential activity for NCSES. There is much that can be done to make that engagement more productive. The panel recommends:

Recommendation 4-1. The National Center for Science and Engineering Statistics should analyze the results of its initial online consumer survey and refine it over time. Using input from other sources, such as regular structured user focus groups and panel-based periodic user surveys, NCSES should regularly and systematically collect and analyze patterns of data use by web users in order to develop a typology of data users and to identify usability issues.

Recommendation 4-2. The National Center for Science and Engineering Statistics should educate users about the data and learn about the needs of users in a structured way by reinstating the program of user workshops and instituting user webinars.

Recommendation 4-3. The National Center for Science and Engineering Statistics should employ user-focused design and user analysis, starting with an initial heuristic evaluation and continuing as a regular and systematic part of its website and tool development.

Recommendation 4-4. The National Science Foundation should sponsor research and development on accessible data visualization tools and approaches and potential other means for browsing and exploring tabular data that can be offered via web, mobile and tablet-based applications, or browser-based ones.

The implementation of the report's recommendations should be undertaken within an overall framework that accords priority to the basic quality of the data and the fundamentals of dissemination, then to significant enhancements that are achievable in the short term, while laying the groundwork for other long-term improvements. The framework could be organized along the following lines (highest priority first):

(1) Focus on collecting the right data (by contractor or otherwise); using appropriate change management and version control to establish data provenance, flag data errors and correct them; annotating those data with sufficient machine-actionable metadata to establish a process for interpreting the data, enabling efficient access to third-party data and to automated NCSES publications; and publishing the data in formats with web-accessible open interfaces for all to use.
(2) Publish methods for combining old data and new data that have been collected under different assumptions or categories or that are disseminated in ways that make them difficult to reintegrate—this is especially necessary for the data from the old and new industry research and development expenditure surveys that will populate the Industrial Research and Development Information System (IRIS).
(3) Provide the essential data reductions and visualizations that NSF's mission requires, for example, when Congress asks for authoritative data on a certain topic, a trusted group must be able to use the data and derived publications to calculate answers.
(4) Provide a growing array of visualizations and printed products tailored for the many different uses and users.

Not every recommendation made in this report can or should be implemented immediately. Some recommendations must build on the implementation of others; for example, development of a database structure that can support accessibility through the semantic web requires that NCSES obtain data from its contractors in different formats than are now received and that it define metadata to accompany the data elements. We therefore suggest a time-phased approach to improving data dissemination, focusing on five major initiatives:

1. Change the means and content of the data received from contractors and actively participate in the development and implementation of the Data.gov compatible metadata standard now being explored by W3C and the SCOPE project.
2. Gain a better understanding of the needs of users of the data—those primary, secondary, and tertiary blocks of users—and then use the information to engage them in an effort to educate them and otherwise meet their needs.
3. Conduct a continuous usability evaluation program, much akin to a program of continuous improvement that is part and parcel of any total quality management program.
4. Provide data in retrievable formats and encourage private-sector providers and individual users to import the data into their visualization tools.
5. Ensure full short-and long-term access to the data by updating its retrieval tools and ensuring proper archiving of its publications and database.

1
The Changing Data Dissemination Landscape

The National Center for Science and Engineering Statistics (NCSES) of the National Science Foundation (NSF) communicates its science and engineering (S&E) information to data users in a very fluid environment in which data dissemination practices, protocols, and technologies, on one hand, and user demands and capabilities, on the other, are changing faster than the agency has been able to accommodate. In this chapter, we discuss how strong forces are driving changing expectations on the part of users of S&E resource and workforce data, as well as how technology and a changing policy and analytical environment in the federal government are forcing NSF to rethink and modernize the manner in which NCSES communicates information to the public.

To help understand how NCSES can respond to the driving forces that we document, we also discuss the environment that it faces and that faces federal statistical agencies in general. For NCSES, this environment is determined by policies established by the Office of Management and Budget (OMB), NSF, and its own policies and procedures that have evolved over the years.

S&E INVESTMENT AND ECONOMIC GROWTH

Much of the pressure that NCSES faces to modernize the way it disseminates information stems from the subject matter itself. It has become increasingly understood that investment in research and development (R&D) creates a platform for innovation and that innovation is a major determinant of national economic competitiveness and growth. It has likewise been increasingly apparent that an associated major determinant is human capital, represented in the output of programs of education and training for the S&E workforce.

The relationship of innovation and science, technical, engineering, and mathematics (STEM) education has been recognized in several major reports, and these reports have formed the basis for major program initiatives. The 2005 report, *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Future* concluded that "a primary driver of the future economy and concomitant creation of jobs will be *innovation,* largely derived from advances in science and engineering" (National Academies, 2010a, p. 2). Underscoring the case

for R&D investment is the conclusion by the National Science Board that "while only four percent of the nation's work force is composed of scientists and engineers, this group disproportionately creates jobs for the other 96 percent" (National Science Board, 2010, Figure 3-3, p. 3-13).

The 2010 follow-up report to the *Gathering Storm* report further concluded that "substantial evidence continues to indicate that over the long term the great majority of newly created jobs are the indirect or direct result of advancements in science and technology, thus making these and related disciplines assume what might be described as disproportionate importance" (National Academies, 2010, p. 18).

The conclusions of these reports are based on analysis that relies heavily on the data that are produced by NCSES. Indeed, the need for good data on science and engineering was recognized as a principle for competitiveness in another recent report, which concluded that "benchmarking national competitiveness across a set of established and forward looking metrics—measuring both inputs such as education, R&D spending, patents and outputs such as job creation, new industries and products, GDP growth and quality of life—is necessary to drive the successful development and implementation of appropriate competitiveness policies" (Global Confederation of Competitiveness Councils, 2010, p. 3).

The three pillars on which the White House Strategy for American Innovation are built—education, research and private sector innovation[1]—are topics on which NCSES now collects data. The White House strategy focuses on educating the next generation with 21st century skills, creating a world-class workforce, and strengthening and broadening American leadership in fundamental research. In order to measure progress in educating the next generation, data are needed on progress in STEM education and its outcomes. The place of American leadership in fundamental research requires data on investments in fundamental science by the public and private sectors, as well as information on the nature and benefits of federally funded investments in research. The White House strategy requires measuring private-sector innovation expenditures (via the Business Research and Development Information Survey).

Recent legislation also underscored the importance of NCSES data. The America Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science (America COMPETES) Reauthorization Act of 2010 requires "a comprehensive study of the economic competitiveness and innovative capacity of the United States." This law, among other initiatives, changed the name and mission of NCSES (see below). It strongly emphasized the need for improvements in the current competitive and innovation performance of the U.S. economy relative to other countries that compete economically with it; coming to grips with regional issues that influence the economic competitiveness and innovation capacity of the United States; and evaluating the effectiveness of the federal government in supporting and promoting economic competitiveness and innovation. All of these initiatives require access to the kind of information that NCSES produces in its data collections.

A BROADER MISSION FOR NCSES

The new emphasis on innovation and competitiveness has been reflected in the new mission statement for NCSES. Not only was the Science Resources Statistics Division (SRS) renamed the National Center for Science and Engineering Statistics by Section 505 of the

[1] See http://www.whitehouse.gov/innovation/strategy/executive-summary.

America COMPETES Act, but also new roles and missions were assigned. Several words in the new mission statement signal this new direction: serve as a "central Federal clearinghouse" for the collection, interpretation, analysis, and "dissemination" of objective data on science, engineering, technology, and research and development. NCSES expects to use the findings and recommendations in this report in determining how best to implement its new dissemination mandate.

According to the America COMPETES Act, the dissemination function is to cover "data related to the science and engineering enterprise in the United States and other nations that is relevant and useful to practitioners, researchers, policymakers, and the public, including statistical data on— (A) research and development trends; (B) the science and engineering workforce; (C) United States competitiveness in science, engineering, technology, and research and development; and (D) the condition and progress of United States STEM education." Data collections related to U.S. competitiveness and STEM education are part of these new responsibilities. We note that these new roles and responsibilities came without additional resources in terms of budget or staff.

The next two sections present examples of the role that NCSES data play in supporting initiatives to develop federal R&D indicators.

SCIENCE OF SCIENCE POLICY

The need for science and engineering metrics has been embedded in the NSF Science of Science and Innovation Policy (SciSIP), as originally articulated by John H. Marburger III, the former director of the Office of Science and Technology Policy (OSTP) and presidential science adviser. According to the agency's description, "the SciSIP program underwrites fundamental research that creates new explanatory models, analytic tools and datasets designed to inform the nation's public and private sectors about the processes through which investments in science and engineering (S&E) research are transformed into social and economic outcomes. Or, put another way, SciSIP aims to foster the development of relevant knowledge, theories, data, tools, and human capital. SciSIP's goals are to understand the contexts, structures and processes of S&E research, to evaluate reliably the tangible and intangible returns from investments in R&D, and to predict the likely returns from future R&D investments within tolerable margins of error and with attention to the full spectrum of potential consequences" (National Science Foundation, 2008c).

The STAR METRICS (Science and Technology for America's Reinvestment: Measuring the EffecTs of Research on Innovation, Competitiveness and Science) program is led by an interagency consortium consisting of the National Institutes of Health (NIH), NSF, and OSTP (Lane and Bertuzzi, 2010). The goal of the program is to create a data infrastructure that will permit the analysis of the impact of science investments using administrative records as well as other electronic sources of data. The program will have two phases. The first phase will use university administrative records to calculate the employment impact of federal science spending through the American Recovery and Reinvestment Act and agencies' existing budgets. The second phase will measure the impact of science investment in four key areas:

- ***Economic growth*** will be measured through such indicators as patents and business start-ups.

- ***Workforce outcomes*** will be measured by student mobility into the workforce and employment markers.
- ***Scientific knowledge*** will be measured through publications and citations.
- ***Social outcomes*** will be measured by the long-term health and environmental impact of funding.

The metrics derived from the NCSES surveys are essential inputs to such science, innovation, and competitiveness metrics. The emphasis on metrics has been adopted and codified as a key element in the NSF Strategic Plan for 2011 to 2016 (National Science Foundation, 2011, p. 9)

RESEARCH AND DEVELOPMENT DASHBOARD

As this report was being prepared, OSTP further underscored the importance of innovation to the economy by announcing the launch of an online tool that permits tracking of U.S. progress in innovation. The *R&D Dashboard* is a website that demonstrates the impacts of federal investments in research and development (R&D). (Koizumi, 2011)

The initial *R&D Dashboard* website presents data on federal R&D awards to research institutions and links those inputs to outputs—specifically publications, patent applications, and patents produced by researchers funded by those investments—from two agencies over the decade from 2000 to 2009: the National Institutes of Health (NIH) and the National Science Foundation (NSF). These two science agencies play a significant role in funding basic research in the United States; more than 80 percent of the federal government's support of university-based research, for example, comes from these two agencies. The site gathers information from two federal sites, USASpending.gov and IT.USASpending.gov, and has information on R&D investments at the state, congressional district, and research institution levels. Information that feeds the *Dashboard* from these two sites, however, is not being updated because of funding cuts.[2]

The OSTP *R&D Dashboard* is designed to answer questions of the following kind: Which institutions by state are performing federally funded research? What fields of science are emphasized locally? Where are the hot spots for robotics, for example, or optical lasers, or advanced textiles resulting from federally funded research? How are federal research grants contributing to the scientific literature by field of science?

The *Dashboard* is looked on as a first step. OSTP plans to explore fundamental changes in how data on R&D are made available to the public. As in other areas included in the push for greater transparency, the emphasis will be on testing models for making R&D-related data from contributing agencies available in ways that are secure, interoperable, and usable by a wide array of potential users. The initial emphasis will be to coordinate further development with coordinating bodies supported by OSTP, including the National Nanotechnology Initiative and the National Coordination Office (NCO) for Networking and Information Technology Research and Development (NITRD).

[2] See http://www.washingtonpost.com/blogs/federal-eye/post/new-cios-role-will-be-belt-tightening/2011/03/23/gIQAdTjbuI_blog.html. Retrieved on August 15, 2011.

INTERNET TRANSFORMS THE DISSEMINATION ENVIRONMENT

In the realm of information dissemination, the Internet has been changing everything for some time. The ongoing radical transformation in the modes of data dissemination has profound implications for NCSES.

More than 15 years ago, the OMB's Federal Committee on Statistical Methodology (FCSM) recognized the growing presence of electronic options for data dissemination in a report entitled *Electronic Dissemination of Statistical Data* (Office of Management and Budget, 1995). The authors of this report were quite prescient in noting that the rapid expansion of computer technology had "led to vast changes in the supply of and demand for Federal statistical data. Technology is no longer the primary barrier between users and information." The authors forecast even further changes with the advent of a national information infrastructure that would have even greater impact. The report concluded that statistical agencies would need to adopt new methods of disseminating statistical information and data to replace the traditional means that used to serve as the principal source of statistical information (p. 1)."

The day foretold by the FCSM committee has long since arrived. The current choices are no longer between paper publications and electronic dissemination, but between various modes of and options for electronic dissemination. Like many other statistical agencies, NCSES has, except for a few special publications, largely abandoned hard-copy publication of its data. Now there are a multitude of choices among electronic means of retrieving reports and data elements—the most prominent of these choices for the federal statistical agencies today are FedStats and Data.gov, which are discussed in Chapter 2.

From a handful of interconnected government and university research computers, the Internet has grown to near ubiquity, and today's users search the web for more information than was available in the past.[3] Moreover, with the increased availability of broadband and high-speed Internet access, dynamic, multimedia-laden websites are replacing formerly static web pages, with the consequence that users have the expectation of being able to interact with the information for which they are searching.

Moreover, a recent poll of the Pew Internet and American Life Project showed that access to the Internet is quickly becoming "untethered"[4] and users are turning to smartphones and other mobile devices for access to the World Wide Web, social networking, and email. As a consequence, those who disseminate information will need to react to these changes, by continuing to leverage the newest means to access and interact with information on the web.

The U.S. government has made a number of mobile applications available on the usa.gov website. Several agencies have developed a mobile edition of their website—an abridged version available to users of smartphones, tablets, personal data assistants (PDAs), and other mobile handheld devices. Taking into account the information that it is beginning to collect in the online survey of data users (described in Chapter 4), NCSES could profitably consider mobile versions of its web presence, perhaps beginning with the development of a mobile application for its announcements of product releases and the InfoBrief series.

FEDERAL GOVERNMENT DATA DISSEMINATION POLICIES

As a federal statistical agency, NCSES operates within a set of OMB guidelines that cover a wide variety of statistical practices, from survey design to data collection to

[3] See http://www.pewinternet.org/~/media//Files/Reports/2008/PIP_Search_Aug08.pdf.
[4] See http://www.pewinternet.org/Commentary/2010/September/Technology-Trends-Among-People-of-Color.aspx.

dissemination. The federal government's policies regarding dissemination of information to the public are promulgated by OMB under the authority of the Paperwork Reduction Act (PRA) of 1980, Public Law 96-511, as amended by the Paperwork Reduction Act of 1995, Public Law 104-13 (44 USC 35). The PRA mandate is broad, calling on agencies to "perform their information activities in an efficient, effective, and economical manner" (Office of Management and Budget, 2000).

Under this authority, published in OMB Circular A-130, NCSES is required to (a) disseminate information in a manner that achieves the best balance between the goals of maximizing the usefulness of the information and minimizing the cost to the government and the public; (b) distribute information dissemination products on equitable and timely terms; (c) take advantage of all dissemination channels, federal and nonfederal, including state and local governments, libraries, and private-sector entities, in discharging agency information dissemination responsibilities; and (d) help the public locate government information maintained by or for the agency.

NCSES is also called on to maintain and implement a management system for all information dissemination products that ensures that members of the public with disabilities, whom the agency has a responsibility to inform, have a reasonable ability to access the U.S. Government Printing Office for distribution to depository libraries. Electronic information dissemination is encouraged.

These broad guidelines of Circular A-130 are further detailed in the OMB standards and guidelines for statistical surveys (Office of Management and Budget, 2006). The standards suggest that, when information products are disseminated, NSF should provide users with access to the following information:

1. Definitions of key variables;
2. Source information, such as a survey form number and description of methodology used to produce the information or links to the methodology;
3. Quality-related documentation, such as conceptual limitations and nonsampling error;
4. Variance estimation documentation;
5. Time period covered by the information and units of measure;
6. Data taken from alternative sources;
7. Point of contact to whom further questions can be directed;
8. Software or links to software needed to read/access the information and installation/operating instructions, if applicable;
9. Date the product was last updated; and
10. Standard dissemination policies and procedures.

NATIONAL SCIENCE FOUNDATION GUIDELINES

As an operating organization in NSF, NCSES must adhere to the NSF guidelines regarding the quality of data disseminated to the public. These guidelines were developed to comply with OMB-issued government-wide guidelines under Section 515 of the Treasury and General Government Appropriations Act for Fiscal Year 2001 (P.L. 106-554), which were designed to ensure and maximize the quality, objectivity, utility, and integrity of information disseminated by federal agencies.

Under NSF guidelines, *utility* is achieved by staying informed of both internal and external information needs and by developing new data or information products when appropriate. This is a multifaceted process, involving keeping abreast of information needs by conducting internal analyses of information requirements, convening and attending conferences, working with advisory committees and committees of visitors, and sponsoring outreach activities. The NSF guidelines require review of ongoing publication series and other information products on a regular basis to ensure that they remain relevant and address current information needs.

Integrity guidelines cover aspects of the security of information from unauthorized access or revision to ensure that the information that is disseminated is not compromised through corruption or falsification. NSF guidelines are designed to ensure that information is protected from unauthorized access, corruption, or revision (i.e., making certain disseminated information is not compromised through corruption or falsification).

NSF also includes *objectivity* in its guidelines. This is a focus on ensuring that information that is disseminated is accurate, reliable, and unbiased and that information products are presented in an accurate, clear, complete, and unbiased manner. Objectivity is achieved by presenting the information in the proper context, identifying the sources of the information (to the extent possible, consistent with confidentiality protections), using reliable data and sound analytical techniques, and preparing information products that are carefully reviewed. These guidelines call for the inclusion of metadata (information about the data), in that all original and supporting data sources used in producing statistical data products should be clearly identified and documented, either in the publication or on each individual table. The metadata will generally include specification of variables used, definitions of variables when appropriate, coverage or population issues, sampling errors, disclosure avoidance rules or techniques, confidentiality constraints, and data collection techniques.

DATA RELEASE POLICY

A decade and a half ago, the predecessor agency to NCSES issued a policy on data release that was based on a consumer survey of data relevance and quality that led to a review by an internal Customer Service Task Force (National Center for Science and Engineering Statistics, 1994). This was the first of two consumer surveys; a second, conducted in 1996, was summarized in *Measuring the Science and Engineering Enterprise: Priorities for the Science Resources Studies Division* (National Research Council, 2000, p. 42). The consumer studies have not been repeated since.

The 1994 data release policy statement declared that its objectives were to encourage the timely release of SRS (NCSES) survey data, ensure that the released data meet SRS standards for "releasability," and ensure that NSF management knows when the data are to be released.

According to this policy statement, the main vehicle for release of timely data was the Data Brief, which is designed to publicize the data and provide a targeted group of users with some understanding of the implications of the data. The goal was to produce timely and accurate data, with accuracy defined as free from such flaws as gross typographical errors or methodological mistakes, and that they appear plausible. Procedures for internal clearance were also outlined.

OTHER GUIDELINES

Finally, the panel suggests that NCSES consider, in conducting its dissemination program, the dissemination guidelines outlined in *Principles and Practices for a Federal Statistical Agency* (National Research Council, 2009). In regard to dissemination, this volume states that a statistical agency should strive for the widest possible dissemination of the data it compiles. Data dissemination should be timely and public. Furthermore, measures should be taken to ensure that data are preserved and accessible for use in future years. Elements of an effective dissemination program include the following:

- An established publications policy that describes, for a data collection program, the types of reports and other data releases to be made available, the audience to be served, and the frequency of release.
- A variety of avenues for data dissemination, chosen to reach as broad a public as reasonably possible. Channels of dissemination include, but are not limited to, an agency's Internet website, government depository libraries, conference exhibits and programs, newsletters and journals, e-mail address lists, and the media for regular communication of major findings.
- Release of data in a variety of formats, including printed reports, easily accessible website displays and databases, public-use microdata [5] and other publicly available computer-readable files, so that the information can be accessed by users with varying skills and needs for data retrieval and analysis. All data releases should be suitably processed to protect confidentiality, with careful and complete documentation.
- For research and other statistical purposes, access to relevant information that is not publicly available through restricted access modes that protect confidentiality. Such modes include protected research data centers, remote monitored online access for special tabulations and analyses, and licensing of individual researchers to allow them to use confidential data on their desktop computers under stringent arrangements to ensure that no one else can access the information.
- Procedures for release of information that preclude actual or perceived political interference. In particular, the content and timing of the statistical agency, and the agency or unit that produces the data should publish in advance and meet release schedules for important indicators to prevent even the appearance of manipulation of release dates for political purposes.
- Policies for the preservation of data that guide what data to retain and how they are to be archived for future secondary analysis."

[5] In the National Research Council report, and throughout this report, the term "microdata" is defined in the statistical sense, that is, microdata are data on the characteristics of units of a population, such as individuals, households, or establishments, collected by a census, survey, or experiment (U.S. Bureau of the Census, 1998).

2
The Current Dissemination Program

The current dissemination program of the National Center for Science and Engineering Statistics (NCSES) is wide-ranging and multifaceted. In order to fulfill its mandate to serve as collector and distributor of information about the science and engineering enterprise for the National Science Foundation (NSF), this relatively small, resource-constrained statistical agency[1] disseminates its publishable data in several formats (hard-copy, mixed, and electronic-only publications); maintains an extensive website; makes its data available for retrieval from the consolidated FedStats database and through the Data.gov portal; provides access to confidential microdata in a protected environment for research purposes; and supports provision of three online communal tools that are used to retrieve data from the NCSES database: the Integrated Science and Engineering Resource Data System (WebCASPAR), the Scientists and Engineers Statistical Data System (SESTAT), and the less known Industrial Research and Development Information System (IRIS) (see Table 2-1).

These diverse outputs and self-maintained tools serve a broad community of information users with widely different data needs, ranging from one-time casual to recurring, highly sophisticated and widely divergent levels of statistical knowledge that extend from rudimentary to very knowledgeable. The user community also has quite different access preferences, as attested by the users who discussed their uses of the data with the panel (see Chapter 5). With limited resources, NCSES attempts to be all things to all users, and because it is spread so thinly, the panel has serious concerns about whether these outputs and tools are optimized for all the tasks to which they are addressed, as well as about whether NCSES is using the most up-to-date technologies and processes to best advantage for the user community.

In this chapter, we assess the status of the NCSES dissemination program. First, we describe the remaining hard-copy publications. We then review the NCSES user interface tools, including WebCASPAR, SESTAT, and IRIS, through which individuals are able to directly access and retrieve tailored outputs from the database. Then we discuss the structure of the databases and their current presentation on the web for downloading and use by third parties. We assess the current status of the program in light of the emerging practices for electronic dissemination, primarily the development of the Semantic Web as a way to facilitate access to

[1] The 2011 budget for NCSES was $41.5 million, down from $45.7 million in fiscal year 2009 and $41.9 million in fiscal year 2010. The agency has only 45 full-time permanent staff members, of whom 21 are statisticians.

information on the Internet. We provide examples of semantic web systems in federal agencies and the possibilities for development of a semantic web structure for science and engineering (S&E) information on the Internet. Finally, we consider the important issue of timeliness—a subject of great concern for users of NCSES data—and the possibility of moving the release and distribution of S&E data to a real-time basis.

TRADITIONAL FORMAT PUBLICATIONS

NCSES continues a few publications using a print-based approach and still has a customer base for them, although that customer base seems to be declining over time. Moreover, although most retrieval of NCSES information is by electronic means, a large part of the offerings are simply electronic depictions of previous hard-copy publications. It is fair to say that NCSES continues to manage its publications program in much the same way as it traditionally has, although the finished products, for the most part, are now sent to the website for posting rather than to a printing facility for production and distribution. The shift to provision of data in electronic format over the years can be characterized as a thin digitization of previously manual products. The format for the database that is made available on the website and that is queried by the NCSES tools is largely a replication of the old tables that found their way into the printed publications.

The major publication is *Science and Engineering Indicators*, a massive (in terms of bulk and effort) biannual product of the National Science Board, to which the NCSES staff makes a substantial commitment. This publication and the S&E indicators program that underscores it are the subject of a companion National Research Council study that is ongoing as our report was being prepared and are therefore not reviewed here. Nonetheless, in interviews with users (see Chapter 4), the volume was well regarded and companion online publications, such as the *Digest*, have also proven popular.[2]

Several annual publications continue to appear in hard copy. Among these publications are some that pertain to specialized audiences: *Women, Minorities, and Persons with Disabilities in Science and Engineering*; *Doctorate Recipients from United States Universities: Summary Report*; and A*cademic Institutional Profiles*. These series have proven to be popular, but their small circulations indicate their limited reach in hard-copy format.

Another series that still has some traction is the *InfoBriefs* series, which is published in both hard copy and on the website. In this series, NCSES highlights key findings of its major statistical programs in summary form, largely to improve the timeliness of data release. Typically, the *InfoBriefs* are followed by publication of a comprehensive set of detailed tables in electronic format (xls, pdf). Again, according to user comments received by the panel, this series is found to be useful and should be retained. The series appears to achieve its purpose of bringing highlights to the attention of the user community. A rudimentary search of the web shows that *InfoBriefs* are often referenced, summarized, or retransmitted in specialty newsletters and blogs.

At the same time, the NCSES approach to dissemination of standing data tables is largely a static electronic analog to its long-standing series of print publications. The approach the

[2] Despite the reported popularity of the print version of some of the publications, even those publications that continue in print have been severely curtailed. The print run for the *Science and Engineering Indicators* volume has been cut from 19,000 to 5,000 in recent years, and NCSES reports plans to cut the number of hard copies further in the future.

agency takes to the release of data tables is relatively unsophisticated when compared with approaches to table access used by other data organizations, such as the Census Bureau's American FactFinder (discussed below).

EMERGING OPTIONS FOR PRINT DISSEMINATION

Although NCSES has taken a number of steps to deemphasize or eliminate the release of its data in print form, to the extent that only a handful of publications are still available in print format, it has not done much to change the way it approaches printing and hard-copy distribution. To meet the needs of the remaining users who require hard-copy publications, there are alternative means of printing and dissemination that may be more efficient for NCSES.

According to Jeff Turner, director of sales and marketing of the U.S. Government Printing Office (GPO), the growing availability and ease of print-on-demand (POD) and electronic books may be an answer to meeting the residual need for print products. In his presentation to the panel he discussed the flexibility of arrangements for POD services.

Turner stated that such agencies as NCSES can directly purchase POD services from vendors by using a simplified purchase agreement through GPO that gives the agency complete control and convenience when looking for ways to quickly procure quality printing and related services. GPO provides training and technical assistance to agencies so they can use vendors certified by GPO.

GPO has also made arrangements for the purchase of printing services from a local Federal Express (FedEx) Office establishment through the GPOExpress contact. Moreover, agencies can choose to provide their publications to the public in POD format through the GPO sales program, wherein GPO manages the contracts and reprints books in response to customer demand, thus saving both the agency and GPO warehousing space and expense.

The GPO eBook program is another innovation. GPO uses the Google Books Partner Program to display titles that have been accepted into the GPO Sales Program, thereby increasing public awareness of federal titles. The eBook program constitutes a step toward focusing additional public attention on federal agency publications and products, but it is less pertinent to the dissemination issues faced by NCSES than the POD program.

TOOLS FOR ACCESSING THE DATABASE

The principal database access tools made available by NCSES to its data users have been in place for some time, and, like many older systems, they are in need of updating. They are best characterized as bespoke tools for individual access, having been developed from scratch to solve specific access problems associated with specific databases and dated user community requests. The resources and effort to maintain the database access tools are high relative to their utility. The capabilities are somewhat limited and technologically dated, in contrast to some of the tools emerging elsewhere among the federal statistical agencies. A brief description of the three main NCSES tools based on information provided on its website follows.

Scientists and Engineers Statistical Data System

SESTAT is a integrated data collection effort capturing information about employment, educational, and demographic characteristics of scientists and engineers in the United States. The

data are collected from three national surveys of this population: the National Survey of College Graduates (NSCG), the National Survey of Recent College Graduates (NSRCG), and the Survey of Doctorate Recipients (SDR). Data are available for download or through the SESTAT Data Tool, which allows users to generate custom data tables.

Integrated Science and Engineering Resources Data System (WebCASPAR)

WebCASPAR is a database system containing information about academic S&E resources that is available on the web. Included in the database is information from four of NCSES's research and development (R&D) expenditure surveys and two of its academic surveys plus information from the Integrated Postsecondary Education Data System (IPEDS) data from National Center for Education Statistics. According to the description, the system provides the user with opportunities to select variables of interest and to specify whether and how information should be aggregated.[3] Information is presented in HTML format and output can be in hard-copy form or in Lotus, Excel, or SAS formats for additional manipulation by the researcher.

Survey of Earned Doctorates Tabulation Engine

As this report was being prepared, NCSES released, on a pilot basis, a new data tool to provide access to selected variables from the Survey of Earned Doctorates (SED). The SED Tabulation Engine complements the WebCASPAR tool by performing tabulations on the 2006 and beyond data. This tool was a consequence of decisions to change the way confidentiality protections are applied to SED data. Beginning with the 2007 SED, data on the race/ethnicity, gender, and citizenship status of doctorate recipients were no longer reported in WebCASPAR. These changes were made with a goal of strengthening the confidentiality protections applied to SED data.

The WebCASPAR system was incapable of employing the new confidentiality protection procedures, so the range of SED variables available in WebCASPAR was reduced. The SED Tabulation Engine was developed so that NCSES could continue to provide data users with access to gender, race/ethnicity, and citizenship data from 2007 onward. This new tool displays estimates that were developed in a way that intends to prevent disclosure of personally identifiable information in tables using gender, race/ethnicity, or citizenship variables. It provides users with the ability to generate statistics using all of the SED variables previously available in WebCASPAR except baccalaureate institution and the highest degree awarded by those institutions. NCSES will explore the possibility of adding the baccalaureate institution variable to the tabulation engine in a future release.[4]

The tabulation engine includes a disclosure control mechanism that is intended to protect the identity of respondents when using the gender, citizenship, and race/ethnicity variables. It displays estimates that are intended not to disclose personally identifiable information and enables users to generate statistics using all of the SED variables previously available in WebCASPAR, except some institutional information. The SED Tabulation Engine was

[3] See http://www.nsf.gov/statistics/database.cfm.

[4] See https://webcaspar.nsf.gov/Help/dataMapHelpDisplay.jsp?subHeader=DataSourceBySubject&type=DS&abbr=DRF&noHeader=1&JS=No. Retrieved August 15, 2011.

developed by NSF through a contract to the National Opinion Research Center at the University of Chicago.[5]

Industrial Research and Development Information System

IRIS links an online interface to a historical database with more than 2,500 statistical tables containing all industrial R&D data published by NSF from 1953 through 1998 when, concurrent with implementation of the new industrial classification system, the series was discontinued. IRIS has recently been updated as an IRIS II version that contains statistics for 1953-2007. The tables that reside in IRIS and IRIS II were drawn from the results of NSF's annual Survey of Industrial Research and Development, the primary source for national-level data on U.S. industrial R&D. This survey was replaced with the Business Research and Development and Innovation Survey, for which there is currently no comparable dedicated access tool. NCSES is contemplating creation of a repository similar to IRIS and IRIS II for the new survey results.[6]

IRIS are in Excel spreadsheet format and are accessible either by defining variables, such as total R&D expenditures, or dimensions, such as size of company, for specific research topics. The data can also be obtained by querying the report in which the tables were first published.

NCSES's three major dissemination tools (SESTAT, WebCASPAR, and IRIS) have been in place without major modification for some time. Some of the data users at the panel's workshop commented that the tools are in need of a retooling. Although the tools retrieve individual and cross-tabulated data elements with some efficiency and produce tabulations as specified by users, they have no ability to enhance data analysis by use of either standardized or user-specified visualizations. Nor can they reach across the data sets to permit integrated retrieval and analysis. Furthermore, they do not offer systematic or complete access to microdata, and they fail to offer any standard means for machine access to the data and metadata, creating substantial barriers to third-party web tools and services.

If NCSES were to consider the best approach to modernizing its tools and access to the available information and data, one step would be to consult and research what other government agencies have done or are doing to improve their dissemination tools; another would be to consider what the private sector has to offer. The first approach and resulting research would enable NCSES to gain knowledge and best practices already available or in process, leveraging and incorporating the learnings into current and future tactical and strategic planning. In addition, in light of the limited resources currently available to it, NCSES should seek to identify other government agencies or private-sector partners that would provide opportunities to join, leverage, or use available toolsets and approaches (see Recommendation 3-5).

Alternative Federal Statistical Agency Tools

Although their databases are constructed in a different manner and the uses are often quite dissimilar, two quite sophisticated retrieval tools now in use and being subject to further development by the Census Bureau should be considered in assessing the adequacy and available functionality of the NCSES tools. The panel invited Census Bureau officials in charge of

[5] See https://ncses.norc.org/NSFTabEngine/#WELCOME. Retrieved May 13, 2011.
[6] Communication with Raymond Wolfe, NCSES, July 22, 2011.

maintaining and upgrading these major tools—American FactFinder and DataWeb—to discuss them at the panel workshop.

American FactFinder

The American FactFinder is the Census Bureau's primary web-based data dissemination vehicle. This tool enables the retrieval of data from the decennial census, the economic census, the American Community Survey, annual economic surveys, and the Population Estimates Program—all very large databases—in tabular, map, or chart-form data products, as well as an online access to archived data (through download).

Jeffrey Sisson, the American FactFinder program manager, reported that the system is being redesigned with several ambitious goals: to increase the effectiveness of user data access; to guide users to their data without forcing them to become experts; to improve turnaround time; to increase the efficiency and flexibility of dissemination operations; to address growing usage and data volume needs; and to provide a platform that evolves over time, avoiding technology obsolescence. The overall goal of the redesign is to make information easier to find, update the look and feel of the site, increase its functionality, implement topic- and geography-based search and navigation, standardize functionality and look across all data products and surveys, implement new and improved table manipulations, and implement charting functionality.

Sisson said that the plan for the redesign was based on stakeholder and user feedback, usability studies, and a usability audit. Based on the usability studies, the Census Bureau selected the following areas for improvement: usability and customer satisfaction, visual elements, conventional layout, consistent structure, and layering of information.

Information received by the panel after the introduction of the redesigned American FactFinder (FactFinder 2) suggests that, even with extensive usability studies, the introduction of a new tool can be a challenging activity. As our report was being prepared, the Census Bureau was continuing to work with users to refine the FactFinder 2 tool to better meet user needs.

Despite the difficulties encountered in the implementation phase, it seems appropriate for NCSES to consider the American FactFinder model, based on formal usability studies in determining how it might better provide improved user access to the large number of standing tables published subsequent to its *InfoBriefs*. The difficulties encountered in implementing the upgrades in the American Factfinder tool are also pertinent to consider when introducing a new tool to the user community.

DataWeb

In his introduction to the discussion of the Census Bureau's DataWeb network, Cavan Capps, chief of DataWeb applications, described the major tasks facing statistical agencies: how to present the right data with the right context to meet users' needs through effective data integration, how to ensure that the most recent and most correct data are displayed, and how to facilitate the efficient reuse of data for different purposes. In his presentation, he stated that Census Bureau met these challenges through the DataWeb network, which consists of three parts. The DataWeb network and its component servers create a web of machine-accessible data, whereas HotReports and DataFerrett provide tools for users to present and manipulate that data.

The DataWeb project was started in 1995 to develop an open-source framework that networks distributed statistical databases together into a seamless unified virtual data warehouse. It was originally funded by the U.S. Census Bureau, with participation at various times by the

Bureau of Labor Statistics, the Centers for Disease Control and Prevention, Harvard University, and a number of nonprofit institutions.

DataWeb is not just an archive or publisher of data; rather, it is a technology infrastructure that reads, normalizes, manipulates, and presents remote data sources from several different agencies in a way that facilitates reuse of the data for policy purposes (American Association for the Advancement of Science, 2003; Bosley and Capps, 2000; Capps et al. 1999). The DataWeb framework is accessed by hundreds of thousands of users to support statistically complex asymmetrical tabulations and visualizations for hundreds of millions of records in seconds, stored in different formats transparently and instantly. The data can be maintained by the sponsoring government agency using its own internal format and processes; thus, the available data are "official" and are updated in real time. This infrastructure is being explored as a way of reviewing data throughout the life cycle of the data creation process, making possible the capture and provision of statistically appropriate metadata that define the appropriate statistical usage and integration.

The software provides a service-oriented architecture that pulls data from different database structures and vendors and normalizes them into a standard stream of data. The normalized stream is intelligent and supports standard transformations, can geographically map itself correctly using the correct vintage of political geography, understands standard code sets so that data can be combined in statistical appropriate ways, understands how to weight survey data appropriately, and understands variance and other statistical behaviors.

Capps described DataWeb as having the capacity for handling different kinds of data in the same environment or framework. It is empowered by statistical intelligence: documentation, statistical usage rules, and data integration rules. Its features include storing the data once, but using it many times. DataFerrett and HotReports both use the DataWeb framework.

DataFerrett is a data web browser that is targeted at sophisticated data users and can present multiple data sets in an integrated way. It speeds analytical tasks by allowing data manipulation, incorporating advanced tabulation and descriptive statistics, and its mapping and business graphics use statistical rules. It has the capability of adding regressions and other advanced statistics.

HotReports are presented much like the NCSES *InfoBriefs*. They are targeted to local decision makers with limited time and statistical background. Designed to bring together relevant variables for local areas, they are topically oriented and updated when needed. They have been developed to be quick to build using a drag-and-drop layout. The main difference is that while InfoBriefs consist of static tables that are generated manually and "pasted" into documents, HotReports are generated from the data itself, its metadata, and publishing rules describing each table. This means first that it is always possible to trace the provenance (including data-editing "footnotes") of any reported summary result to the existing data, and that the table is "dynamic"—offering live updates (if desired), drill-down, and integration with other data sources.

The DataWeb system demonstrates the feasibility of integrating data from multiple federal agencies for rich reporting and analysis. It also demonstrates how metadata can be used to make data products, such as reports, both more reproducible and more dynamic.

It seems appropriate, then, for NCSES to look at DataWeb as a resource as it considers a new approach to data retrieval. It can consider redesigning its retrieval tools through incorporating aspects of DataWeb design and functionality as well as making its data available through DataWeb.

PRIVATE-SECTOR TOOLS AND INFRASTRUCTURE FOR DISSEMINATION

Information presented to the panel at its workshop emphasized that this is an exciting, fast-changing time for electronic data dissemination in the public sector. Indeed, many of the tools and applications that were discussed in the workshop in late 2010 have been substantially revised since then. Nonetheless, the panel thinks that the following summary discussion of the trends in data visualization, data publication, and data sharing is foundational, in that it points to developments that need to be taken into account by NCSES as the agency considers updating its data dissemination program.

The major inputs to the following discussion of what was then the state of practice were (a) a presentation by panel member Micah Altman, who summarized the state of current practice in terms of publicly available systems for online numeric presentation and for web-based data visualization, data publication, and data sharing; (b) a presentation describing a tool called Google Public Data Explorer by Jürgen Schwärzler, statistician, and Benjamin Yolken, project leader for this program; (c) a presentation by panel member Christiaan Laevaert on the practical aspects of using the Google Public Data Explorer tool and the significant improvements in the overall visibility of the data offerings of the Statistical Office of the European Union (EUROSTAT); and (d) a presentation by Steve McDougall, product manager, and Stephan Jou, technical architect for IBM, who described the lessons that have been learned concerning the Many Eyes website, wherein users can experiment with, download, and create visualizations of data sets.

The current set of tools for online data access include special-purpose tools for visualization, tools for one-way data publication, and tools for public data sharing and exchange. These can be further classified as open source and closed source. Some leading examples were discussed in each category.

State of the Practice in Online Data Visualization

The panel heard presentations on three toolkits that are examples of the advanced visualization that can be made possible when data are available in machine-understandable formats using open standards and metadata.

Protovis and its associated tool, Data Driven Documents (D3), are toolkits for dynamic visualization of complex data. These open-source tools handle small-sized databases. They support a partial grammar of graphics in high-level abstractions (D3 adds capacity for animation, interaction, and dynamic visualizations) (Bostock and Herr, 2009).

Similarly, **Processing** and **Prefuse Flare** are open-source toolkits built to support advanced web-based visualizations. Processing is both a framework and an open-source language that was originally based on Java. The Processing tool uses a function-based visualization model, whereas Flare is built on Flash and uses an object-based model (Herr et al., 2005; Reas and Fry, 2007).

State of the Practice in Online Data Publication

Google also has a number of offerings, including Google Docs (formerly Google Sheets), which is an Excel-type tool that has application programming interfaces (APIs) for integration

and handles small data. Fusion Tables focuses on data sharing, linking, and merging. The Google Public Data Explorer is used for data publication, and the (now defunct) Google Palimpsest had aimed to provide scientific data sharing and preservation. Despite being under the Google umbrella, each of these tools is essentially a standalone system, using its own user interfaces, with its own business model and term of services.

The most pertinent tool for public data use is the Google Public Data Explorer, which, as described to the panel by the Google development team, searches across data elements and has some visualization capability. It was launched as a Google product in March 2010. It is designed to make large, public-interest data sets easy to explore, visualize, and communicate. In the standard Google visualizations, charts and maps animate over time, and changes become easier to perceive and understand. It is designed for users who are not data experts. In a short time, users can navigate between different views, make their own comparisons, and share findings.

The Google Public Data Explorer includes a number of data sets, all of which are provided by third-party data providers, such as international organizations, national statistical offices, nongovernmental organizations, and research institutions. These providers are responsible for creating and maintaining all of the content that appears in the product.

The potential of the Google Public Data Explorer tool was discussed by panel member Christiaan Laevaert. Eurostat has been rethinking its approach to visualization tools, adapting procedures that are able, with minimal effort, to supply data in formats required by emerging tools or standards on the Internet. Free access to and reuse of data are a cornerstone of Eurostat's dissemination policy, and it is precisely the reuse of its data—in all kinds of commercial and noncommercial projects—that gives Eurostat much higher visibility than it could achieve solely through its own dissemination products. As an example, working with Google resulted not only in data's being featured on the Google Public Data Explorer but also in the integration of data into Google search with Onebox. The Google search integration makes data sets searchable in 34 languages and ensures the highest ranking in search results. Currently, four Eurostat data sets have been integrated, which has significantly improved the overall visibility of its data.

The Organization of Economic Cooperation and Development (OECD) also recently upgraded its statistics retrieval and display capabilities with the introduction of the *Statistics from A-Z—Beta Version* tool. Users can identify series with the use of keywords and obtain an instant retrieval of excel files or real-time data in a variety of formats with capacity of production of tailored charts.[7]

Although the Eurostat applications on Google Public Data Explorer and the OECD-developed Statistics from A-Z represent new and interesting efforts in the international arena, other tools have been developed by private-sector businesses that have extensive track records of developing platforms and services for data publication. These include the Nesstar Publisher, Ivation Beyond 20/20, Socrata, and the Tableau system.

Tableau is a particularly interesting example of the state of the practice in data extraction. Like the Google Public Data Explorer product, it can be used to publish data for web-scale online use. In contrast, Tableau handles data with tens of millions of rows (which is smaller than high-end SQL databases but far exceeds the capability of Google Public Data Explorer), supports a wide variety of linked visualizations, and provides a easy-to-use graphical user interface for nonexpert users to publish data. Google Public Data Explorer provides an XML API, but no

[7] See http://www.oecd.org/document/0,3746,en_2649_201185_46462759_1_1_1_1,00.html. Retrieved October 3, 2011.

configuration tools. Moreover, Tableau supports not only visualizations but also direct downloads of data extracts and of derivative "print" works, such as reports and HTML tables. Nevertheless, Google's ability to leverage its search engine dominance and redirect key search terms to Google Public Data Explorer data visualizations can provide publishers using this tool with unparalleled visibility among users.

State of the Practice in Data Sharing

Data sharing platforms go beyond data publication to allow the wider user community to comment and correct data provided through the system, add value through integrated visualizations or tags, and even provide additional data for comparison and integration. At the time our report was being prepared, there was one open-source data sharing platform, the Dataverse Network. Several competing closed commercial platforms have been developed over the last several years, including the now-defunct Google Palimpsest, Graphwise, Swivel, Dabble, and Verifiable data sharing services, as well as the operational Data360, Factual, Many Eyes, and BuzzData services. The existing services that are listed are all of note for different reasons. More new services, such as FigShare and Numbrary, have emerged recently or are on the horizon but have yet to achieve significant uptake.

The DataVerse Network (DVN) software is the only open-source system currently available specifically designed for data sharing (King, 2007). It is designed to provide access to research data and to facilitate data sharing through standard/open tools, such as DDI, Dublin Core, and USMARC metadata; Z39.50, LOCKSS, and OAI-PMH search and harvesting; and Creative Commons licensing. It replaces the Virtual Data Center software, which was developed under the NSF DLI-2 program (Altman et al., 2001). It facilitates the public preservation and distribution of persistent, citeable, authorized, and verifiable research data, with powerful but easy-to-use technology. The project increases scholarly recognition and distributed control for authors, journals, and others who make data available; improves data access and analysis; and still enables professional archives to provide integrated preservation and other services. It is a leading example of standards-based open systems.

The DataVerse Network also serves as a federated catalog, allowing users to find and access data across dozens of remote sources, including the Interuniversity Consortium for Political Social Science Research, DataWeb, and the National Archives and Records Administration. Already accessible through the DVN is the largest collection of social science data in the world, through a partnership with the Data Preservation Alliance for the Social Science (Data-PASS) (Altman et al., 2009; Gutman et al., 2009). This includes integrated access to hundreds of large government data sets.

Of these systems, the Dataverse Network is unique in being designed to explicitly support long-term access and permanent preservation. To this end, the system supports best practices, such as format migration, human-understandable formats and metadata, persistent identifier assignment, and semantic fixity checking. In addition, many threats to long-term access can be fully addressed only by collaborative stewardship of content, and the system supports distributed, policy-based replication of its content across multiple collaborating institutions, to ensure the long-term stewardship of the data against budgetary and other institutional threats (see Altman and Crabtree, 2011).

Making data available in machine-understandable formats using open standards and metadata also enables the media or other data redistributors to easily pick up the data and

integrate it into their own specific visualization tools for further dissemination. This enhances the visibility of the data and allows a statistical agency to reach a much broader audience with tools specifically targeted for such audiences. As an example, *The Guardian*, a British newspaper, has published a visualization tool based on data from Eurostat that explains to European citizens "Who we are, how we live and what it costs."[8]

Data360, created in 2004, is the oldest closed-source data sharing service still operational. Its stated aim was to make data available for better public policy. It now contains thousands of data sets and offers static and dynamic visualizations, direct access to data, and generated reports (Macdonald, 2009, p. 4).

Factual is a data manipulation developed in the commercial sector. It is closed source, runs as a proprietary service, and handles only moderate-sized databases. It extensively supports collaborative data manipulation in such functions as data linking, aggregation, and filtering, and it has extensive mashup support, with Google RESTful and Java JSON APIs for extraction and interrogation of data sets. It also integrates with Google charts and maps. It is a leading example of collaborative data editing. Factual contains a relatively small collection but has the aim of eventually loading all the Data.gov files.[9] If this aim is achieved, several of the NCSES data files that reside in Data.gov will be available in this tool.

Many Eyes is a website that permits users to enter their own data sets and produce tailored visualizations from a stock of sample visualizations on demand (Viegas, 2007). Many Eyes is largely uncurated, and as a result it hosts over 200,000 data sets, the vast majority of which are tiny, undocumented, and with unknown provenance. In part this is because the goal of the site is not to create a data collection or archive but to make visualization a catalyst for discussion and collective insight about data. Many Eyes is particularly notable for its prototype work involving accessibility for people with disabilities. (In contrast, none of the other visualization tools described provides accessible components or analogs.) By employing a processing design that carefully separates data manipulation and data analysis from presentation (see, for example, Wilkinson, 2005) and deferring visualization to the final stage of the chain of computation, the Many Eyes prototype was able to offer powerful data manipulation and analysis functions that were potentially accessible to a visually impaired audience. Although this is not yet in production, it shows that data analytics for the visually impaired can go far beyond those typically offered.

BuzzData is a relatively new entry to the data sharing offerings in which a community of interest for a data set is formed and each dataset has tabs for tracking versions, visualizations, related articles, attachments, and comments. The idea is that users using the data will build value to the data set, thereby creating a social network around it (Howard, 2011).

Trends in Data Access Tools and Infrastructure

Data dissemination is a rapidly developing area, in which players, technologies, and vocations are changing rapidly. The above review of emerging public and private-sector tools reveals a number of general trends and patterns, which are summarized below:

[8] See http://www.guardian.co.uk/world/interactive/2011/mar/14/new-europe-statistics-interactive.
[9] See http://www.factual.com/topic/government.

- In the private sector, no dominant business model, company, or commercial product has emerged. To the contrary, many commercial services in this area have failed, and business models for data sharing remain unclear.
- The availability, usability, and features of third-party systems have raised user expectations for access to data. Increasingly, users are expecting access to data in real time and at a fine level of detail. They want access to data that are machine understandable and that can be imported or mashed up using third-party services. Data.gov is a prime example of this trend applied to the public sector.
- Mega-scale online analysis, social integration, metadata exchange of catalog information, collaboration features, and ad hoc support for data manipulation are "solved problems" and well within the state of the practice. However, many services fail to adhere to good practices.
- Extremely powerful (peta-scale) online analysis, interactive statistical disclosure limitation, semantic harmonization, dynamic linking of data across different data sources with different data collection designs, and data analysis and browsing support for the visually impaired remain research problems.
- None of the commercial services is designed with preservation or long-term access.
- Both private-sector and public production services currently available fall short of providing rich access to visually impaired users.

Overall, these patterns strongly suggest that NCSES should not adopt a single service or technology for data visualization and sharing, nor should it develop another bespoke system, but instead should make data available in open formats and protocols, and with sufficient documentation and metadata, to enable the easy inclusion of these data in third-party catalogs and services. It would benefit from exploring mashups (a mashup occurs when a web page or application uses and combines data, presentation, or functionality from two or more sources to create new services) with ongoing public-sector dissemination tool sets, such as DataWeb, in order to quickly transform its electronic dissemination platforms and refine its participation in government-wide portals (see Recommendation 3-4).

DISSEMINATION BY MEANS OF GOVERNMENT-WIDE PORTALS

In addition to data dissemination through its own website and possible utilization of such tools as DataWeb, NCSES has options for disseminating data through two major government-wide initiatives. It has a presence through both portals, but they both fall short of serving as comprehensive platforms for featuring and disseminating S&E information in electronic form.

FedStats

An early, once-ambitious government-side data access service, FedStats has been available online since 1997. FedStats is a portal that was designed to be a one-stop gateway through which users can retrieve a full range of official statistical information produced by the federal government without having to know in advance which federal agency produces which particular statistic. It has searching and linking capabilities to data from agencies that provide data and trend information on such topics as economic and population trends, crime, education, health care, S&E workforce and expenditures, farm production, and more. Data can be retrieved by searching by subject matter, program area, or agency.

NCSES has been a part of FedStats from the beginning. Currently, the tool drives a user who is searching by subject matter (topic) or press releases to the NCSES website, from whence the search continues using the existing NCSES search and retrieval tools. Searching by agency is a bit problematic—the site had not been updated to incorporate the new name of NCSES as of September 2011.

Data.gov

A promising new portal for disseminating federal government information in the form of raw data and applications (apps) has more recently been developed. Data.gov is a major component of a spate of recent open-government initiatives that have been designed to serve as a catalyst for increasing transparency. NCSES has been a member of this federal government open-government initiative from its beginning in May 2009. The SESTAT tool is one of the apps that can be accessed through Data.gov, although the WebCASPAR, IRIS, and SED Tabulation Engine tools were not being made available through this portal at the time this report was being prepared.

Workshop presenter Alan Vander Mallie**,** program manager in the General Services Administration, stated that Data.gov aims to promote accountability and provide information for citizens on what their government is doing with tools to enable collaboration across all levels of government. It is a one-stop website for free access to data produced or held by the federal government, designed to make it easy to find, download, and use, including databases, data feeds, graphics, and other data visualizations.

Vander Mallie reported that, at its inception in 2009, Data.gov consisted of 47 raw data sets and 27 tools to assist in accessing the data in some of the complex data stores. At the time of the workshop, the program supported 2,895 raw data sets and 638 tools, which are accessed through raw data and tool catalogues. (The number of raw data sets and geographic data sets claimed on the Data.gov website home page had grown to nearly 390,000 by fall 2011.) This increase is primarily the result of linking and rebranding the Geospatial One Stop (geodata.gov) service as part of the Data.gov site. The catalog of raw data sets (http://explore.data.gov/catalog/raw/) available has increased to roughly 3,602, based on a catalog search. Raw data are defined as machine-readable data at the lowest level of aggregation in structured data sets with multiple purposes. The raw data sets are designed to be mashed up—that is, linked and otherwise put in specific contexts using web programming techniques and technologies. Following the workshop, Socrata, which provides an open government software solution, has introduced a new Data.gov web site designed to help government agencies publish and distribute data in new ways, including interactive charts, maps, and lists. At the time this report was being prepared, this software was available only to participating government agencies and was not accessible to the panel.

In the future, Vander Mallie said, Data.gov is slated to continue to expand its coverage of data sets and tools and to continue to support communities of interest by building community pages that collect related data sets and other information to help users find data on a single topic in one location. One continuing objective is to make data available through the application programming interface, permitting the public and developers to directly source their data from Data.gov.

Expansion into the Semantic Web, an emerging standardized way of expressing the relationships between web pages so the meaning of hyperlinked information can be understood, is also part of the future plan for Data.gov. The objective is to enable the public and developers

to create a new generation of "linked data" mashups. Working toward this goal, Data.gov has an indexed set of resource framework documents that are available and is working with the W3C to promote international standards for persistent government data (and metadata) on the web. Plans are also in place for expanding mobile applications, improved "meta-tagging" (short descriptions of an HTML web page that describe the content and facilitate implementation of standards to describe the data), and enhancing data visualization across agencies. In short, the idea is to give agencies a powerful new tool for disseminating their data and a one-stop locale for the public to access them. Efforts also exist to create government-wide or agency-specific data catalogs and dictionaries, which would be published along with the available data sets.

Suzanne Acar, senior information architect for the U.S. Department of the Interior and cochair of the Federal Data Architecture Subcommittee of the Chief Information Officer Council (see www.cio.gov), put the current and future Data.gov into context. She discussed the evolution of Enterprise Data/Information Management (EIM)—a framework of functions that can be tailored to fit the strategic information goals of any organization. For agencies like NSF to benefit from the capabilities of Web 2.0 and Web 3.0, it is important to ensure consistent quality of information and official designations of authoritative data sources.

While this report was being prepared, the future of Data.gov remained somewhat uncertain because of the threat of budget cuts (Lipowicz, 2011). Nonetheless, the development of Data.gov was heading in an additional direction—a direction that could be promising for improved dissemination of S&E data. The Office of Management and Budget is setting up a number of community-based, topic-specific Data.gov sites. The initial sites cover information on energy, law, and health.[10] In conjunction with the Office of Science and Technology Policy, NCSES might consider setting up such a topic-specific site for the science and technology community, particularly as it is now a clearinghouse for data dissemination. Overall, the sense of the panel was that Data.gov was a useful channel for disseminating NCSES data, but that NCSES should not rely on it as the only solution for disseminating data in open formats and through open APIs.

EXPANDING ACCESS TO THE NCSES DATABASE

In addition to making its database available to the public through use of the SESTAT, WebCASPAR, and IRIS tools as well as through FedStats and Data.gov, NCSES makes the microdata available under carefully controlled circumstances for download and use by outside organizations and developers. NCSES, like all federal agencies, is bound by the Privacy Act of 1974 to protect the confidentiality of the records it maintains about individuals and other statutory requirements for the protection of confidential statistical information under Title V of the 2002 E-Government Act, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), and the NSF's own statutory provisions. These statutes require NCSES to establish protocols and procedures to protect the information the agency collects. In addition, CIPSEA requires that data collected under a pledge of confidentiality be used solely for statistical purposes and thus not be disclosed in identifiable form.

This confidentiality protection is afforded to the data in several ways. Some are fairly straightforward, such as deleting identifying information (such as name and address) from the records. In other cases, however, such straightforward methods may not be adequate. This is true

[10] See http://www.data.gov/energy; http://www.whitehouse.gov/blog/2011/06/30/invitation-our-latest-open-innovation-ecosystem-energydatagov. Retrieved August 15, 2011.

for most of NCSES's microdata files that contain information about individuals. In those cases, NCSES attempts to develop a public-use file that provides researchers with as much microdata as feasible, given the need to protect respondent confidentiality. It achieves this goal by suppressing selected fields and/or recoding variables. These suppressions, however, may render the resulting data of little use to analysts and researchers.

When NCSES believes that protection of respondent confidentiality would require such extensive recoding that the resulting file would have little, if any, research utility, the agency has developed a variety of methods to assist individuals in using the data in such a situation. In some cases, researchers are able to state their needs for tabulations or other statistics with sufficient specificity that necessary summary information can be provided without the need for access to microdata. In other cases, NSF and the researcher can execute a license agreement that permits the researcher to use the data files at the NSF offices in Arlington, Virginia, or, under rigorously restricted conditions, at the researcher's academic institution.

Microdata files for three surveys may be obtained under a license agreement with NSF: the Survey of Earned Doctorates, the Survey of Doctorate Recipients, and the National Survey of Recent College Graduates. The SESTAT Integrated Data File can also be obtained in this manner.

For two of these surveys—the Survey of Earned Doctorates and the Survey of Doctorate Recipients—plans are under way to provide authorized researchers with remote access to microdata using the most secure methods to protect confidentiality. This online environment is called the NORC Data Enclave. The enclave seeks to implement technological security, statistical protections, legal requirements, and researcher training in one package. The NORC Data Enclave intends to aid in preserving data for the long term by documenting the data using Data Documentation Initiative–compliant metadata standards. When implemented, the enclave intends to set up a research "collaboratory"—an arrangement that would develop a knowledge infrastructure around each data set, enabling geographically dispersed researchers to share information through wikis and blogs. This is an expanding and innovative program for the agency, one intended to both protect confidential data and enhance the usability of the data for research and analytical purposes.

Otherwise confidential data from the 2008 Business R&D and Innovation Survey (BRDIS), sponsored by NCSES and conducted by the U.S. Census Bureau, has been made available to qualified researchers on approved projects through the Census Bureau's Research Data Centers (RDCs). This survey is a successor to the Survey of Industrial Research and Development. Data available in the RDC network are business domestic and global R&D expenditures and workforce that are collected information from a nationally representative sample of about 40,000 manufacturing and nonmanufacturing industries. There are plans to create an onsite RDC at NCSES so program staff can have access to the confidential data under controlled circumstances.

Although respondent privacy must be protected, the current NCSES approach is neither transparent, nor does it appear systematic. As the recent introduction of the SED Tabulation Engine illustrates, data from the same series survey may be split across different, nonintegrated systems. The private NCSES collection is not made available under a consistent set of terms of use (which vary by database), nor a consistent mechanism (i.e., some data sets are not available at all, some are available through the NORC enclave, and some only through the Census Bureau), nor are the methods of disclosure risk analysis used publicly documented.

Statistical and technical methods for protecting confidentiality are rapidly changing. Maximizing research utility requires a regular review of methods, consistent license agreements, and providing data in many forms, including public-use data and restricted data enclaves (National Research Council, 2005).

In addition, the need to provide confidentiality in the present does not eliminate the responsibility to provide for long-term access. The risk of reidentification changes as time elapses. As discussed in Chapter 3, all NCSES data, even confidential data, should be stewarded for long-term access and permanent preservation.

REAL-TIME DISSEMINATION AS A GOAL

One of the most common user criticisms that the panel heard about the dissemination program was the length of time between the survey reference periods and when NCSES released data from those surveys. In an era when users are increasingly being treated to real-time or near-real-time economic and social information, the lengthy delays in publication of NCSES survey results are not very well understood. The lack of timeliness is discussed here as a dissemination issue, though, in reality, timeliness problems have to do more with data gathering, statistical methodology, and processing practices, some of which have been addressed in previous National Research Council reports (National Research Council, 2004, pp-. 105, 114, 131, 147, 159-160; National Research Council, 2010, p. 21).

It was reported to the panel by the NCSES leadership that there have been initiatives by NCSES over the years to shorten the publication time by reducing reliance on printed reports and to make more use of relatively quick-turnaround formats, such as *InfoBriefs*. These have successfully put the major data series in the hands of users more quickly than in the past. However, users still have to wait too long after the reference period to get access to the detailed publication tabulations that are necessary for sophisticated analysis from a major NCSES survey; for example, detailed data from the new Survey of Industrial Research and Development for the years 2006 and 2007 were released in June 2011, a year after less detailed summaries of data from the BRDIS for 2008 were released in May 2010.

The delay in other reports, as indicated by new releases announced on the NCSES website, are similarly problematic:

- Science and Engineering Research Facilities: Fiscal Year 2007 (released September 23, 2011)
- Characteristics of Scientists and Engineers in the United States: 2006 (released September 14, 2011)
- U.S. Exports of Advanced Technology Products Declined Less Than Other U.S.Exports in 2009 (released September 1, 2011)
- Science and Engineering Doctorate Awards: 2007-08 (released August 22, 2011)
- Industrial Research & Development Information System (IRIS) 1953-07 data (released July 26, 2011)

As mentioned earlier in this chapter, the shift to provision of data in electronic format has been simply a digitization of previously manual products. The format for the website database is a replication of the old tables that found their way into the printed publications, so the laborious and time-consuming processes that were required for production of the manual products are still

necessary. Another source of the timeliness problem stems from the fact that NCSES has largely shifted to electronic dissemination but without systematic machine-understandable metadata and change control. This means that a great deal of NCSES time still must be spent in painstakingly checking data and formatting the data for print and electronic publication in order to check the accuracy and reliability of the published products. For example, each page of the hard copy must be checked by someone looking at the source data. This effort comes at the expense of ensuring data integrity at the source, and it takes an inordinate amount of scarce staff time.

TABLE 2-1 Summary of Selected Characteristics of NSF Science and Engineering Surveys

Survey	Current Contractor	Database Retrieval Tool / Publication	Variables Available	Availability of Microdata	Series Initiated/Archiving
Survey of Earned Doctorates	National Opinion Research Center (NORC)	WebCASPAR; InfoBriefs; Science and Engineering Degrees; Science and Engineering Indicators; Women, Minorities, and Persons with Disabilities in Science and Engineering; Doctorate Recipients from United States Universities: Summary Report; Academic Institutional Profiles.	Academic institution of doctorate; baccalaureate-origin institution (U.S. and foreign); Birth year; Citizenship status at graduation; Country of birth and citizenship; Disability status; Educational attainment of parents; Educational history in college; Field of degrees (N=292); Graduate and undergraduate educational debt; Marital status, number/age of dependents; Postgraduation plans (work, postdoctorate, other study/training); Primary and secondary work activities; Source and type of financial support for postdoctoral study/research; Type and location of employer; Race/ethnicity; Sex; Sources of financial support during graduate school; Type of academic institution (e.g., historically black institutions, Carnegie codes, control) awarding the doctorate	Access to restricted microdata can be arranged through a licensing agreement. A secure data access facility/data enclave providing restricted microdata access is under development with NORC.	1957 (conducted annually, limited data available 1920-1956)
Survey of Graduate Students and Postdoctorates in Science and Engineering	RTI International	WebCASPAR; InfoBriefs; Graduate Students and Postdoctorates in Science and Engineering; Science and Engineering Indicators; Women, Minorities, and Persons With Disabilities in Science and Engineering; Academic Institutional Profiles	The number and characteristics of graduate students; postdoctoral appointees; and doctorate-holding nonfaculty researchers in science, engineering and health (SEH) fields.	Data for the years 1972–2008 are available in a public-use file format	1975 (conducted annually)

Survey	Current Contractor	Database Retrieval Tool / Publication	Variables Available	Availability of Microdata	Series Initiated/Archiving
Survey of Doctorate Recipients	NORC	SESTAT; InfoBriefs; Characteristics of Doctoral Scientists and Engineers in the United States; Science and Engineering Indicators; Women, Minorities, and Persons With Disabilities in Science and Engineering; Science and Engineering State Profiles	Citizenship status; country of birth; country of citizenship; date of birth; disability status; educational history (for each degree held: field, level, institution, when received); employment status (unemployed, employed part time, or employed full time); geographic place of employment; marital status; number of children; occupation (current or past job); primary work activity (e.g., teaching, basic research, etc.); postdoctorate status (current and/or 3 most recent postdoctoral appointments); race/ethnicity; salary; satisfaction and importance of various aspects of job; school enrollment status; sector of employment (e.g., academia, industry, government, etc.); Sex; work-related training	Access to restricted data for researchers interested in analyzing microdata can be arranged through a licensing agreement. The date available on-line though the enclave arrangement discussed above.	1973 (conducted biennially)

Survey	Current Contractor	Database Retrieval Tool / Publication	Variables Available	Availability of Microdata	Series Initiated/Archiving
National Survey of Recent College Graduates	Mathematica Policy Research, Inc. and Census Bureau	SESTAT; InfoBriefs; Characteristics of Recent Science and Engineering Graduates; Science and Engineering Indicators; Women, Minorities, and Persons With Disabilities in Science and Engineering	For individuals who recently received bachelor's or master's degrees in an SEH field from a U.S. institution: age; citizenship status; country of birth; country of citizenship; disability status; educational history (for each degree held: field, level, when received); employment status (unemployed, employed part time, or employed full time); educational attainment of parents; financial support and debt amount for undergraduate and graduate degree; geographic place of employment; marital status; number of children; occupation (current or previous job); Place of birth; work activity (e.g., teaching, basic research, etc.); Race/ethnicity; Salary; Overall satisfaction with principal job; School enrollment status; Sector of employment (e.g., academia, industry, government, etc.); Sex; Work-related training	Access to restricted data for researchers interested in analyzing microdata can be arranged through a licensing agreement.	1976 (conducted biennially)

Survey	Current Contractor	Database Retrieval Tool / Publication	Variables Available	Availability of Microdata	Series Initiated/Archiving
National Survey of College Graduates	Census Bureau	SESTAT; InfoBriefs; Science and Engineering Indicators; Women, Minorities, and Persons With Disabilities in Science and Engineering	For individuals holding a bachelor's or higher degree in any field: academic employment (positions, rank and tenure); age; citizenship status; country of birth; country of citizenship; disability status; educational history (for each degree held: field, level, when received); employment status (unemployed, employed full time, or employed part time); Geographic place of employment; immigrant module (year of entry, type of entry visa, reason(s) for coming to U.S., etc.); Labor force status; marital status; number of children; Occupation (current or past job); primary work activity (e.g., teaching, basic research, etc.); publication and patent activities; race/ethnicity; salary; Satisfaction and importance of various aspects of job; school enrollment status; sector of employment (academia, industry, government); sex; work-related training	Public use data files are available upon request.	1962 (conducted biennially)
Business Research and Development and Innovation Survey (BRDIS)	Census Bureau	IRIS; InfoBrief; Business and Industrial R&D; Science and Engineering Indicators; National Patterns of Research and Development Resources; Science and Engineering State Profiles	Financial measures of R&D activity; company R&D activity funded by others; R&D employment; R&D management and strategy; and intellectual property, technology transfer, and innovation.	Census Research Data Centers	1953 (conducted annually); a new series began in 2008 when the survey was changed

Survey	Current Contractor	Database Retrieval Tool / Publication	Variables Available	Availability of Microdata	Series Initiated/Archiving
Survey of Federal Funds for Research and Development	Synectics for Management Decisions, Inc.	WebCASPAR; InfoBrief; Federal Funds for Research and Development; Science and Engineering State Profiles; Science and Engineering Indicators; National Patterns of Research and Development Resources	Federal obligations by the following key variables: character of work; basic research; Applied research; Development; federal agency; federally funded research and development centers (FFRDCs); field of science and engineering; geographic location (within the United States and foreign country); performer (type of organization doing the work); R&D plant federal outlays by: character of work; basic research; applied research; development; R&D plant	Data tables	1952 (conducted annually)
Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions	Synectics for Management Decisions, Inc.	WebCASPAR; InfoBrief; Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions; Science and Engineering State Profiles; Science and Engineering Indicators; National Patterns of Research and Development Resources	Data by federal agency, academic institutions and location: R&D; Fellowships, traineeships, and training grants; R&D plant; Facilities and equipment for instruction in science and engineering; General support for science and engineering; Type of academic institution (i.e., historically black colleges and universities (HBCUs), tribal institutions, high-Hispanic-enrollment institutions, minority institutions); Type of institutional control (public versus private)	Data tables only	1965 (conducted annually)
Survey of R&D Expenditures at Federally Funded R&D Centers (FFRDCs)	ICF Macro	WebCASPAR; InfoBrief; R&D Expenditures at Federally Funded R&D Centers; Academic Research and Development Expenditures Science and Engineering Indicators; National Patterns of Research and Development Resources.	FFRDC R&D expenditures by source of funds (federal, state and local, industry, institutional, or other); and character of work (basic research, applied research, or development)	Data tables only	1965 (conducted annually)

Survey	Current Contractor	Database Retrieval Tool / Publication	Variables Available	Availability of Microdata	Series Initiated/Archiving
Survey of Research and Development Expenditures at Universities and Colleges	ICF Macro	WebCASPAR; InfoBrief; Academic Research and Development Expenditures; Science and Engineering Indicators; National Patterns of Research and Development Resources; Science and Engineering State Profiles; Academic Institutional Profiles	Institution R&D expenditures by source of funds (federal, state and local, industry, institutional, or other); character of work (basic research versus applied research and development); pass throughs to sub-recipients; receipts as a sub-recipient; S&E field; non-S&E field; R&D equipment expenditures by S&E field; federal agency; type of degree granted, historically black college or university (HBCU), public or private control); geographic location (within the United States)	Data tables (selected items)	1972 (conducted annually, limited data available for various years for 1954-1970)
Survey of State Research and Development Expenditures	Census Bureau	InfoBrief; State Government R&D Expenditures; Science and Engineering Indicators.	State agency or department; State R&D expenditures; Internal performers; External performers; Basic research; Source of funds (federal, state, other); R&D facilities	Data tables	1964 (conducted occasionally)
Survey of Science and Engineering Research Facilities	RTI International	WebCASPAR; Scientific and Engineering Research Facilities; Science and Engineering Indicators	Status of research facilities at academic institutions and nonprofit biomedical research organizations and hospitals by: Amount and type of science and engineering research space; Current expenditures for projects to construct and repair/renovate research facilities; Condition of research facilities; Planned construction and repair/renovation of research facilities; Source of funds (federal, state and local, institutional) for construction and repair/renovation of research facilities; Research animal facilities; Bandwidth speeds and high performance network connections; Fiber; High performance computing; Wireless connections	Microdata from this survey for the years 1988 through 2001 are not available.	1986 (conducted biennially)

Survey	Current Contractor	Database Retrieval Tool / Publication	Variables Available	Availability of Microdata	Series Initiated/Archiving
Survey of Public Attitudes Toward and Understanding of Science and Technology	NORC, via an S&T module on the General Social Survey	Science and Engineering Indicators	Demographic, behavioral, and attitudinal by: How information about S&T is obtained; Interest in science-related issues; Visits to informal science institutions; S&T knowledge; Attitudes toward science-related issues	Data tables	ICPSR, 1979-2001; CD,1979-2004; (conducted biennially)

3
Strategy for Modernizing Data Storage, Retrieval, and Dissemination

In this chapter, we propose a strategy for modernizing the infrastructure and processes that support the dissemination function of the National Center for Science and Engineering Statistics (NCSES). Several rather significant actions need to be taken in order to capitalize on the new technologies and processes that would facilitate this modernization. We make six recommendations for action, ranging from revising the format in which science and engineering (S&E) data are received from the survey contractors to more attention on archiving the data for long-term access and preservation.

Capacity of NCSES to Take Advantage of New Technologies

Emerging technologies for data capture, storage, retrieval, and exchange will dramatically change the context in which NCSES will provide data to users in the future. These technologies will further increase efficiency, permitting users to access the data interactively and to dynamically integrate it with other information. For NCSES, the key to being able to take advantage of these technologies is to begin with a sharp focus on modernizing procedures for collection and ingestion of raw data and information about the data (metadata) into the data system. This is no simple task because of the likelihood that modernization will call for accommodating infrastructure changes. Whether the existing systems will have the capacity to ingest the metadata and individual record data in formats that support the new technologies is not certain.

In order to take full advantage of many of the emerging data sharing and visualization tools described in Chapter 2, it is important that the incoming data be collected and ingested into the NCSES data processing system in as disaggregated a form as possible. The data should be accompanied by sufficient information about the data items (metadata) to support future analyses and comparability with previous analyses, and there should be an appropriate versioning/change management system to ensure that the ability to trace the origin and history of the data (provenance) is incorporated. This is challenging to NCSES because, for the most part, the agency data are collected, updated, and accessed by contractors to NCSES. Since the collection, tabulation, and front-end activities are controlled by contractors, NCSES must specify the requirements for data inputs that are compatible with retrieval in open data formats and suitable for retrieval in formats that support common tools that software developers use to process data.

The data also need to be in formats that enable taking advantage of the web development capabilities embedded in Data.gov and other emerging dissemination means. The data must be capable of mashup with other data sources. These capabilities require that access to the data be available through an open application programming interface (API) that exposes the disaggregated data, along with its metadata, in machine-understandable form. The result is to enrich results and enhance the value of the data to users.

It is critically important that the data be accompanied by the machine-actionable documentation (metadata) needed to establish the data's history of origin and ownership (provenance) and include a record of any modifications made during data editing and clean-up. The documentation also needs to include the measurement properties of the data with sufficient detail and accuracy to enable publication-ready tables to be automatically generated in a statistically consistent manner.

Furthermore, it is critically important that a formal automated capability for tracking and controlling changes to a project's files—in particular to source code, documentation, and web pages (version control)—and formal change management procedures be applied to data collected by contractors. This establishes a reliable data provenance and ensures that all previous publications can be automatically verified and replicated.

In the panel's judgment, NCSES is not very well positioned to meet the above preconditions for taking advantage of emerging technologies. The survey data that are entered into the center's database are received from the survey contractors in tabular format mainly though machine-readable tabulations, rather than in a more easily accessible microdata format.

This situation is not unique to the S&E data that are received from contactors by NCSES. Suzanne Acar (representing the U.S. Federal Bureau of Investigation and the Federal Data Architecture Subcommittee of the Chief Information Officers Council) stated that difficulty in fully utilizing emerging technologies is a government-wide issue, one that will be taken up by a group of the World Wide Web Consortium (W3C) and other standards organizations.[1] W3C has plans to develop contract templates to enable governmental organizations to properly specify the format for receipt of the data from their contractors.

According to Ron Bianchi (representing the Economic Research Service of the U.S. Department of Agriculture), barriers to taking advantage of emerging technologies is a widespread issue across the federal statistical system and has been identified as a major concern for the newly formed Statistical Community of Practice and Engagement (SCOPE). This coordinating activity involves most of the large federal statistical agencies. The initial plans for the SCOPE initiative have included developing a template for contract deliverables specifications for data formats and accompanying metadata.

Recommendation 3-1. The National Center for Science and Engineering Statistics should incorporate provisions in contracts with data providers for the receipt of versioned microdata, at the level of detail originally collected, in open machine-actionable formats.

Implementing this recommendation will be no simple task for NCSES. Currently, NCSES manages 13 major surveys that involve contracts with five private-sector organizations and the U.S. Census Bureau (see Table 2-1). Furthermore, adding this requirement may initially

[1] W3C is an international community of member organizations that develops web standards: see http://www.w3c.org. Retrieved November 2010.

incur additional costs to support a shift from the current practice of formatting the data after they are received to requiring contactors to input the data in a new format. Some consideration will have to be made for reformatting the existing historical databases to be compatible with the new open formats and structures, when possible, so data can be manipulated across current and prior survey results.

To enable the receipt of metadata from contractors in a universally accessible format, NCSES should consider adopting an electronic data interchange (EDI) metadata transfer standard. The selection and adoption of a metadata transfer standard would be more effective if NCSES accomplished it through participation in a government-wide initiative, such as the W3C contract template development or the SCOPE effort, which is more focused on the federal statistical agencies.

Improving Data Delivery, Presentation, and Quality

In their presentations to the panel, the NCSES staff produced a large hard-copy stack of tabulations, noting that the stack represented just one of the center's periodic reports. The staff also noted that, even though the center has largely shifted to electronic dissemination, the dictates of data accuracy and reliability require that a great deal of NCSES time is spent in checking data and formatting them for print and electronic publication.[2] For example, each page of the hard copy must be checked by someone looking at the source data. This effort comes at the expense of ensuring data integrity at the source. We think this emphasis is misplaced.

Although it will never be possible to fully avoid edit and quality checks, because errors are prone to creep into data at any stage in processing, there is much to be gained by focusing primarily on the quality of the incoming raw data from the source. This approach is best ensured by adopting a comprehensive database management framework for the process, rather than the current primary focus on review of the tabular presentation. A framework that ensures integrity at the source of the data, buttressed by the availability of metadata, is the necessary foundation of real improvement in data dissemination. Adoption of such an approach should have further benefits. By changing to a dissemination framework from a review framework, NCSES could free up some existing resources or be able to reduce contractor involvement, which would allow for the realignment of resources and funding to focus on making further process improvements.

Recommendation 3-2. The National Center for Science and Engineering Statistics should transition to a dissemination framework that emphasizes database management rather than data presentation and strive to use auditable machine-actionable means, such as version control, to ensure integrity of the data and make the provenance of the data used in publications verifiable and transparent.

All of the tables published by NCSES are selections, aggregations, and projections of the underlying micro-level observations. Recommendation 3-2 envisions that, whenever possible, published tables should be defined explicitly in these terms and produced by an automated process that includes metadata.

The panel acknowledges that in some cases—such as the NCSES's *Science and Engineering Indicators*—this approach may not be immediately feasible, since an extensive data

[2] This information is based on the National Science Foundation presentation to the panel, October 27, 2010 (slide numbers 14-16).

appendix is necessary to support the analysis in the report. However, in general (following the practice that NCSES currently employs for the most detailed statistical tables), a web release of the raw data will reduce the burden on the NCSES staff related to manually check publications and will form the basis of a transition from tables to information and provide the users with more timely information. This structured approach to release of data will also provide transparency in the process, increase replicability, and assuage any user concerns about the delay between data collection and their availability.

It is important that the data provided by contractors to NCSES include machine-readable metadata that capture the statistical properties of the data and of the collection and research design. The appropriate form and content of these metadata are being considered in the SCOPE initiative. It is likely that such metadata are produced in the data collection process, since computer-assisted telephone interviewing (CATI) and other related survey tools use much of this information in their operations. However, metadata are currently not included in the required deliverables to the National Science Foundation (NSF) from contractors.

The shift to increased user capacity to produce customized output from the raw data is potentially a major and significant enhancement, which has the potential to offer great direct benefit, but such a change will also require consideration of second-order effects. Care will need to be taken to ensure that data confidentiality is ensured when providing users with cross-source microdata: consequently, rules about publishable cell size, for example, will have to be carefully considered.[3] The greater transparency inherent in making more of the raw data available also increases the risk that users could juxtapose data in ways that lead to invalid interpretations, although this danger can certainly be reduced by the accessibility of robust metadata that explain the meaning (and limitations) of the data.

Recommendation 3-3. The National Center for Science and Engineering Statistics should require that data received from contractors be accompanied by machine-actionable metadata so as to allow for automated production of NCSES publications, comparability with previous analysis, and efficient access for third-party visualization, integration, and analysis tools.

Another positive benefit of providing transparency and tools for exploratory access to data is that users will be in a position to identify errors in the data. NCSES should be prepared to solicit and accept error reports and make corrections as necessary. In contemporary terms, this would be an application of "crowd sourcing"—a focused attempt to tap into the collective intelligence of the users of the data. Clearly, when the general public has access and tools to combine data across data sources, there may be additional questions about data accuracy and usefulness, and NCSES will need to do its best to educate users and respond to their discoveries.

In its presentations, NCSES staff stressed that they are a comparatively small organization with limited resources. One way that these limited resources could be stretched is for NCSES to consider digital distribution channels, including enhanced use of pdf files and,

[3] Several reports of the Committee on National Statistics address the need to maintain the confidentiality of data provided to government agencies in confidence: *Privacy and Confidentiality as Factors in Survey Response* (National Research Council, 1979); *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics* (National Research Council, 1993); *Protecting Student Records and Facilitating Education Research* (National Research Council, 2009); and *Protecting and Accessing Data from the Survey of Earned Doctorates* (National Research Council, 2010).

after investigation of cost and benefits, perhaps facilitating print-on-demand (POD) publication. NCSES may wish to consider turning to POD technology of the U.S. Government Printing Office (GPO) as a potential means of controlling the costs associated with printing and distributing the few remaining hard-copy reports that it produces (see Chapter 2 for details).

VISUALIZATION OF S&E DATA

Just as a picture may be worth a thousand words, so can the best data visualizations replace a ream of tabular output and written analysis. Applications of data visualization—or as Edward Tufte (1992) characterizes it, the visual display of quantitative information—are growing profusely. (See Ware, 2004, for a contemporary treatment of this area.) Data visualizations are increasingly being used by federal data-producing agencies and others to analytically depict large data sets, such as those produced by NCSES. Two of the larger statistical agencies—the Census Bureau and the Bureau of Economic Analysis—and other federal agencies maintain visualization sites that are suggestive of approaches that NCSES might profitably take.[4]

Indeed, assisted by NCSES, the National Science Board has provided visualized data in the form of charts and graphs, and it maps its printed and online digest published in support of the 2010 *Science and Engineering Indicators* volume (National Science Board, 2010b). These static displays of information have been chosen by NSF staff for their ability to clarify relationships and trends in visually pleasing and interesting ways. They are appropriately considered first-generation visualizations, since they are not associated with an electronic database and thus are not susceptible to manipulation by data users who want to interactively illustrate aspects of the data for their own analysis.

The field of data visualization is quite dynamic, with new approaches and technologies being offered in the form of online sites and applications by both private and public sectors, as well as nuanced approaches to building a community of analysts around visualized subject matter. Because of the shortage of staff resources and the fast-changing data visualization landscape, the panel suggests that NCSES choose several deliberate approaches that can be taken in order to make progress toward improving visualization of the NCSES data. NCSES could (a) confederate with other federal statistical agencies that are already moving forward with visualization programs under an umbrella, such as SCOPE; (b) work with private-sector vendors, such as the Google Public Data Explorer, to expand the potential for visualization of the NCSES data sets (taking much the same approach as Eurostat); or (c) continue to develop a select set of straightforward visualizations, such as those offered in the 2010 *Digest* but continuously update those visualizations and post them to the Internet when new data become available.

As discussed in Chapter 2, a complementary approach would be to provide the data in machine-understandable formats using open standards and with appropriate metadata so that users can develop their own visualizations using the increasingly sophisticated private vendor visualization tools that are on the market. NCSES could take advantage of the rapidly emerging services that make data easier to find, aggregate, interpret, integrate, and link.

[4] See http://blogs.census.gov/censusblog/2011/07/visualizing-foreign-trade-data.html; http://lehd.did.census.gov/led/datatools/visualization.html; http://www.bea.gov/newsreleases/glance.htm; http://www.bea.gov/itable/index.cfm; http://www.uspto.gov/dashboards/patents/main.dashxml (retrieved on August 15, 2011).

Recommendation 3-4. The National Center for Science and Engineering Statistics should proceed to make its data available through open interfaces and in open formats compatible with efficient access for third-party visualization, integration, and analysis tools.

RETRIEVAL AND DISSEMINATION TOOLS

Adopting a new approach to data management and distribution will open up many exciting opportunities for low-cost solutions to data retrieval and dissemination. These opportunities would expand utilization of emerging government and private-sector resources to go beyond the capabilities offered by the current Scientists and Engineers Statistical Data System (SESTAT), the Integrated Science and Engineering Resource Data System (WebCASPAR), the Industrial Research and Development Information System (IRIS), and the Survey of Earned Doctorates (SED) Tabulation Engine tools.

As discussed in Chapter 2, once the conditions are established for dissemination of data, the public services, such as Data.gov, and private services, such as the Google Public Data Explorer, can bear much of the burden of dissemination. A caveat is in order here, however. Although using private-sector tools for dissemination is a promising solution for NCSES, dissemination tool development is extremely dynamic in the private sector, as panel member Micah Altman observed at the panel workshop. Many of the start-up dissemination and data sharing services have gone out of business. In view of this uncertainty, his advice is that users should mitigate the risk of using any of these systems by opting for open-source software whenever possible, retaining preservation copies of files in other institutions, limiting use to dissemination only (not data management), and leveraging metadata and APIs to create one data source that is then disseminated through multiple sources.

Another caution was voiced at the panel workshop by Myron P. Gutmann, director of NSF's Directorate of Social, Behavioral and Economic Sciences, with regard to such private-sector services as the Google Public Data Exchange. He warned that it could be dangerous to overrely on these private-sector dissemination tools, since the conditions of service or even the continued provision of service are corporate decisions that could significantly change or even end the dissemination mode. He also expressed a concern that distribution in a nongovernment-owned system could open the possibility of unauthorized changes in the data set unless there were strict controls in place within the dissemination tool and a policy that the data be anchored back to the originating federal agency source.

Altman identified research challenges and gaps between the state of the art and the state of the practice. Research challenges in this area include peta-scale online analysis, interactive statistical disclosure limitation, business models for long-term preservation, and data analysis tools for the visually impaired. Closable gaps include managing nontabular complex data and metadata-driven harmonization and linkage across data resources.

Recommendation 3-5. The National Center for Science and Engineering Statistics should develop a plan for redesign of its retrieval tools utilizing the emerging, sustainable capabilities of other government and private-sector resources.

PRESERVING ACCESS TO S&E DATA

When considering data release and management, it is important to have a long-term data management plan. Yet according to staff, the current NCSES approach to archival issues is ad hoc. In view of the importance of these data for historical reference, long-term access and permanent archival preservation are needed, and these could be ensured through proper policies and practices.

At a minimum, all of the collected data and the electronic and hard-copy publications that are produced should be scheduled for retention by the National Archives and Records Administration (NARA). In this regard, the NSF Sustainable Digital Data Preservation and Access Network Partners (DataNet) initiative is a ready in-house source of information on best practices and tools for implementing an active archival program.

NARA ELECTRONIC RECORDS PROGRAM

The National Archives and Records Administration has responsibility for the custody and retrieval of federal government records for which they have received a transfer of legal custody of records for the originating agency. A growing part of the NARA collections are in the form of electronic records. Because of the panel's interest in ensuring the long-term retention and retrieval of NCSES data, we invited Margaret O. Adams, manager of the Archival Services Program, and Theodore J. Hull, senior archivist of accessions, to discuss the NARA reference services for electronic records.

The process for identifying records for archiving is a collaborative one. NCSES is required by law to manage records created or received in the course of business, and it does so by completing a form (Standard Form 115) that outlines the holdings and requests records disposition authority. Through a records scheduling and appraisal process, the archivist of the United States determines which federal records have temporary value and may be destroyed and which federal records have permanent value and must be preserved and transferred to the National Archives of the United States. The archivist's determination constitutes mandatory authority for the final disposition of all federal records (36 CFR 1220.12). Only a very small percentage of records identified for permanent retention are actually accessed by NARA, but the kind of electronic records that are produced by NCSES have a very high chance of being appraised for permanent retention—that is, social and economic microdata collected for input into periodic and onetime studies and statistical reports, including information filed to comply with government regulations, as well as summary statistical data from national or special censuses and surveys.

According to Hull, a good part of the accessioning work is done by NARA. When records, documentation, and accession documents (SF-258) are received, NARA conducts a preliminary assessment, which can involve converting files to ASCII, contacting agency for replacements or additional documentation, verifying file formats, and selecting only permanent files for retention. Only then are records are archived using NARA's Archival Preservation System (APS).

After they are accessed, they may be researched and retrieved using descriptions of the electronic records series in NARA's online Archival Research Catalog (ARC).[5] (This source

[5] See www.archives.gov/research/arc

will be replaced by NARA's Online Public Access [OPA] system in coming months.) ARC includes descriptions for approximately 68 percent of NARA's holdings nationwide and about 99 percent of accessioned electronic records.

The NARA records system is a very large system. As of January 2011, there were 717 series and 6.6 billion logical data records contributed by over 150 source agencies described in the ARC. The ARC search supports filtering by type of records (data files), and copies of fully releasable data files are provided on removable media for cost recovery. The Online Public Access system currently under development aims to support direct download of electronic records files

In her presentation, Adams referred to the Committee on National Statistics publication, *Principles and Practices for a Federal Statistical Agency* (National Research Council, 2009, p. 27), which states that "a good dissemination program also uses a variety of channels to inform the broadest possible audience of potential users about available data products and how to obtain them…Agencies should also arrange for archiving of data with the National Archives and Records Administration and other data archives, as appropriate, so that data are available for historical research in future years."

As mentioned above, the archiving process begins with the identification of holdings and the request for records disposition authority by the agency. This is sometimes a challenging task, particularly with the growth of electronic versus hard-copy holdings. In the case of NCSES, the process of identifying and completing a records disposition authority request was last completed in 1995. Several types of records were then identified for permanent retention, including final published surveys and studies; electronic micro-level survey data, final edited versions of all electronic survey microdata, databases, spreadsheets, detailed tables, charts, statistical data, and other micro-level respondent information created prior to compiling, condensing, or summarizing the survey responses into the final summarized or published product; electronic text and detailed statistical tables, data analyses, and related records; electronic copies of survey reports, including the text of the final report and all other electronic records related to the report, such as detailed tables, charts, statistical data analyses, and spreadsheets; and technical information regarding data format and structure and other related computer program and system documentation, including codebooks, file layouts, data fields, data dictionaries, and other records that are necessary to understand the microdata. For most of these items, NCSES is instructed to retain them at the agency level for 10 years and then forward them to NARA.

Much has happened in terms of data collection, processing, and dissemination in the years since 1995. It is appropriate that NCSES review and refile, if necessary, a request for records disposition authority.

Recommendation 3-6. The National Center for Science and Engineering Statistics should work with National Archives and Records Administration to ensure long-term access and preservation of all of its publications and all data necessary to replicate these publications. As a necessary step, the National Center for Science and Engineering Statistics should review and update the request for disposition authority that is filed with the National Archives and Records Administration to ensure prompt and complete disposition of records and should regularly review the status of compliance with the records retention directive.

4
Engaging Data Users

In Chapter 1, the dynamic and growing role of the Internet as a force for change in National Center for Science and Engineering Statistics (NCSES) dissemination practices was briefly discussed. Rosabeth Moss Kanter has made the point that, although the Internet offers new challenges and opportunities, the quality of the customer experience remains centrally important to the success of many (Kanter, 2011). In developing dissemination policies and procedures, fulfilling the needs of data users in a manner that exceeds expectations of the user should be a key goal for NCSES.

Although NCSES has long been committed to serving the needs of data users, it has not gathered sufficient information on who its users are, how they use its data, and how well it is meeting their needs. Although NCSES has made several notable attempts to gather this intelligence about user needs, it does not have a formal, systematic, consistent, structured, and continuing program for doing so.

One problem for NCSES is that there are multiple communities of users for which products must be developed. Furthermore, the breadth and diversity of NCSES data users will expand as it orients itself to the broader mission mandated by the America COMPETES Act. For the most part, outreach efforts have been addressed to those whom NCSES perceives to be in its main user community. The user community consists mostly of researchers and analysts of research and development (R&D) expenditures and the R&D workforce, particularly those concerned with federal science policy.

The panel heard from key data users in the course of its workshop and through interviews conducted by panel members and staff. These users were representing the legislative and administrative branches of the federal government, the organizations that support federal government science and engineering (S&E) analysis, the academic community, and regional economic development analysts. In the presentations and interviews, these users were asked to address, from their perspective, the current practices of the National Science Foundation (NSF) for communicating and disseminating information in hard-copy publication format as well as on the Internet through the NCSES website, and the Integrated Science and Engineering Resource Data System (WebCASPAR), and the Scientists and Engineers Statistical Data System (SESTAT) database retrieval systems.

Congressional Committee Staff

Panel and staff members met with staff of the House Subcommittee on Research and Science Education to discuss congressional staff uses of the NSF S&E information. Staff work in support of the committee is a fast-turnaround operation, requiring speed in retrieving data and easy access. In fulfilling its work, the committee staff makes extensive use of *S&E Indicators* in hard copy. The staff relies on the report narrative to help them interpret the data; the analysis helps them put the numbers into perspective. They expressed the view that data tables lacking explanation are subject to misinterpretation. Like other user groups interviewed by the panel, the congressional staff expressed concern about the timeliness and frequency of the survey-based data.

The main use of the website occurs when the staff is away from the office and hard copies of the publications. They most often use Google as the search engine for discovering S&E information, commenting that the search capability of the NSF site is cumbersome and unreliable. In response to a question about use of WebCASPAR, there seemed to be confusion as to what WebCASPAR is and whether, in fact, they did use it at all. The staff often turns to the American Association for the Advancement of Science web database when they need NSF statistics, because it is readily available and comprehensive.

The House committee staff would like to have access to *Indicators* in June rather than in the following spring, and the committee had proposed legislation to make that happen; the legislation was not supported in the Senate.

Staff also expressed a need for more usability tools, such as the ability to link to other data. This capability may be available in Data.gov, but the staff has not used Data.gov very much. They were also interested in the possibility of visualization tools for the data. Some data needed for support of legislative initiatives are not presented in the aggregation (i.e., tables and cross-cuts) they desire. For example, the staff would like disaggregated S&E workforce and science, technology, engineering, and mathematical education data by occupation, industry, and geography. Also, they need more data broken out by field of science and engineering.

Congressional Research Service

As an arm of the Congress, the Congressional Research Service (CRS) responds to members of Congress and the congressional committees. In meeting the requirements of Congress for objective and impartial analysis, CRS publishes periodic reports on trends in federal support for R&D, as well as reports on special topics in R&D funding. Both types of studies rely heavily on data from NSF, both as originally published and as summarized in such publications as *Science and Engineering Indicators* and as extracted from the NSF website. The panel met with Christine Matthews, specialist in science and technology policy in the Resources, Science and Industry Division of the Congressional Research Service. She is the primary staff contact with NSF. Her recent publications include *The U.S. Science and Technology Workforce* (2009), *Science, Engineering, and Mathematics Education: Status and Issues* (2007), and *National Science Foundation: Major Research Equipment and Facility Construction* (2009).

Matthews is a frequent user of NSF information. She makes 8-10 visits to the NSF website each day and is a listserv subscriber. Although she visits the NSF website often during a given day, many of those searches are on the general NSF awards site and sites for divisions other than NCSES.

In addition to general information on S&E expenditures and workforce, she specifically references data on academic R&D for historically black colleges and universities (HBCUs) and information on R&D facilities and equipment. She directs most of her specific inquiries through the NSF congressional liaison office, mentioning staff member George Wilson.

She commented that her use of the data is limited by the curtailment in the amount of published information in NSF reports that accompanied the shift from hard-copy to electronic dissemination of several of the key reports. The HBCU data, for example, was located in a special report with analysis and extensive tables, but now they appear only as an *InfoBrief* and in data tables.

Most of her data requests are filled by data readily available on the website. She has requested special data runs only a few times, noting that not everyone has the ability to request special data runs. Her experience with WebCASPAR is positive, as it is user-friendly. She has not used SESTAT.

The timelines of the data is not a particular problem for her. She recognizes that the data require time for collection and processing. For most of her uses, the data are sufficiently timely. She is able to satisfactorily explain the lags to congressional staff members when pressed. She does not generally use visualizations of NCSES data, but when she does, she would prefer visualizations in color.

Office of Science and Technology Policy

Representing the Office of Science and Technology Policy (OSTP), Kei Koizumi summarized the extensive use of NSF S&E information by this agency of the Executive Office of the President. He typically accesses the NCSES data primarily through the NCSES website, through the detailed statistical tables for individual surveys. He commented that the *InfoBrief* series is useful in that it informs him about which data are new. He reads each *InfoBrief* and explores some of the data further. For data outside his core area (R&D expenditures data), he often looks for the data in *S&E Indicators*, and, if needed, he goes to the most current data on the NCSES website. He uses WebCASPAR to access historical data and long time series.

His overall comments focused attention on the timeliness of the data, suggesting that, to users, the data are never timely enough, although some of the lags are understandable. He remains optimistic that next year the data will be available earlier. He expressed concerns over the quality of the data, and the methodology employed in the federal funds survey, which were summarized in a recent National Research Council report (National Research Council, 2010).

Science and Technology Policy Institute

The Science and Technology Policy Institute (STPI) was created by Congress in 1991 to provide rigorous objective advice and analysis to OSTP and other executive branch agencies, offices, and councils. Bhavya Lal and Asha Balakrishnan reported on the activities and interests of STPI, which can be considered a very sophisticated user of NSF S&E information. STPI supports sponsors in three broad areas: strategic planning; portfolio, program, and technology evaluation; and policy analysis and assessment.

In their presentation, Lal and Balakrishnan reported on several specific examples of the attempts by STPI to use NSF S&E information. In one task, investigators sought to determine the amount of research funded by government and industry for specific subfields of interest (i.e.,

networking and information technology). They were able to obtain percentage basic research of R&D "by source" and "by performer" for government and industry, but not broken out by specific fields or sectors of interest as broad as networking and information technology. They were able to get data on industry R&D by fields (i.e., North American Industry Classification System codes), but without the breakdown of basic research, applied research, and development funding. Based on this experience, the investigators recommended that NSF provide access to the data in a raw format.

Their overall view was that access to NSF-NCSES data tables and briefs is extremely helpful in STPI's support of OSTP and executive agencies. However, access to the data in a raw format would better enable assessment of emerging fields. The STPI researchers would like to obtain the data sets that underlie the special tabulations related to publications, patents, and other complex data. Similarly, they would like access to more notes on conversions, particularly to international data, to understand underlying assumptions; for example, China's S&E doctoral degrees. For their work, they requested more detail on R&D funding/R&D obligations by field of science and by agency, although, for their needs, those data need not be publicly available.

Academic Uses

Paula Stephan of Georgia State University, who classifies herself as a "chronic" user of NSF S&E information, summarized her uses of the data. She has a license with NSF, and about 40 to 50 times a year she uses restricted files pertaining to SDR, SED and SESTAT, *InfoBriefs*, and the *Science and Engineering Indicators* appendix tables. She accesses data through WebCASPAR. Graduate students use WebCASPAR to build tables and create such variables as stocks of R&D, stocks of graduate students, and stocks of postdoctorates by university and field. She reported that WebCASPAR can be difficult for new users to navigate, but they have to use WebCASPAR because the NCSES web page does not always have the most up-to-date links to data. For example, the number of doctorates for 2007 and 2008 is available only from WebCASPAR.

She commented that the S&E indicators appendix tables are easy to use and that the tables are very well named, so it is easy to find data. The ability to export the data to Excel allows one to easily analyze data.

Stephan noted that she does not use table tools, but her colleague, Henry Sauermann, did so for a study, and he reported that table tools provided him with exactly what he needed (starting salaries for industry life scientists). She pointed out that the NSF staff have been very responsive to user needs. For example, in 2002 users recommended that NCSES collect information on starting salaries of new Ph.D.s in the SED, and, beginning in 2007, the question was on the SED.

She suggested a need for more user support. Data workshops were held for three years that brought together users and potential users of licensed data. This same approach could be useful for acclimating users to web-based data. It would be a good way to find out how people use the data and to find out difficulties with or questions that people have about the data.

Like other users, Stephan commented that a major problem with the data is timeliness. The lack of timeliness affects the ability of researchers to assess current issues, such as the effect of the 2008-2010 recession on salaries, the availability of positions, the length of time individuals stay in postdoctoral status, and the mobility of S&E personnel. As an example of the lag, she pointed out that the 2008 SDR will be publicly released in November 2010 but the

restricted data will not be released for licensed use until sometime in 2011 (the data were collected in October 2008). Owing to this lag, the data will provide little useful information about how the recession affected careers: analysts will have to wait until fall 2012 to get the 2010 data and will have to wait until sometime in 2013 to get the restricted data.

Similarly, the earliest SED data collected during recession—for July 1, 2008, to June 30, 2009—were not scheduled to be released until November 2010 (note: the data release was subsequently delayed to allow for correction of data quality issues in the racial/ethnic data). So it is "early" recession data, although it will be analytically important because it will be the third year for which salary data have been collected in SED: when these SED salary data are available, analysts will be able to learn a good deal comparing the data with earlier years. However, such analyses will have to wait until November 2011 when the 2010 SED (July 1, 2009 to June 30, 2010) data are released (and assuming that salary data are made available).

Stephan pointed out the timeliness is not a new issue. She quoted a 2000 National Research Council report: "SRS must substantially reduce the period of time between the reference date and data release date for each of its surveys to improve the relevance and usefulness of its data" (National Research Council, 2000, p. 5).

Regional Economic Development Users

Jeffrey Alexander, a senior science and technology policy analyst with SRI International, is a frequent user of NSF S&E information and a contractor to NSF. In his presentation, he summarized his previous private-sector uses of the information, mainly focused on uses of the data for analysis of technology applications at the state policy level.

He accessed data from the website and through use of WebCASPAR. He stated a major caution about the comparability of data sources and noted that good metadata (data about the data) are not generally available for NCSES data. In particular, he said there is a need for more detailed geographic coding of the data so one can be confident in matching NSF data with data from the Bureau of Labor Statistics and other sources. Like other users, he expressed a concern with the timeliness of the data and said that timeliness is a key factor in the relevance of the data.

With regard to access, Alexander said he often needs trend data, so he most generally goes to the tables on the web page to extract specific data items. He has problems in downloading multiple files, and he finds that the WebCASPAR and SESTAT tools are not very user-friendly. A useful enhancement would be to enable searches for variables across the various surveys. He does not use the printed publications, although he finds that the *InfoBriefs* are very useful in announcing and highlighting new products.

Alexander suggested that the NCSES needs to become a center of information for the user community, and it should devote more attention to reaching out to larger users with information about how to access data as well as to seek input for improvements.

Limitations of User Analysis

The input received in the workshop and in the interviews was very helpful to the panel in framing its analysis of user needs. The users of NCSES data can conveniently, if imprecisely, be classified as primary users (those who directly use NCSES data in their research and analysis); secondary users (those who indirectly rely on NCSES products to understand and gauge the

implications for programs, policy, and advocacy, and those who assist others in obtaining access to the data); and tertiary users (the public).

The input of primary users was extensively provided in the panel workshops and in interview sessions, and some information was gathered from secondary users, but information from tertiary users was less systematically gathered and is given less attention in this report. Only since NCSES has begun to conduct consumer surveys is information about the needs of all user groups becoming known.

It is incumbent on NCSES to consider the needs of all of these groups and the technology platforms they use to access the data as the agency considers the program of measurement and outreach discussed in this report. NCSES could consider novel means of harvesting information about data use to analyze usage patterns, such as reviewing citations to NCSES data in publications, periodicals, and news items. For example, to get a sense of users who are citing *S&E Indicators*, a panel member did a Web of Science "cited reference search" on *NAT SCI BOARD and (sci eng ind*). This exercise yielded a list of 691 publications going back to 1988, shortly after *S&E Indicators* was introduced under that name. Google Scholar is another potential source of such information.

Reaching out to a wide variety of data users by means of surveys or interviews would be another worthwhile initiative. Moreover, such interactions would inform NCSES not only about user dissemination needs, but also about their substantive data needs, such as subject, variables, and level of geography. A list of organizations that could be contacted to assist in obtaining input on uses of S&E information would include the American Association for the Advancement of Science, the American Economic Association, the Association of Public and Land-Grant Universities, the Association of Public Policy Analysis and Management, the Association for University Business and Economic Research, the Council for Community and Economic Research, the Industry Studies Association, the International Association for Social Science Information Services and Technology, the Interuniversity Consortium for Political and Social Science Research, the National Association for Business Economics, the Special Libraries Association Divisions of Biomedical and Life Sciences and Engineering, and the State Science and Technology Institute. One means of ensuring that the needs of the secondary and tertiary data users are met is to ensure that programs of outreach are specially directed to members of the media—those who rerelease the NCSES data and interpret them to the public.

Among the tools that NCSES has used to assess user needs, according to John Gawalt, NCSES program director for information and technology services at the time of the workshop, are URCHIN, a web statistics analysis program that analyzes web server log file content and displays the traffic information on the basis of the log data, and WebTrends, software that collects and presents information about user behavior on its website. With proper permissions and protections, NCSES is also contemplating using cookies to identify return users and increase the efficiency of filling data requests.

In April 2011, NCSES took another step in the direction of obtaining user feedback when it placed a link on the website that directs users to a short customer survey to formally measure satisfaction and initiated an email-based sample survey, sent to customers who had requested electronic notification of new NCSES reports. As of mid-August 2011, the agency had received 44 responses to the website survey and 20 responses to the email survey. Most of those responding to both surveys were researchers, students, and teachers, with a smaller number of librarians, reporters, and policy makers, including legislative staff members.

The respondents viewed the organization of the home page in positive or neutral terms, reporting that they could find what they were looking using the current topical groupings or that they could find what they needed even though the organization was not satisfactory. Not surprisingly, researchers tended to want more in-depth reports with extensive data and analysis and detailed data tables, whereas reporters and policy makers were more likely to be satisfied with short, topical reports with summary data and analysis. Students and teachers varied in their needs and were about split between wanting short, topical reports and wanting more in-depth reports. Detailed data tables were commonly requested from this subset of customers as well. The staff of NCSES reports that it will continue to solicit the views of visitors to the website and to periodically solicit views from a sample of requestors of electronic notification of NCSES reports in the future.

Recommendation 4-1. The National Center for Science and Engineering Statistics should analyze the results of its initial online consumer survey and refine it over time. Using input from other sources, such as regular structured user focus groups and panel-based periodic user surveys, NCSES should regularly and systematically collect and analyze patterns of data use by web users in order to develop a typology of data users and to identify usability issues.

The surveys are a useful start, but there is much more that can be accomplished by way of seeking the input of data users. In seeking a model for outreach to users, NCSES could consider modeling its efforts on the very aggressive program of Statistics Canada, described at the workshop by panel member Diane Fournier. Statistics Canada uses a combination of online questionnaires, focus groups, and usability testing to assess user needs and the usability of its website. One advantage of this approach, although it is resource intensive, is the possibility of gathering use information from a wide range of users, both from regular users, who are knowledgeable, and from secondary and tertiary users, who are less familiar with the data.

Another initiative that NCSES could undertake to better determine user needs is to renew the data workshops that it conducted for several years but have been discontinued. Those workshops brought together users and potential users of licensed data. This same approach could be useful for acclimating users to web-based data and to introduce frequent users to changes in data dissemination practices and procedures. Such data workshops would be a good way to find out how knowledgeable data users use NCSES data and to find out what concerns users have about the data. These workshops could be conducted onsite, in remote locations (perhaps in conjunction with meetings of interested associations), or by means of webinars (perhaps hosted by interested associations).

The input received in the workshop and in the interviews was very helpful to the panel in framing its analysis of user needs. We recognize that the analysis relies mainly on the input of primary and, to a lesser extent, secondary users. The panel was not able in the time allowed to systematically gather much information from tertiary users (such as policy makers, the media, and librarians). Nonetheless, the panel thinks that it is incumbent on NCSES to consider the needs of all three of these groups and the technology platforms that they use to access the data as it considers the program of measurement and outreach discussed in this report.

The agency can begin by developing a concrete typology of its data users. One approach to this would be to develop *user personas*—that is, stereotypical characters who represent the variety of user types for the science and engineering data (Pruit and Adlin, 2006, p. 3). These

personas are usually developed by distilling data collected in interviews with users, much as the panel has tried to do in this report. The personas could be formalized in short descriptions to aid data dissemination designers, in that they provide a common description of the needs, skills, and the environment faced by the various user persona.

A related approach would be to develop a typology of *user interaction scenarios* that describe what users do with the online resources. The user scenario would provide a concrete and flexibly detailed representation of the tasks that users will try to carry out with the systems (Rosson and Carroll, 2002). These two aids (personas and scenarios) provide for a user-centered integration of the system life cycle. Once done, they will serve as a reference for subsequent redesign and they help to focus the design of usability tests and user assistance programs.

Recommendation 4-2. The National Center for Science and Engineering Statistics should educate users about the data and learn about the needs of users in a structured way by reinstating the program of user workshops and instituting user webinars.

The outreach activities discussed in this chapter, along with the development of a formal typology of users, will assist NCSES to better understand and respond to user needs. These activities will also assist the agency in allocating its scarce resources to the groups and needs that have the greatest return to the dissemination investment.

USING WIKI AND OTHER COLLABORATION TOOLS FOR COMMUNICATION WITH USERS

Another means of obtaining user input is offered by means of online collaboration tools, or wikis. Wikis have greatly improved the ability of federal agencies to establish open lines of communication and engage communities interested in their activities (Schroeder et al., 2009).

The most widely used wiki tool is Wikipedia, the collaboratively created online encyclopedia. The Wikipedia Foundation provides the computing infrastructure, the server, wiki software, general rules for entries, and style guidelines. Content is generated by anyone who has access to an Internet browser. Users can edit existing content pages or create new pages on topics not yet covered. The Wikipedia wiki software provides the online editing environment, tracks the changes made to pages, and allows contributors to engage in an online discussion about the content of pages. Page and text formatting is accomplished by simple specialized mark-up tags.

Wiki software tools have increasingly been adopted by government agencies as a platform for sharing information and as a means of encouraging the sharing of best practices and other types of information. Wiki software is available from commercial software vendors and as open-source software. Standard tools include software for group editing of online content, blog pages, threaded discussions, and file management for group access to files and images.

A version of Wiki has served as the foundation of Eurostat's dissemination system, called "Statistics Explained."[1] This is a new way of publishing European statistics, explaining what they mean, what is behind the figures, and how they can be of use, in an easily understandable language. Statistics Explained looks similar to Wikipedia, but unlike it, information can be updated only by Eurostat staff, thus ensuring the authenticity and reliability of the content. The latest data can be accessed through hyperlinks available in each statistical article.

[1] See http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Main_Page.

The U.S. General Services Administration (GSA) operates a wiki environment to encourage communication across governmental entities. The GSA site emphasizes a "community of practice" model for taking advantage of wiki software. People who have some engagement in a particular subject or project can benefit from a central online point of contact rather than attempting communication through a series of email conversations.

Wikis and other online collaboration tools can help maintain a dialog with academics and outside experts. Wiki pages on technical issues related to the database could generate a valuable two-way flow of information about technical issues between outside researchers and staff experts at NCSES.

KEEPING USERS INFORMED

The current NCSES websites and published reports appropriately point users to technical descriptions of the data collections and identify staff who are ready and able to assist users in their use of the data. However, a perusal of other federal statistical agency websites identifies useful information sharing. For example, the Census Bureau's Manufacturing and Construction Division, which manages the Business Research and Development and Innovation Survey (BRDIS) for NCSES, includes on its website (http://www.census.gov/mcd/clearance/) a listing of the open opportunities for public comment noted in the *Federal Register*, identifies planned changes, and includes copies of the forms and the supporting documents as submitted to the Office of Management and Budget (OMB). The technical information in these OMB clearance packages can assist users in understanding the strengths and weaknesses of the data.

ENHANCING USABILITY OF NCSES DATA

Usability is generally understood to be the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use (ISO 9241-11). The field of website usability is developing rapidly and now includes sophisticated methods to gather feedback from users about their interactions with websites.

Although there is no single broad measure of web-site usability, some very useful guidelines are contained in a federal government publication, *Research-Based Web Design and Usability Guidelines*, prepared jointly by the U.S. Department of Health and Human Services and the General Services Administration (U.S. Department of Health and Human Services and the General Services Administration, undated). Identified on the web as Usability.gov, the publication contains guidelines that emphasize the need to design websites "to facilitate and encourage efficient and effective human-computer interactions." The guidelines call on website designers to strive to reduce the user's workload by taking advantage of the computer's capabilities on the premise that users will make the best use of websites when information is displayed in a directly usable format and content organization is highly intuitive.

The guidelines make the point that task sequences make a difference. The sequencing of user tasks need to be consistent with how users typically do the tasks for which they have visited the site, using techniques that do not require them to remember information for more than a few seconds, employing terminology that is understandable, and taking care to refrain from overloading users with information. Likewise, users should never be sent to unsolicited windows or unhelpful graphics. The guidelines emphasize that speed in accomplishing a task is

important, and users should not have to wait for more than a few seconds for a page to load, and, while waiting, should be supplied with appropriate feedback. Tasks like printing information should be made easy.[2]

Evaluation of the NCSES Website

In order to assess how well the current NCSES web site (http://www.nsf.gov/statistics/) fulfills these basic usability guidelines and criteria, the panel conducted an evaluation of the site as it appeared in May 2011. This review was by no means exhaustive. Rather, the goal was to stimulate the development of a formal usability process by briefly reviewing the current design. Clearly, further user research would be necessary prior to making improvements to the current design. Any decision to change the website's design, including content and organization, must be based on user feedback and a usability evaluation testing strategy, which is presented at the end of this review.

It is apparent that having the NCSES web pages as a subsite of the NSF website poses limitations for NCSES website designers. If not treated carefully, this fact of life could increase the difficultly of navigating the site for NCSES data users. For example, the design of the NCSES tab "Statistics" is a path to a different site altogether. The visual cues indicating for users that they are still within the NSF website is the use of the same visual design template (same header, footer, and title format with image), which is crucial.

However, the main issue with this design is that users can have some difficulty finding NCSES if their point of entry is the NSF home page. From the NSF home page, users are expected to find what they are looking for by exploring the site through main and secondary navigation.

Once users find the NCSES subsite (from the NSF home page or directly via an Internet search engine or bookmark), users are faced with an organization-centric site rather than a user-centric site based on tasks. The current design appears to try to educate return visitors on how to navigate the site, but it would be best to organize the site in a way that all users (frequent, infrequent, or new users) can quickly and efficiently accomplish the task they are setting out to do. Suggestions for reorganizing the NCSES subsite appear in Appendix B.

On the whole, the evaluation of the NSF website points to the need for more systematic user-centered design and more regular usability evaluation. There are a number of methods in use, including expert heuristic evaluation, usability testing with small samples of actual users, and large-scale web browsing analytics.

A heuristic evaluation is recommended as an initial approach. It is one lightweight method that web designers use for discovering usability problems in a user interface design (electronic or paper prototypes), so that problems can be addressed throughout an iterative design process. The evaluation is usually employed early in the design process—the earlier, the better. Jakob Nielson, an expert in this field, recommends having between three and five evaluators separately review the interface design. The number of issues discovered increases with each evaluator, but the cost-benefits begin to decrease after five (Nielsen, 1994; Nielsen and Molich, 1990). Along with the customer surveys and focus groups recommended in Recommendation 4-1, the heuristic evaluation can intelligently inform the process of designing a more effective and efficient website.

[2] See http://www.usability.gov/pdfs/chapter2.pdf.

Recommendation 4-3. The National Center for Science and Engineering Statistics should employ user-focused design and user analysis, starting with an initial heuristic evaluation and continuing as a regular and systematic part of its website and tool development.

Meeting Compliance Standards

Websites should be designed to ensure that everyone, including users who have difficulty seeing, hearing, and making precise movements, can use them. Generally, this means ensuring that websites facilitate the use of common assistive technologies. As a federal government agency, NSF is governed by the Section 508 regulations. These amendments to the Rehabilitation Act require federal agencies to make their electronic and information technology accessible to people with disabilities. Section 508 was enacted to eliminate barriers in information technology, to make available new opportunities for people with disabilities, and to encourage development of technologies that will help achieve those goals. The U.S. Access Board has responsibility for the Section 508 standards and has announced its intention to harmonize the web portions of its Section 508 regulations with Web Content Accessibility Guidelines (WCAG) 2.0, for which the Web Accessibility Initiative (WAI) has responsibility. Statistical Policy Directive Number 4 (March 2008) directs statistical agencies to make information available to all in forms that are readily accessible.[3]

Some of the major accessibility issues to be dealt with include:

- Provide text equivalents for nontext elements;
- Ensure that scripts allow accessibility;
- Provide frame titles;
- Enable users to skip repetitive navigation links;
- Ensure that plug-ins and applets meet the requirements for accessibility; and
- Synchronize all multimedia elements.

When it is not possible to ensure that all pages of a site are accessible, designers should provide equivalent information to ensure that all users have equal access to all information.[4] Other standards include the "web accessibility initiative" of the World Wide Web Consortium (W3C), which provides guidance and tools for a range of websites and applications. Even more significant, given the possibility for rich dynamic interaction with these data resources, is that W3C has also developed standards for access to dynamic content, with specific guidelines in four categories:

1. *Accessible rich Internet applications*: address accessibility of dynamic web content, such as those developed with Ajax, dynamic HTML, or other such technologies;
2. *Authoring tool accessibility guidelines*: address the accessibility of the tools used to create websites;
3. *User agent accessibility guidelines*: address assistive technology for web browsers and media players; and

[3] A summary of Section 508 is available online at http://www.section508.gov/index.cfm?fuseAction=stdsSum (retrieved November 2010).
[4] U.S. Department of Health and Human Services, *Research-based Web Design and Usability Guidelines*, P. 23. 2008. See http://www.usability.gov/guidelines/guidelines_book.pdf (retrieved on May 9, 2011).

4. *Web content accessibility guidelines*: address the information in a website, including text, images, forms, and sounds.

The convention when considering web design for individuals with disabilities is to ensure that the site is accessible to those who are visually impaired. However, there is a much wider range of ways in which someone's access to information should be considered when developing websites and web applications. For example, a chart that is color-coded may not be readily interpreted by someone with color blindness, multimedia files may not be accessible to someone with deafness unless they are accompanied by transcripts, and someone with a cognitive disability, such as attention deficit disorder, may find websites that lack a clear and consistent organization difficult to navigate.[5]

Data Accessibility Issues

The accessibility of tabular data and data visualization is an open research question. Although W3C has pioneered standards for accessibility of dynamic user interfaces, many other issues, including table navigation, navigation of large numeric data sets, and dynamic data visualization, raise computer-human interaction challenges that have been explored only peripherally. The issue of accessibility is a clear opportunity for NSF to partner with scientists with disabilities and those who work on interface design and so lead by example.

In order for NSF S&E information to be used, it must be accessible to users. By nearly eliminating the hard-copy publication of the data in favor of electronic dissemination, mainly through the web, NSF is committed to the provision of web-based data in an accessible format, not only for trained sophisticated users, but also for users who are less confident of their ability to access data on the Internet. Importantly, the user population includes people with disabilities for whom, by law and right, special accommodations need to be made.

The panel benefited from a presentation by Judy Brewer, who directs the WAI at W3C. W3C hosts the WAI to develop standards, guidelines, and resources to make the web accessible for people with disabilities; ensure accessibility of W3C technologies (20-30 per year); and develop educational resources to support web accessibility.

Brewer stated that Web 2.0 adds new opportunities for persons with disabilities, and that data visualization is a key to effective communication. However, people with disabilities face a number of barriers to web accessibility, including missing alternative text for images, missing captions for audio, forms that "time out" before they can submit them, images that flash and may cause seizures, text that moves or refreshes before they can interact with it, and websites that do not work with assistive technologies that many people with disabilities rely on.

In response to a question, Brewer addressed the continued problem of making tabular information accessible, and she requested input on where the WAI should go in this area. She referred to a workshop held by the National Institute of Standards and Technology on complex tabular information that resulted in several recommendations.

Brewer argued for publishing existing S&E data in compliance with Section 508 requirements, while continuing R&D on accessibility techniques for new technologies, improved accessibility supports for cognitive disabilities, and more affordable assistive technologies, such

[5] Presentation of Judy Brewer, director of the Web Accessibility Initiative at the W3C, on the issue of accessibility of information on the web.

as tablets. She said WAI would partner with agencies to ensure that dissemination tools are accessible.

Recommendation 4-4. The National Science Foundation should sponsor research and development on accessible data visualization tools and approaches and potential other means for browsing and exploring tabular data that can be offered via web, mobile, and tablet-based applications, or browser-based ones.

5
The Way Ahead

These are exciting and challenging times for federal government statistical agencies responsible for disseminating their data products to their user communities. The times are especially challenging for the National Center for Science and Engineering Statistics (NCSES), which is finding the importance of its data magnified many fold by the growing recognition of the role that science and engineering (S&E) investment is playing as a source of economic and social growth and prosperity But these are also uncertain times for federal government agencies like NCSES that are concerned over the future of their programs in light of fixed or declining budgets associated with the need to restrain government spending. There is a simultaneous growth in pressure to carefully evaluate all government activities to ensure their efficiency and cost-effectiveness. A key component of efficiency and effectiveness is a well managed and responsive data dissemination program.

The environment for the data dissemination program for NCSES is also in flux. The agency is confronting new roles and missions as directed in the America COMPETES Act, which changed the agency's name and added significant new responsibilities. For example, the newly specified role of serving as a central federal clearinghouse for the collection, interpretation, analysis, and dissemination of objective data on science, engineering, technology, research and development, and innovation suggests a need for the agency to become more strategic in its outlook. NCSES will be venturing into new territory and will need to support a broader range of data users, particularly in areas of competitiveness and innovation, even as it seeks to modernize the dissemination services it now provides. The key to accomplishing these ends in an era of expected budget shortfalls and in view of the limited staff resources in the agency, including some of the technological skills that will be required to modernize the data processing and dissemination systems, is to take advantage of consortia opportunities and to proceed within a framework that accords priority to the most essential tasks.

STRENGTH IN NUMBERS

The task of developing and implementing a dissemination improvement plan is a tall order for NCSES to take on by itself. The agency is already stressed, with its constrained staff and budget resources, to meet the growing demand for its data and implement the several new areas of responsibility that have recently been added to its roles and missions.

One of several possible approaches to meet the needs of data users as well as to encouraging and expanding development of tools and applications that would facilitate the dissemination of its information by developers and dissemination channels is to take the necessary steps in concert with other agencies in the federal statistical community. The federal statistical agencies, as a group, have begun to organize to enhance dissemination of their data in the project called the Statistical Community of Practice and Engagement (SCOPE). SCOPE is an important beginning. There are efficiencies for both the agencies and users from more cross-agency collaboration, harmonization of definitions and terminology, identification of best practices, and sharing of the development of common tools that support best practices. As a participant in this community of practice, NCSES could maximize use of the capacity of Data.gov for service as a primary public interface and dissemination platform/portal, retrieval of datasets on the Data.gov dataset hosting platform that is currently being developed, and harness Data.gov cloud computing power.

NCSES should also consider taking advantage of commonly-developed, user-friendly data delivery and data display tools that have largely been developed by the World Wide Web Consortium (W3C) community. These tools address 508 compliant alternatives to tabular displays, develop displays of complex sample survey data while protecting confidential microdata, and develop visualization tools for multifaceted statistical designs. And it can benefit from such projects as promoting data harmonization and integration through the development of metadata and data exchange. Specifically, SCOPE will take the fundamental steps of developing and implementing Stats Metadata 1.0 (for delivery in fiscal 2012) and establishing common definitions to facilitate data exchange and interoperability (by fiscal 2013). The goal is to promote development and use of common platforms for data collection and data analysis and to suggest research on solutions to the "data mosaic" problem in the current technology environment and support the creation of an open-source development community.

TIME-PHASED DISSEMINATION IMPROVEMENT PLAN

The panel understands that not every recommendation made in this report can or should be implemented immediately. Some recommendations must build on the implementation of others; for example, development of an open data base structure that can support accessibility and dissemination through the use of open standards and formats requires that NCSES obtain from its contractors the data sufficient to make the results reproducible, in a format enabling automatic reproduction of all published tables, along with metadata sufficient to interpret the data elements and results.

The implementation of the report's recommendations should be undertaken within an overall framework that accords priority to the basic quality of the data and the fundamentals of dissemination, then to significant enhancements that are achievable in the short term, while laying the groundwork for other long-term improvements. The framework could be organized along the following lines (highest priority first):

(1) Focus on collecting the right data (by contractor or otherwise); using appropriate change management and version control to establish data provenance, flag data errors and correct them; annotating that data with sufficient machine-actionable metadata to establish a process for interpreting the data, enabling efficient access to third-party data

and to automated NCSES publications; and publishing the data in formats with web-accessible open interfaces for all to use.

(2) Publish methods for combining old data and new data that have been collected under different assumptions or categories or that are disseminated in ways that make them difficult to reintegrate—this is especially necessary for the data from the old and new industry research and development expenditure surveys that will populate the Industrial Research and Development Information System (IRIS).

(3) Provide the essential data reductions and visualizations that the mission of the National Science Foundation (NSF) requires, for example, when Congress asks for authoritative data on a certain topic, a trusted group must be able to use the data and derived publications to calculate answers.

(4) Provide a growing array of visualizations and printed products tailored for the many different uses and users.

Within this overall framework, three parallel tracks are suggested with concrete steps to improve data dissemination. The first track involves improving the transparency and reproducibility of published and disseminated results by obtaining complete, reliably versioned, well-documented, and machine-understandable data from contractors. This will require the modification of current contractual arrangements and procurements as referenced in the panel's recommendations. The second track involves improving use of the NCSES products by establishing a formal, systematic, and continuous program for evaluating user needs and the usability of NCSES products via the web and other means of delivery. The third track involves ensuring full short- and long-term access to NCSES content by providing open data, offering machine-accessible protocols for access to data and other products, and establishing a continuous process for replicating or archiving releases by the National Archives and Records Administration for long-term preservation and access.

IMPROVING THE TRANSPARENCY AND REPRODUCIBILITY OF PUBLISHED RESULTS

As noted in earlier chapters, it is not currently possible to automatically and systematically reproduce or validate all tables and results in NCSES published products from the raw data. There are many contributing causes: not all data are made available to NCSES at the level of detail at which they were collected, data are not accompanied by machine-readable metadata, and there is a lack of a systematic version control/change-management process for the data prior to final delivery by contractors.

The root cause of this problem, as we have identified, is insufficient accountability from contractors. Contractors are not delivering the data and metadata in the detail most needed, and they are not supplying sufficient metadata, provenance information, or change management. Strengthening accountability from contractors is a first step to any improvement in transparency.

This should be followed by more systematic development of metadata standards, change management and versioning, and provenance tracking. These need not be perfect; any open, transparent, machine-understandable, automatic method could be used. And these can be then improved.

As part of improving metadata standards, NCSES should actively participate in the development and implementation of the Data.gov compatible metadata standard now being

explored by W3C and the SCOPE project. Implementation of this standard, as discussed in this report, will require revamping the specifications for data delivery now in the contracts of the agency's data collectors.

ESTABLISHING A FORMAL, SYSTEMATIC, AND CONTINUOUS USE AND USABILITY EVALUATION PROGRAM

We have pointed to the need for a continuous use and usability evaluation program, much akin to pointing to the need for a program of continuous improvement that is part and parcel of any total quality management program. We focus on use and usability because, like other federal statistical agencies, as NCSES continues to shed its hard-copy publication programs in favor of providing its data through web applications, usability will become a more important issue, and new uses and users have begun to be identified.

A first step is to develop a clearer understanding of requirements. In the first instance, the requirements for an NCSES dissemination program are essentially determined by the environment facing the agency, its legislative mandate, and guidance and directives from above. These are assessed in Chapter 1. The more difficult, but nonetheless important part of establishing a requirement is to understand the needs of its customers—the data users. As discussed in Chapter 4, NCSES today has only a rudimentary understanding of the range of its users and their data needs. Thus, the first step in the plan must be to gain a better understanding of the users of the data—those primary, secondary and tertiary blocks of users—and then to engage them in an effort to understand their needs. Some steps have already been taken to enhance engagement of user groups. The measures of website use and the new online survey of web users are important and necessary first steps, but they are by no means sufficient to provide the kind of detailed knowledge NCSES needs. Agency leadership would be well advised to monitor the maturing space of web metrics and analytics. These, along with customer service programs, would enable continuous input, evaluation, and understanding of all users and their products.

The learnings from these outreach activities should then be widely shared. One possible activity would be to glean and post some kind of listing of user sites that have distilled the NCSES basic data, aggregated them, or combined them with other data. Although these derived forms cannot carry the NSF imprimatur of accuracy, they can be very helpful.

A suggested next step is to review the initiatives taken by Statistics Canada to evaluate the usability of its delivery methods. Tied in with usability, we urge attention to issues of accessibility for all users, with the understanding that 508 compliance is a necessary but insufficient first step.

We make several suggestions in Chapter 4 and Appendix B for enhancing the visitor's experience with the NCSES website. Some of these suggestions can be implemented by NCSES; others will require coordination with the NSF organizations that establish the basic look and feel of the website.

ENSURING FULL SHORT- AND LONG-TERM ACCESS

As discussed in Chapter 3, the Internet changes the meaning of access. Ensuring full access in today's environment requires that, as much as possible, machine-understandable

microdata and metadata be made accessible via standard open protocols to any third party for use without restriction.

The power of visualization tools to retrieve and explain the data leads to the suggestion that a major emphasis throughout the implementation period should be on providing data that can be easily accessed by visualization tools. We do suggest that NCSES develop visualizations beyond the kind of rudimentary ones that it already provides in the *Science and Engineering Indicators Digest*. Rather, the agency should provide data in machine-accessible formats and explore partner relationships in the private sector to identify opportunities to leverage developing or existing tools/applications, along with maintaining open data formats and standards to allow individual users to import the data into their visualization sets. By adopting an approach that stresses the basics of data provision (common formats with appropriate metadata) and partnerships with the private sector as opportunities become available, the NCSES will avoid the issue of rapid obsolescence associated with rapid change in the particular tools and systems offered by the private sector.

Ensuring long-term access requires that both the NCSES publications and all of the data necessary to fully replicate them be archived. NSF should work with the National Archives and Records Administration, as the archive of record, to ensure that copies of all products and data, including those created by contractors, are efficiently delivered for long-term stewardship.

RAPID ITERATIVE IMPROVMENTS

The recommendations in this report will take several years to implement. However, the groundwork can be laid, and many improvements made, in a relatively short amount of time, even in the first year. We suggest that at least the following be accomplished in the first year:

- Establish an ongoing archiving process;
- Revise contracts with data providers to ensure accountability for delivery of full microdata in machine-understandable format with change control;
- Perform a heuristic evaluation of the website;
- Initiate a process of continuous usage/user data needs collection; and
- Disseminate existing microdata available using standard open machine protocols.

We expect that improvement will be iterative, and will primarily stem from development of further technologies, methods, and standards and from the collection of systematic information on user behavior and needs. In light of this, other recommended tasks can be deferred, awaiting further developments in technology or methods. For example:

- Redesigning the NCSES website can await heuristic evaluation.
- Developing a detailed metadata standard, can await a candidate metadata standard from the SCOPE and World Wide We b Consortium initiatives.
- Creating a capacity for user-influenced visualizations can await further developments in accessible visualization technology.

The future well-being of the U.S. economy depends on the nation's capacity to generate, and take economic advantage of, technology-driven innovations across all industries, particularly those that compete internationally. This capacity in turn depends on choices that market actors,

including the federal government, firms large and small, educational and research institutions, state and regional technology-based development agencies, workers, and students, make with regard to research and development, development of the science, technology, engineering, and mathematical workforce, and the commercialization of innovation. The data generated by NCSES will guide these choices. The data dissemination strategy of the agency, then, will have a substantial influence on the nation's future economic path.

Technology is opening the door to significant advances in the ability to communicate data and analytical products to data users. The promise of such services as Data.gov and the potential for third-party services, such as the Google Public Data Explorer, and federated catalogs, such as the DataVerse Network, to add value to the data and make them accessible to new groups of users and for new uses are just becoming recognized. The emerging Semantic Web (Web 3.0), expanded and new tools and approaches, open standards and platforms, the potential for mashups, and community-based platforms (including participative input, transparency by means of wikis and open government movements) show a more distant promise of communicating data to users in entirely new ways, much to the advantage of users and the federal agencies themselves.

To avail itself of the opportunities afforded in these new approaches, NCSES needs to adopt a vision of the future that supports access to data directly through the agency and through the many third-party services and catalogs that are emerging. NCSES also needs to have a plan that will lead to making its data available through open interfaces and open formats, accompanied by open metadata, and to develop the necessary infrastructure to exploit these advances. These evolving technologies could open opportunities for addressing the visualization experience and overcoming accessibility limitations more effectively that the current browser-based experiences.

References

Altman, M., and J. Crabtree
2011 Using the SafeArchive System: TRAC-based auditing of LOCKSS. In *Proceedings of Archiving 2011*. Society for Imaging Science and Technology.

Altman, M., M. Adams, J. Crabtree, D. Donakowski, M. Maynard, A. Pienta, and C. Young
2009 Digital preservation through archival collaboration: The Data Preservation Alliance for the Social Sciences. *American Archivist* 72(1):169-182.

Altman, M., L. Andreev, M. Diggory, E. Kolster, M. Krot, G. King, D. Kiskis, A. Sone, and S.Verba
2002 An introduction to the Virtual Data Center Project and Software. Pp. 203-204 in *Proceedings of the The First ACM+IEEE Joint Conference on Digital Libraries* (JCDL 01), 2001. Roanoke, Virginia: ACM Press.

American Association for the Advancement of Science
2003 All American data trove. *Science* 301(5636):1025.

Berners-Lee, T., J. Hendler, and O. Lassila
2001 The Semantic Web. *Scientific American Magazine.* May.

Bostock, M., and J. Heer
2009 Protovis: A graphical toolkit for visualization. Pp. 1121-1128 in *IEEE Transactions on Visualization and Computer Graphics*.

Boskin, M.J., and L.J. Lau
1992 Capital, technology, and economic growth. In N. Rosenberg, R. Landau, and D.C. Mowery, eds., *Technology and the Wealth of Nations*. Stanford, CA: Stanford University Press.

Bosley, J., and C. Capps
2000 *Adapting Usability Test Methods to Improve a Data Extraction Tool (FERRETT)*. BLS Statistical Survey Papers. Retrieved on October 14, 2011 from http://www.bls.gov/ore/abstract/st/st000200.htm.

Capps, C., A. Green, and M. Wallace
1999 The vision of integrated access to statistics: The data web. *Of Significance* 1(2):42-47.

Global Confederation of Competitiveness Councils
2010 *Global Competitiveness Principles, 2010*. Washington, DC. Retrieved on April 27, 2011 from: http://www.compete.org/images/uploads/File/PDF%20Files/Embargoed_2010_Global_Competitiveness_Principles_(2).pdf.

Gutmann, M., M. Abrahamson, M.O. Adams, M. Altman, C. Arms, K. Bollen, M. Carlson, J. Crabtree, D. Donakowski, G. King, J. Lyle, M. Maynard, A. Pienta, R. Rockwell, L. Timms-Ferrara, and C. Young
2009 From preserving the past to preserving the future: The Data-PASS Project and the challenges of preserving digital social science data. *Library Trends* 57(3):315-337.

Heer, J., S.K Card, and J.A. Landay
2005 Prefuse: A toolkit for interactive information visualization. Pp. 421-430 in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery.

Howard, A.
2011 *BuzzData: Come for the Data, Stay for the Community*. O'Reilly Radar. September 20,
2011 Retrieved on September 30, 2011 from: http://radar.oreilly.com/2011/09/buzzdata-data-community.html.

Kanter, R.
2011 *The Internet Changes Everything—Except Four Things*. Retrieved on June 10, 2011, from http://blogs.hbr.org/kanter/201/05/the-internet-changes-everythin.html.

King, G.
2007 An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research* 36:173–99.

Koizumi, K.
2011 *R&D Dashboard Makes Federal R&D Data Transparent and Accessible.* Retrieved on June 10, 2011, from http://www.whitehouse.gov/blog/2011/02/10/rd-dashboard-makes-federal-rd-data-transparent-and-accessible.

Lane, J., and S. Bertuzzi
2010 The STAR METRICS Project: Current and future uses for S&E workforce data. Retrieved on April 27, 2011, from http://www.nsf.gov/sbe/sosp/workforce/lane.pdf.

Macdonald, S.
2009 *Data Visualization Tools: Part 1: Numeric Data in a Web 2.0 Environment*. Retrieved on September 30, 2011 from: http://ie-repository.jisc.ac.uk/303/1/Numeric_data_mashup.pdf.

Miller, J., Kimmel, L. and ORC Macro.
2004 National Science Foundation Surveys of Public Attitudes Toward and Understanding of Science and Technology, 1979-2001. National Science Foundation, Division of Science Resources Statistics. Arlington, VA. Obtained from: Inter-university Consortium for Political and Social Research, 2005. doi:10.3886/ICPSR04029.v1.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine
2005 *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Future.* Washington, DC: The National Academies Press.
2010 *Rising Above the Gathering Storm, Revisited: Rapidly Approaching Category 5*. Washington, DC: The National Academies Press.

National Center for Science and Engineering Statistics
1994 *Policy on Data Release*. SRS Memorandum, January 10. National Science Foundation.
No Date *NSF's Information Quality Guidelines for Section 515*. Retrieved on June 10, 2010, from http://www.nsf.gov/policies/nsfinfoqual.pdf.

National Research Council
2000 *Measuring the Science and Engineering Enterprise: Priorities for the Science Resources Studies Division*. Washington, DC: National Academy Press.
2004 *Measuring Research and Development Expenditures in the U.S. Economy*. Panel on Research and Development Statistics at the National Science Foundation. Washington, DC: The National Academies Press.
2005 *Expanding Access to Research Data: Reconciling Risks and Opportunities.* Panel on Data Access for Research Purposes. Washington, DC: The National Academies Press.
2009 *Principles and Practices for a Federal Statistical Agency, Fourth Edition*. Committee on National Statistics, C.F. Citro, M.E. Martin, and M.L. Straf, eds. Washington, DC: The National Academies Press.
2010 *Data on Federal Research and Development Investments: A Pathway to Modernization*. Panel on Modernizing the Infrastructure of the National Science Foundation Federal Funds Survey. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
2011 *Communicating National Science Foundation Science and Engineering Data to Users*: Letter Report. March 9, 2011. Retrieved on March 18, 2011, from http://www.nap.edu/catalog.php?record_id=13120.

National Science Board
2010a *Science and Engineering Indicators 2010*. Arlington, VA. NSB 10-01. Retrieved on July 22, 2011, from http://www.nsf.gov/statistics/seind10/.
2010b *Key Science and Engineering Indicators, 2010 Digest*. NSB 10-2. Retrieved on July 22, 2011, from http://www.nsf.gov/statistics/digest10/nsb1002.pdf.

National Science Foundation
2011 *NSF Strategic Plan for Fiscal Years (FY) 2011-2016*. April. Retrieved on June 27, 2011, from http://www.nsf.gov/news/strategicplan/nsfstrategicplan_2011_2016.pdf.

Office of Management and Budget
2000 *Management of Federal Information Resources, Circular A-130*. Retrieved on May 4, 2011, from http://www.whitehouse.gov/omb/circulars_a130_a130trans4.
2006 *Standards and Guidelines for Statistical Surveys, September*. Retrieved on May 4, 2011, from http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.

Powers, S.
2003 *Practical RDF*. Sebastopol, CA: O'Reilly.

Pruit, J., and T. Adlin
2006 *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. San Francisco: Morgan Kaufmann.

Rosson, M.B., and J.M. Carroll
2002 Usability Engineering: Scenario-based Development of Human-Computer Interaction. London: Academic Press.

Tufte, E.
2004 *The Visual Display of Quantitative Information, 2nd ed.* Chesire, CT: Graphics Press.

Reas, C., and B. Fry
2007 *Processing: A Programming Handbook for Visual Designers and Artists*. Cambridge, MA: MIT Press.

U.S. Department of Health and Human Services and General Services Administration
No Date *Research-Based Web Design and Usability Guidelines.* Retrieved on August 15, 2011, from http://www.usability.gov/guidelines/guidelines_book.pdf.

Viegas, F., M.. Wattenberg, F. van Ham, J. Kriss, and M. McKeon
2007 ManyEyes: A site for visualization at Internet scale. *IEEE Transactions on Visualization and Computer Graphics* 13(6):1121-1128. Retrieved on July 25, 2011, from: http://www.computer.org/portal/web/csdl/doi/10.1109/tvcg.2007.70577.

Ware, C.
2004 *Information Visualization, Second Edition: Perception for Design*. San Francisco: Morgan Kaufman.

Wilkinson, L., D. Wills, D. Rope, A. Norton, and R. Dubbs
2005 *The Grammar of Graphics, 2nd ed.* New York, NY: Springer.

World Wide Web Consortium
2011 *W3C Semantic Web Activity*. Retrieved June 22, 2011, from http://www.w3.org/2001/sw/.

Appendix A
Acronyms and Abbreviations

API	Application Programming Interface
ARC	Archival Research Catalogue
BRDIS	Business Research and Development and Innovation Survey
CATI	Computer Assisted Telephone Interviewing
EDI	Electronic Data Interchange
EIM	Enterprise Data/Information Management
EUROSTAT	Statistical Office of the European Community
GDP	Gross Domestic Product
GSA	General Services Administration
GPO	Government Printing Office
HTML	Hyper Text Markup Language
ICPSR	Interuniversity Consortium for Political and Social Research
IRIS	Industrial Research and Development Information System
NARA	National Archives and Records Administration
NCSES	National Center for Science and Engineering Statistics
NIH	National Institutes of Health
NORC	National Opinion Research Center
NSCG	National Survey of College Graduates
NSF	National Science Foundation
NSRCG	National Survey of Recent College Graduates
NTIS	National Technical Information Service
OMB	Office of Management and Budget
OPA	Online Public Access
OSTP	Office of Science and Technology Policy
OWL	Web Ontology Language
PRA	Paperwork Reduction Act
R&D	Research and Development
RDF	Resource Description Framework
S&E	Science and Engineering
SciSIP	Science of Science and Innovation Policy
SDR	Survey of Doctoral Recipients

SCOPE	Statistical Community of Practice and Engagement
SED	Survey of Earned Doctorates
SESTAT	Scientists and Engineers Statistical Data System
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WebCASPAR	Integrated Science and Engineering Resource Data System
XML	Extensible Markup Language

Appendix B
Suggestions for Improving the Website[1]

The information currently visible to users on the landing (home) page of the National Center for Science and Engineering Statistics (NCSES) includes the following elements:

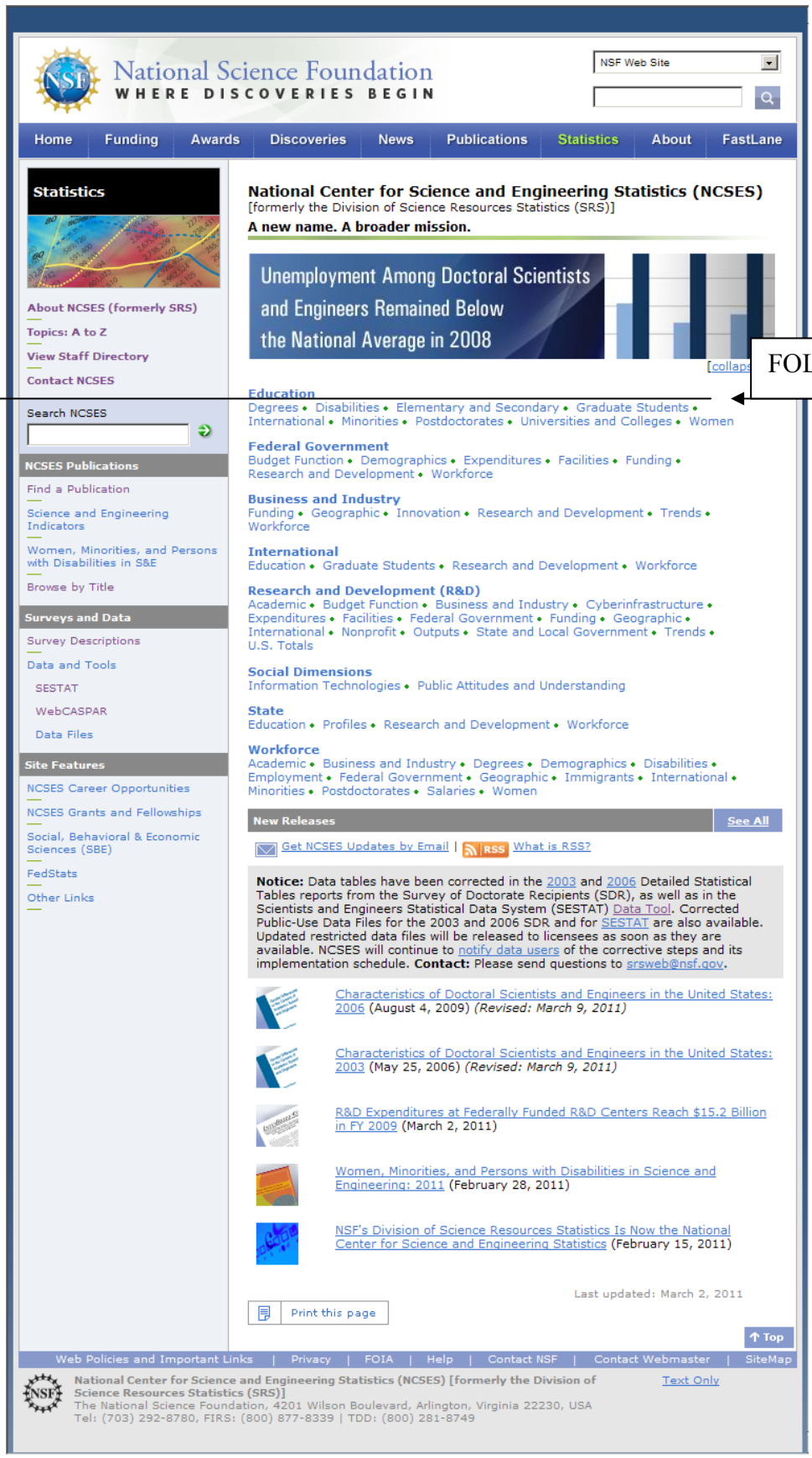
- The title
- The rotating banner
- The new name and mission
- *Women, Minorities, and Persons with Disabilities (report)*
- *Unemployment Among Doctoral Scientists and Engineers (InfoBrief)*
- The left-side navigation
- About NCSES
- Other links for contacting NCSES

To find the appropriate links for accomplishing major user-identified tasks beyond accessing the report and *InfoBrief*, users have to scroll "below the fold."[2]

Suggestion: NCSES should place the most important information, based on user feedback, at the top of the page.

[1] This appendix was prepared by panel member Dianne Fournier.

[2] The part of a web page that is visible in the web browser window when the page first loads is described as being "above the fold." See *The Motive Web Design Glossary,* August 6, 2009. Retrieved on August 10, 2011, from: www.motive.co.nz.


National Science Foundation
WHERE DISCOVERIES BEGIN
NSF Web Site
Home
Funding
Awards
Discoveries
News
Publications
Statistics
About
FastLane
Statistics
About NCSES (formerly SRS)
Topics: A to Z
View Staff Directory
Contact NCSES
Search NCSES
NCSES Publications
Find a Publication
Science and Engineering Indicators
Women, Minorities, and Persons with Disabilities in S&E
Browse by Title
Surveys and Data
Survey Descriptions
Data and Tools
SESTAT
WebCASPAR
Data Files
Site Features
NCSES Career Opportunities
NCSES Grants and Fellowships
Social, Behavioral & Economic Sciences (SBE)
FedStats
Other Links
National Center for Science and Engineering Statistics (NCSES)
[formerly the Division of Science Resources Statistics (SRS)]
A new name. A broader mission.
Unemployment Among Doctoral Scientists and Engineers Remained Below the National Average in 2008
[collaps
FOLD
Education
Degrees • Disabilities • Elementary and Secondary • Graduate Students • International • Minorities • Postdoctorates • Universities and Colleges • Women
Federal Government
Budget Function • Demographics • Expenditures • Facilities • Funding • Research and Development • Workforce
Business and Industry
Funding • Geographic • Innovation • Research and Development • Trends • Workforce
International
Education • Graduate Students • Research and Development • Workforce
Research and Development (R&D)
Academic • Budget Function • Business and Industry • Cyberinfrastructure • Expenditures • Facilities • Federal Government • Funding • Geographic • International • Nonprofit • Outputs • State and Local Government • Trends • U.S. Totals
Social Dimensions
Information Technologies • Public Attitudes and Understanding
State
Education • Profiles • Research and Development • Workforce
Workforce
Academic • Business and Industry • Degrees • Demographics • Disabilities • Employment • Federal Government • Geographic • Immigrants • International • Minorities • Postdoctorates • Salaries • Women
New Releases
See All
Get NCSES Updates by Email | RSS What is RSS?
Notice: Data tables have been corrected in the 2003 and 2006 Detailed Statistical Tables reports from the Survey of Doctorate Recipients (SDR), as well as in the Scientists and Engineers Statistical Data System (SESTAT) Data Tool. Corrected Public-Use Data Files for the 2003 and 2006 SDR and for SESTAT are also available. Updated restricted data files will be released to licensees as soon as they are available. NCSES will continue to notify data users of the corrective steps and its implementation schedule. Contact: Please send questions to srsweb@nsf.gov.
Characteristics of Doctoral Scientists and Engineers in the United States: 2006 (August 4, 2009) (Revised: March 9, 2011)
Characteristics of Doctoral Scientists and Engineers in the United States: 2003 (May 25, 2006) (Revised: March 9, 2011)
R&D Expenditures at Federally Funded R&D Centers Reach $15.2 Billion in FY 2009 (March 2, 2011)
Women, Minorities, and Persons with Disabilities in Science and Engineering: 2011 (February 28, 2011)
NSF's Division of Science Resources Statistics Is Now the National Center for Science and Engineering Statistics (February 15, 2011)
Last updated: March 2, 2011
Print this page
Top
Web Policies and Important Links | Privacy | FOIA | Help | Contact NSF | Contact Webmaster | SiteMap
National Center for Science and Engineering Statistics (NCSES) [formerly the Division of Science Resources Statistics (SRS)]
The National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230, USA
Tel: (703) 292-8780, FIRS: (800) 877-8339 | TDD: (800) 281-8749
Text Only

NCSES landing (home) page: three different titles are used to describe the NCSES subsite: "Statistics" (National Science Foundation, NSF, tab title), "Science and Engineering Statistics" (NSF home page), and the "National Center for Science and Engineering Statistics (NCSES)." This is a problem for users who are asked to make the leap between "Statistics" and "NCSES."

Although the first item in the "Statistics" drop-down menu, "Science & Engineering Statistics," explains more fully the type of statistics found on this site by using a more descriptive title, this link is not needed, since it links to the current landing page. This behavior does not respect two usability standards: (1) a hyperlink on a page should not send users to the current page and (2) two hyperlinks that are named differently should bring users to different content pages.

Suggestion: Consider testing the current version with users (frequent and infrequent) through task completion exercises to find out whether navigation is difficult for them given that the information scent[3] is reduced with the use of different titles. Test with a second version in which the tab name "Statistics" is changed to "Science and Engineering Statistics." This would be consistent with the name on the right-side navigation of the NSF home page (http://www.nsf.gov/) and would allow the user to eliminate the first item in the drop-down menu or eliminate the drop-down menu altogether, since Search statistics and About statistics are covered elsewhere.

Breadcrumb trail: a breadcrumb trail is a row of links showing how the site is structured. It is usually located at the top of the page. It prevents users from feeling lost in the site, especially if they arrive deep into the site from an Internet search engine or from a saved bookmark and do not know where they are in relation to the full NSF site.

Suggestion: Add a breadcrumb trail at the top of the page.

Left-side navigation: the first thing that catches the eye is the title with the graphic. Although the graphic provides a splash of color on the page, it is not necessarily the best use of this prime real estate, as it is not communicating information. Consider removing the image or using the image as a background to the title bar. (It is recognized that this will require changes to the NSF design.)

Suggestion: If, through user testing, NCSES finds that users respond better to changing the title "Science and Engineering Statistics," then use the title here. NCSES should consider removing the image if the NSF template allows for it to be removed. This will bring the menu items further up the page, and more links will be visible above the fold. The new design should consider removing the hyperlink from the "Statistics" left-side navigation title.

[3] The information scent predicts a path's success. The navigation of the page with good information scent will signal to the user that they have reached or are nearing their goal. Retrieved October 22, 2004, from *The Motive Web Design Glossary,* www.motive.co.nz.

Positioning of the links: the links "About NCSES (formerly SRS)," "Topics: A to Z," "View Staff Directory," and "Contact NCSES," although critical on any web page, are usually placed at the bottom of the page in the footer or in the left-side navigation because they are applicable to the whole website. Such commonplace links should be left out of the primary navigation and placed in secondary navigation at the bottom of the page, out of the way of primary tasks. These links "About NCSES," "Topics A to Z," and the like are not competing tasks so they should be places at the bottom of the page. If user-centered design guidelines are followed, the number of users who need to contact NCSES for help will decrease.

Suggestions: Place the "About NCSES," "Topics: A to Z," "View Staff Directory," and "Contact NCSES" at the bottom of the menu, since the footer is already occupied with NSF Help and Contact information. Add "About NCSES" to the NSF "About" tab drop-down menu.

Search NSF and NCSES: usability standards suggest that there should only be one search box on a page. It can be confusing for users to have both the site search and the NCSES search boxes on the NCSES landing page.

Currently the NSF site search box can accommodate 23 characters. Standards suggest that the minimum should be 27 characters.

Another usability standard recommends that different hyperlinks on a page with the same name, or suggesting the same name, should link to the same page. Currently, the "Search NCSES" in the left-side navigation and the "Search Statistics" in the tab drop-down menu bring users to different applications. At least the pages look different, suggesting that the applications are different. It would be best to make NCSES search an option in the NSF site search drop-down menu. That way, the search bar in the left-side navigation and the "Search Statistics" link could be removed. It is essential that users can locate and use search functionality effortlessly.

Suggestion: Consider removing the NCSES search box, unless the NSF search engine does not contain NCSES content; consider changing the arrow button of the site search to a larger button that is raised with the word "Search" on it. (This adds affordance by ensuring that a button with round corners makes it look like it is clickable.) Consider removing the NSF Search drop-down menu, since it does not contain the NCSES search option or, alternatively, add the NCSES search option.

NCSES publications: between the NSF site search, the NCSES subsite search, the NSF Publication search, and the NCSES publication search, there are too many search options available on the site. Which one should a user try first? Which one will provide the expected results? How many search options will a user try if they are unsuccessful with their first choice, their second choice, and so forth? The methods for searching and browsing with each one are very different.

The use of facets when finding a NCSES publication works very well. Combining such facets with the NSF site search could be very interesting. It is therefore unclear why more than one publication search is necessary.

Similarly, it is unclear why the "Science and Engineering Indicators" and the "Women, Minorities, and Persons with Disabilities in S&E" links are available under the NCSES Publications section of the menu, inserted between the "Find a Publication" and "Browse by

Title" links. These high-demand publications could be more effectively featured by adding the pointers as features on the rotating banner and/or in the center of the home page.

Suggestion: Consider deleting this section of the left-side navigation if NCSES publications can be searched or browsed under the NSF site search. Allow users to quickly link to the most sought-after publications by including them in the rotating banner or in the center page.

Surveys and Data: in this area of the left-side navigation, the "Data and Tools" heading links to the same page as the "Data Files" link. As stated above, two different hyperlinks should bring users to different content pages. Having all of the data and tools listed in the left-side navigation would allow users to directly link to these data and tools without having to go through their description first. This could simplify the navigational path for frequent users.

Suggestion: Remove "Data and Tools" heading because it leads to the same page as "Data Files." Change the links under this heading to read:

- Survey Descriptions
- Data and Tools
    - Microdata files
    - Public-use files
    - Integrated Science and Engineering Resource Data System (WebCASPAR)
    - Industrial Research and Development Information System (IRIS)
    - Survey of Earned Doctorates (SED) Tabulation Engine
    - Scientists and Engineers Statistical Data System (SESTAT)

Subject section in the left-side navigation: a subject section in the left-side navigation could contain the sections currently on the landing page. Doing so allows for the main tasks to appear in the center of the page.

Suggestion: Consider moving the subjects from the center of the page to the left-side navigation to ensure that subjects are available to users regardless of the page they are viewing.

Site features: it would be preferable to change "Other Links" to "Links to Other Sites," because it should be clear to users that taking this navigational path will eventually lead the user away from the NSF site. The FedStats link does not indicate that users will leave the NSF site. NCSES might want to use an icon to indicate this.

Suggestion: Consider changing "Other Links" to "Links to Other Sites." Consider adding an image indicating that users will be leaving the NSF site if they click on "FedStats."

Rotating banner: the banner currently contains three features. It is unclear why these were chosen and how often they remain as features.

A collapse option for the banner is a really good idea because many users find rotating banners distracting. Adding an indicator as to the number of features as well as a pause button, giving the user further control of rotation, would be beneficial for users. However, when viewing

on a smartphone, the rotating banner is not displayed, leaving the [collapse] link floating on the page without purpose.

Suggestion: Consider testing the visual design on different platforms.

Minimalist Design: content adds noise to a web page, and too much noise reduces the usability of the page because of the overload of information. Therefore, dialogue that is rarely needed and competes with relevant information should be deleted or made available to users as they progressively drill down through the different layers of detail.

When writing for the web, the linear narrative is considered as filler content by users, slowing down their ability to jump around the page; it is usually best to find the key words that they are looking for to complete their task.

Suggestion: Consider reducing the amount of text on text-heavy pages, because users generally do not read it—rather they scan for information looking to choose their next navigational path.

Review data tools: NCSES data tool users have openly said that building and retrieving data tables are not simple tasks. Additional user consultation should take place to simplify these tools. Also, consider consolidating the data tools into one data table builder or more clearly indicating the differences between them.

Suggestion: Consider conducting ethnographic interviews with users at their place of work or conduct usability testing with specific tasks to understand user's difficulties when using NCSES tools. Consider developing one data tool for meeting one-stop shopping needs.

SUGGESTED READINGS

Nielsen, J. Writing Style for Print vs. Web, Alertbox, June 9, 2008, http://www.useit.com/alertbox/print-vs-online-content.html.

Nielsen, J., and H. Loranger. (2006). *Prioritizing Web Usability.* Berkeley, CA: New Riders.

Shaikh, D., and K. Lenz. Where's the Search? Re-examining User Expectations of Web Objects. *Usability News*, February 2006, Vol. 8. http://www.surl.org/usabilitynews/81/webobjects.asp.

Appendix C
Biographical Sketches of Panel Members and Staff

Kevin Novak (*Chair*) is vice president of integrated web strategy and technology for the American Institute of Architects (AIA), where he oversees the Web, eKnowledge, and Technology departments on behalf of the institute's 86,000 members. In addition to this work, Novak is cochair of the electronic government workgroup of the World Wide Web Consortium (W3C) and former cochair of the Internet in Developing Countries Task Force at the MOBI Foundation. Prior to joining AIA, he served as director of web services at the Library of Congress, where he led the development of the its award-winning 22 million-item online multimedia collection, one of the world's largest websites. This work included launching initiatives like the World Digital Library and the Library of Congress Experience and oversight of the THOMAS legislative information service. Novak began his Internet career as the electronic government manager for Montgomery County in Maryland. He has an M.A. in technology management from the University of Maryland and a B.A. from the University of Pittsburgh.

Micah Altman is senior research scientist in the Institute for Quantitative Social Science in the Faculty of Arts and Sciences at Harvard University, archival director of the Henry A. Murray Research Archive, and nonresident senior fellow at the Brookings Institution. He conducts research in social science informatics, social science research methodology, and American politics, focusing on the intersection of information, technology, and politics; and on the dissemination, preservation, and reliability of scientific knowledge. His work has been recognized with awards from the American Political Science Association, citations by the U.S. Supreme Court, and coverage by numerous local and national media organizations. And his many publications and six open-source software packages span informatics, statistics, computer science, political science, and other social science disciplines. He holds a Ph.D. from the California Institute of Technology.

Elana Broch is assistant population research librarian in the Stokes Library for Public and International Affairs and the Ansley J. Coale Population Research Collection at Princeton University. She has done work in visualization of statistical information and presentation of

statistical inference. She provides current awareness service to faculty, students, post-doctorate students, and visiting researchers associated with Princeton's Office of Population Research. Previously she was measurement statistician at the Educational Testing Service in Princeton, New Jersey. She has a Ph.D. in psychometric methods from the University of Minnesota.

John M. Carroll is Edward Frymoyer professor of information sciences and technology at Pennsylvania State University. His research interests include methods and theory in human-computer interaction, particularly as applied to networking tools for collaborative learning and problem solving, and design of interactive information systems. He is the author of *Making Use* (2000), *HCI in the New Millennium* (2001), and *Learning in Communities* (Springer, 2009). Carroll serves on several editorial boards for journals, handbooks, and series and as editor-in-chief of the *ACM Transactions on Computer-Human Interactions*. He received the Rigo Award and the CHI Lifetime Achievement Award from the Association for Computing Machinery (ACM), the Silver Core Award from the International Federation for Information Processing, and the Goldsmith Award from IEEE. He is a fellow of ACM, IEEE, and the Human Factors and Ergonomics Society. He has a Ph.D. in psychology from Columbia University.

Patrick J. Clemins is director of the R&D Budget and Policy Program of the American Association for the Advancement of Science (AAAS). In this role, he serves as an international expert on the U.S. federal research and development investment, disseminating data and analyses through presentations, publications, and web content to a variety of audiences that include national and international policy makers, scientific associations, journalists, and the research community. Prior to joining AAAS, he was an AAAS Science and Technology Policy Fellow at the National Science Foundation in the Directorate for Biological Sciences. In the Division of Biological Infrastructure, he focused on fostering collaboration between the biological sciences and the computing and engineering research communities and the use of computing technologies for outreach and community building. Previously he was a systems engineer for Techteriors, LLC, a home automation firm, designing, programming, and managing client projects and heading a team that designed a new touch panel interface. He has B.S., M.S., and Ph.D. degrees in electrical and computer engineering from Marquette University, focusing on machine learning, digital signal processing, and bioacoustics.

Diane Fournier is a senior analyst for qualitative and quantitative research activities in the Communications Division at Statistics Canada. She specializes in qualitative research and is managing a group of facilitators. Having been involved in client consultation at Statistics Canada since 2004, she is now expert on the use of qualitative research methods that include focus groups, usability testing, and ethnographic interviewing. She graduated from Carleton University in 1990 with an M.A. During her studies she investigated the strategies of adjustment adopted by women and men in farm-based households using an ethnographic interviewing approach, in which she collected individual oral histories. Her main focus is to consult with Statistics Canada website users and test different parts of the website to heighten the user experience. Her current research involves working with interdepartmental experts on the topic of website user design and experience for the review of past and emerging federal government website designs.

Christiaan Laevaert is responsible for the management of the website of Eurostat, the statistical office of the European Union—a position he has held since he joined the Dissemination Unit of Eurostat in 2005. He coordinates the functional specifications as well as the technical implementation of the website, the associated visualization tools, and the content structure. The website was completely revamped in April 2009. He is an active member of the Dissemination Working Group in the European Statistical System, which discusses and exchanges best practices in the area of dissemination of statistical information. He has been an official of the European Commission since 1987 and was involved in various projects in the field of informatics engineering as well as in the institution's Data Centre.

Emily Ann Meyer (*Co-Study Director*) is a program officer and study director at the Computer Science and Telecommunications Board (CSTB). She was a study director for the National Materials Advisory Board and the Board on Manufacturing and Engineering Design. At CSTB she is directing a report on Depiciting Innovation in Information Technology (which updates the iconic "tiretracks" diagram) and co-directing a study on systems modernization for the Centers for Medicare and Medicaid Services. Emily holds a J.D. from Hamline University School of Law; and a B.A. (magna cum laude) in Political Science from Virginia Wesleyan College where she also minored in German.

Thomas Plewes (*Co-Study Director*) is a senior program officer for the Committee on National Statistics and was study director for earlier National Research Council studies of research and development statistics at the National Science Foundation. Previously he was associate commissioner for employment and unemployment statistics of the Bureau of Labor Statistics. He was a member of the Federal Committee on Statistical Methodology. He is a fellow of the American Statistical Association. He has a B.A. in economics from Hope College and an M.A. in economics from the George Washington University.

Andrew Reamer is research professor at the George Washington University Institute of Public Policy. He focuses on policies that promote U.S. competitiveness; his areas of interest include innovation; regional, economic, and workforce development; and economic statistics. He serves as chair of the Bureau of Labor Statistics Data User Advisory Committee; a member of the Bureau of Economic Analysis Advisory Committee; past president of the Association of Public Data Users; and a board member of the Council for Community and Economic Research. Previously, he was a fellow at the Brookings Institution's Metropolitan Policy Program and deputy director of its Urban Markets Initiative. He founded the Federal Data Project, which sought to improve the availability and accessibility of federal socioeconomic data for states, metropolitan areas, and cities. He also coauthored the policy brief that served as the basis for the Regional Innovation Program authorized by Congress in 2010. He currently is a nonresident senior fellow at Brookings. He has a Ph.D. in economic development and public policy and a M.C.P. (master of city planning) from the Massachusetts Institute of Technology.