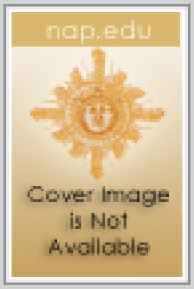

Communicating National Science Foundation Science and Engineering Information to Data Users: Letter Report

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

ISBN
978-0-309-21003-4

32 pages
8 1/2 x 11
2011

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

Committee on National Statistics
Panel on Communicating NSF Science and Engineering Information to Data Users

The Keck Center
500 Fifth Street, NW
Washington, DC 20001
Phone: 202 334 3096
Fax: 202 334 3751
www.national-academies.org/cnstat

March 9, 2011

Dr. Lynda Carlson
Director
National Center for Science and Engineering Statistics
National Science Foundation
4201 Wilson Boulevard, Suite 965
Arlington, VA 22230

Dear Dr. Carlson,

This letter report from the Panel on Communicating National Science Foundation (NSF) Science and Engineering Information to Data Users recommends action by the National Center for Science and Engineering Statistics (NCSES), formerly the Division of Science Resources Statistics (SRS), 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 panel members are listed in Appendix A.

The recommended actions in this letter report can be considered as preliminary steps for NCSES to prepare for initiatives that will foster a transition from current practices and approaches to an improved program of data dissemination. These and other issues will be discussed in the panel's final study report, which is scheduled to be issued in mid-2011. This letter report also includes a summary of the workshop that was held on October 27–28, 2010; see Appendix B. The workshop focused on the several aspects of the NCSES's current approaches to communicating and disseminating statistical information—including NCSES's information products, website, and database systems. It included presentations from NCSES staff and representatives of key user groups—including the academic research, private nonprofit research, and federal government policy-making communities; see Appendix C for the workshop agenda.

PANEL CHARGE AND ISSUES

The NCSES, as a means of fulfilling its mandate to collect and distribute information about science and engineering enterprise for the National Science Foundation, conducts an ambitious program of data dissemination in several formats: hard-copy and electronic-only publications, an extensive NCSES website, and two tools that are used to retrieve data from the NCSES database: the Integrated Science and Engineering Resource Data System (WebCASPAR) and the Scientists and Engineers Statistical Data System (SESTAT). These outputs and tools serve a broad community of information users, with wide-ranging data needs, statistical knowledge, access preferences, and technical abilities.

Our panel was asked to review the NCSES communication and dissemination program that is concerned with the collection and distribution of information on science and engineering and recommend future directions for the program. Specifically, we were asked to:

- Review NCSES's existing approaches to communicating and disseminating statistical information, including the center's information products, website, and database systems. [This review will be conducted in the context of both current "best practices" and new and emerging techniques and approaches.]
- 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.
- Consider the impact that current federal and NSF website guidance and policies have on the design and management of NCSES's online (Internet) communication and dissemination program.
- 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).
- Consider the impact of government-wide activities and initiatives (such as FedStats and Data.gov) and the emerging user capability for online retrieval of government statistics.

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 the data for print and electronic publication.[1] 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 believe 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 of the data at the source of the data, buttressed by the availability of metadata (that is, data about the data), is the necessary foundation of real improvement in data dissemination.

RECOMMENDATION 1: 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 ensure integrity of the data at the source.

All of the tables published by NCSES are selections, aggregations, and projections of the underlying microlevel observations. The recommendation above envisions that, wherever

[1] This information is based on the NSF presentation to the panel, slide numbers 14–16.

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 feasible since an extensive data appendix is necessary to support the analysis in the report. However, in general, a web release (following the practice that NCSES currently employs for the most detailed statistical tables) of the raw data will reduce the burden on NCSES's staff 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, and assuage any user concerns about the delay between data collection and its availability.

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

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 federal statistical agencywide Statistical Community of Practice and Engagement (SCOPE) initiative, which was discussed by Ron Bianchi (representing the Economic Research Service of the U.S. Department of Agriculture) at the workshop (see Appendix B). 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 NSF from contractors.

The shift to increased provision of raw data is potentially a major and significant enhancement, that 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 assured when providing users with cross-source microdata: consequently, rules about publishable cell size, for example, will have to be carefully considered.[2] The greater transparency inherent in making more raw data available also increases the risk that users could juxtapose data in ways that lead to invalid interpretations, though this danger can certainly be lessened by the accessibility of robust metadata that explain the meaning (and limitations) of the data. The current state of metadata technology permits tagging the data items with permanent uniform resource locator (URL) and uniform resource identifier (URI) codes that enable identifying the source and meaning of the data items.

[2] Several reports of the Committee on National Statistics address the need to maintain the confidentiality of data provided to government agencies in confidence: National Research Council (1979), *Privacy and Confidentiality as Factors in Survey Response* (Washington, DC: National Academy Press); National Research Council (1993), *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, Panel on Confidentiality and Data Access, G.T. Duncan, T.B. Jabine, and V.A. DeWolf, eds. (Washington, DC: National Academy Press); National Research Council (2009), *Protecting Student Records and Facilitating Education Research: A Workshop Summary*, M. Hilton, rapporteur (Washington, DC: The National Academies Press); and National Research Council (2010), *Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary*, T.J. Plewes, rapporteur (Washington, DC: The National Academies Press).

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, and NCSES should be prepared to solicit and accept error reports and make corrections as necessary. Clearly, when the general public has access and tools to combine data across data sources there will be additional questions about data accuracy and usefulness, and NCSES will need to do its best to educate users and respond to users' discoveries.

Finally, when considering data release and management, it is important to have a long-term data management plan. Yet according to staff, the NCSES's current approach to archival issues is ad hoc. In view of the importance of these data for historical reference, long-term archival access is needed, and it could be assured through proper policies and practices. At a minimum, all of the collected data and publications should be scheduled for retention by the National Archives and Records Administration. 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.

MODERNIZING DATA CAPTURE, STORAGE, AND RETRIEVAL

Emerging technologies for data capture, storage, and retrieval will dramatically change the context in which NCSES will provide data to users in the future. For NCSES, the key to taking advantage of these technologies—which are designed to increase the efficiency of data capture, storage, and retrieval and to permit users to access the data interactively (such as with Web 2.0)—is to focus on procedures for entry of the raw data into the system. It is critically important that data enter the system in as disaggregated form as possible.

Furthermore, 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 (know as provenance) and to 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.

The data also need to be able to take advantage of the web development capabilities embedded in data.gov and other emerging dissemination means to "mashup" data sources (a web page or application that uses and combines data, presentation, or functionality from two or more sources to create new services). These capabilities, incorporating such tools as open Application Programming Interface (API), enrich results and enhance the value of the data to data users. Unfortunately, NCSES is not very well positioned to take advantage of these new developments because the survey data that are entered into the center's database are received from the survey contractors in tabular format (though machine readable) rather than in an easily accessible microdata format. This is not unique to the science and engineering data. The committee heard from Suzanne Acar (representing the U.S. Department of the Interior and the Federal Data Architecture Subcommittee) that this is a governmentwide issue, one which will be taken up by a group of the World Wide Web Consortium (W3C),[3] which has plans to develop contract templates to enable governmental organizations to properly specify the format for receipt of the data from their contractors. Ron Bianchi (representing the Economic Research Service of the

[3] W3C is an international community of member organizations, a full-time staff, and the public to develop web standards: see http://www.w3c.org [November 2010].

U.S. Department of Agriculture) stated that this is also a concern for the newly formed Statistical Community of Practice and Engagement (SCOPE), and that current plans for this coordinating activity that involves most of the large federal statistical agencies include developing a template for contract deliverables specifications.

> **RECOMMENDATION 2: The National Center for Science and Engineering Statistics should incorporate provisions in contracts with data providers for the receipt of data in formats, and accompanied by metadata, that will allow efficient access for third-party visualization and integration and the use of analysis tools.**

Implementing this recommendation will be no simple task for NCSES. Currently, NCSES manages 13 major surveys, involving contracts with five private sector organizations and the U.S. Census Bureau: see Table 1. Furthermore, adding this requirement may initially incur additional costs to support a shift from the current practice of formatting the data after it is received to requiring contactors to input the data in a new format.

To enable the receipt of metadata from contractors in a universally-accessible format, NCSES will need to adopt an electronic data interchange (EDI) metadata transfer standard. The selection and adoption of a metadata transfer standard would be more effective if NCSES accomplished this through participation in a governmentwide initiative, such as the W3C contract template development, or the SCOPE effort focused on the federal statistical agencies.

UNDERSTANDING DATA USERS AND THEIR EMERGING NEEDS

NCSES is strongly committed to serving the needs of data users, but it has little evidence of how well it is meeting those needs. NCSES has made several notable attempts to gather this intelligence about user needs, but it does not have a formal, consistent, structured, and continuing program for doing so.

One problem for NCSES is that there are multiple levels of users for which products must be developed. For the most part, outreach efforts have been addressed to primary users, who are mostly researchers and analysts of research and development (R&D) expenditures and the R&D workforce. These users represent the most visible of users–researchers and analysts in the federal agencies that support government science and engineering analysis, in academia, and in the private sector. (Several of these users gave presentations at the workshop). However, the needs of a group of secondary users (those who rely on NCSES products to understand and gauge the implications for programs, policy, and advocacy) and tertiary users (such as policy makers and librarians) are given less attention, mostly because outreach to these groups is so difficult.

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 NCSES considers the program of measurement and outreach discussed in this letter. 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. Reaching out to document librarians and other secondary users by means of surveys or interviews would be another worthwhile initiative. One means of assisting in assuring that the needs of the secondary and tertiary data users are met is to assure that programs of outreach are specially directed to members of the media—those who re-release 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 a web statistics analysis program that analyzes web server log file content and displays the traffic information on the basis of the log data (URCHIN) and software that collects and presents information about user behavior on its website (WebTrends). With proper permissions and protections, NCSES is also contemplating using cookies to identify return users and increase the efficiency of filling data requests. It also has plans to sponsor and field a customer survey to formally measure satisfaction.

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 and focus groups to assess user needs and the usability of its website. The information has been used by Statistics Canada to develop a profile of its users and how they access the database. 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 those who are knowledgeable and regular users 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 NCSES 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.

> **RECOMMENDATION 3: The National Center for Science and Engineering Statistics should proceed with its plans to conduct a customer survey through use of an online questionnaire; should analyze patterns of data use by web users; and should consider reinstating the program of user workshops in order to educate users about the data and to learn about the needs of users in a structured way.**

DATA ACCESSIBILITY

The panel heard from Judy Brewer (director of the Web Accessibility Initiative [WAI] at the W3C) on the issue of accessibility of information on the web. Her presentation and the discussion that followed the presentation raised several important issues.

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 accessibility 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 who is deaf 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.

A range of guidance materials are available for developing accessible websites. Section 508 of the 1998 amendments to the U.S. Rehabilitation Act (29 U.S.C. 794d) governs the accessibility of electronic and information technology for people with disabilities, and specifies the *minimum* standards for accessibility.[4] Other standards include the "web accessibility initiative" of 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:

- *accessible rich Internet applications*—address accessibility of dynamic web content, such as those developed with Ajax, Dynamic HTML, or other such technologies;
- *authoring tool accessibility guidelines*—address the accessibility of the tools used to create websites;
- *user agent accessibility guidelines*—address assistive technology for web browsers and media players; and
- *web content accessibility guidelines*—address the information in a website, including text, images, forms, and sounds.

In addition to issues of the usability and navigability of websites and web applications, there are issues related to the use and navigation of the datasets. Tabular data formats can be difficult to understand for those who must use screen readers, and data that are not organized in a transparent or immediately understandable way may be of limited or no utility for users with cognitive disabilities.

Through this presentation and the panel's subsequent discussion, it became clear that the issue of the accessibility of tabular data and data visualization is a research question. Although W3C has pioneered standards for accessibility of dynamic user interfaces, many other issues including table navigation, navigation of large numeric datasets, 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.

RECOMMENDATION 4: The National Science Foundation should sponsor research and development on accessible visualization and other means for exploring tabular data.

The panel recognizes that such a far-reaching initiative is almost certainly beyond the capability and resources of NCSES, and so our recommendation is to the National Science Foundation. A program of research and development might fit well into the portfolio of other NSF units and could be considered for funding under such programs as the Broadening Participation Research Initiation Grants in Engineering Program. The importance of promoting scientific visualization as an aid to usability and accessibility is a recognized component of NSF's cyberinfrastructure vision[5] and is inherent in its human-centered computing cluster initiatives. Although NSF is not in a position to pioneer tool development, identifying the appropriate research areas is something that will have a major impact in the field.

NCSES could share the knowledge gained through these research and development activities with the broader community of practice and so have a major impact on a wide range of

[4] A summary of Section 508 is available online at http://www.section508.gov/index.cfm?fuseAction=stdsSum [November 2010].

[5] National Science Foundation, *Cyberinfrastructure Vision for 21st Century Discovery*, March 2007, p. 28.

potential users of NCSES's data (as well as other statistical datasets), thus making the data available to potential users for whom they are now inaccessible.

As a final note, the panel wishes to commend NCSES for encouraging this review of its dissemination practices. This is a particularly opportune time for incorporating lessons learned in several U.S. and international initiatives that are designed to increase the transparency, usability and accessibility of government data. Implementing the recommendations in this interim report will go a long way toward laying the basis for significant improvements in the way center data are disseminated.

Sincerely yours,
Kevin Novak, *Chair*
Panel on Communicating NSF Science and
Engineering Information to Data Users

Attachments:
Appendix A—Panel Members
Appendix B—Workshop Summary
Appendix C—Workshop Agenda

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

Survey	**Current Contractor**	**Database Retrieval Tool/ Publication**	**Availability of Microdata**	**Series Initiated/Archiving**
		Education of Scientists and Engineers		
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	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	Data for the years 1972–2008 are available in a public-use file format	1975 (conducted annually)
		Science and Engineering Workforce		
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*	Access to restricted data for researchers interested in analyzing microdata can be arranged through a licensing agreement	1973 (conducted biennially)
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*	Access to restricted data for researchers interested in analyzing microdata can be arranged through a licensing agreement	1976 (conducted biennially)
National Survey of College Graduates	Census Bureau	SESTAT; InfoBriefs; *Science and Engineering Indicators*; *Women, Minorities, and Persons With Disabilities in Science and Engineering*	Public use data files are available upon request	1962 (conducted biennially)

Research and Development Funding and Expenditures				
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*	Census Research Data Centers	1953 (conducted annually)
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*	Data tables only	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 tables only	1965 (conducted annually)
Survey of R&D Expenditures at Federally Funded R&D Centers	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*	Data tables only	1965 (conducted annually)
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	Data tables (selected items) only	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*	Data tables only	1964 (conducted occasionally)

Science and Engineering Research Facilities				
Survey of Science and Engineering Research Facilities	RTI International	WebCASPAR; *Scientific and Engineering Research Facilities*; *Science and Engineering Indicators*	Microdata from this survey for the years 1988–2001 are not available	1986 (conducted biennially)
Other Surveys				
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*	Data tables only	ICPSR, 1979–2001; CD, 1979–2004; (conducted biennially)

Appendix A

PANEL ON COMMUNICATING NATIONAL SCIENCE FOUNDATION SCIENCE AND ENGINEERING INFORMATION TO DATA USERS

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

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

Appendix B
Workshop Summary

INTRODUCTION

The National Center for Science and Engineering Statistics (NCSES), formerly the Division of Science Resources Statistics (SRS), of the National Science Foundation (NSF), in seeking to fulfill its mandate for collecting and disseminating information about science and engineering, conducts an ambitious program of data dissemination. The program includes a variety of release formats, including numerous hard-copy and electronic-only publications. The extensive NCSES website also houses the Integrated Science and Engineering Resource Data System (WebCASPAR) and the Scientists and Engineers Statistical Data System (SESTAT), which are tools used to retrieve data from NCSES databases. These formats and tools serve a broad community of information users, with differing data needs, statistical knowledge, access preferences, and technical abilities.

As part of a program of periodic review of all of its activities, NCSES requested that the Committee on National Statistics of the National Research Council appoint an expert group to conduct a study of its dissemination program. As part of its work, the Panel on Communicating National Science Foundation Science and Engineering Information to Data Users organized a workshop, which was held in Washington, DC, on October 27–28, 2010. The purpose of the workshop was to bring together NCSES leadership and staff, data users, and representatives of the data storage, retrieval, and dissemination community to gather information for several of the panel's study tasks:

- document current (traditional) practices of NCSES for communicating and disseminating information in hard copy publication format as well as on the Internet through the NCSES website, and the WebCASPAR and SESTAT database retrieval systems;
- consider the needs of significant data users;
- evaluate the website design from the perspective of usability, including consideration of navigation aids, user-centered design techniques, and interactivity;
- consider the impact on the ability of NCSES to improve communication and dissemination, given NSF website policy and governmentwide policies;
- evaluate the impact of current governmentwide initiatives, such as data.gov, for retrieval of government statistics; and
- consider new and emerging tools for retrieval and display of information, including using the semantic web for linking with other databases.

The first section of this summary covers U.S. agency initiatives on data dissemination. The second section looks at international initiatives. The final section considers accessibility issues. See Appendix B for the workshop agenda.

CURRENT STATUS OF THE DISSEMINATION PROGRAM

With regard to the dissemination of its varied and important data, NCSES has been in a change mode for some time. NCSES has taken steps to move from, in the words of center director Lynda Carlson, “spewing out reams of data tables to providing our customers information.” For example, few of the data series and analytical products are currently released as printed matter, in sharp contrast to a decade ago, when paper was the primary mode of dissemination.

Although progress has been made, the leadership of NCSES is concerned that much more needs to be done to improve the dissemination of data, and for this reason, Carlson stated, NCSES requested this panel’s study to carefully review NCSES’s approach to communicating and disseminating statistical information in the context of current “best practices” and new and emerging technologies. This review was asked to consider the needs of different types of users in NCSES’s user community, current federal and NSF website guidance and policies, existing staff and contract resources, and broader governmentwide activities and initiatives. In her introductory remarks to the workshop, Carlson requested that the panel consider what NCSES might do differently in terms of publication types, online report formats, data access (online data tools and data files), documentation of survey methods, data quality, accessibility, and outreach and notifications.

This is a tall order, particularly when considering the wide-ranging mandate of this rather small statistical agency. NCSES’s permanent staff of about 45 employees and its contractors have responsibility to collect data and maintain databases on research and development (R&D), science and engineering (S&E) education, the S&E workforce, and related areas; designing and conducting major surveys (currently, 11 of them, several dating back to the early 1950s); collecting and analyzing science and technology-relevant data from other agencies and organizations (both domestic and international); providing a global context for U.S. data; enabling comparisons and benchmarking through collaboration with the Organisation for Economic Co-operation and Development (OECD) and other international and national statistical agencies; producing the biennial *Science and Engineering Indicators* products (under the guidance of the National Science Board) and congressionally mandated reports (such as *Women, Minorities, and Persons with Disabilities in Science and Engineering*); and, finally, conducting a program of dissemination of data and analyses to a wide range of data users.

Much of the internal responsibility for the dissemination program of NCSES falls to John Gawalt, deputy director and acting program director of its Information and Technology Service Program. In his presentation to the workshop, Gawalt summarized the current status of the dissemination program and discussed NCSES’s various products and publications. Two of the products appear in both print and electronic formats: *InfoBrief*, which is published when needed to highlight results from recent surveys and analyses, and special analytical reports, such as the reports on women, minorities and persons with disabilities in science and engineering and *Science and Engineering Indicators*. Other products are in electronic form only. They include a series of extensive tables and technical material for each of the 11 periodic surveys, as well as special reports (*National Patterns of R&D Resources,* Academic Institutional Profiles*,* and *S&E State Profiles*), working papers, and methodology reports. Several other specialized products, such as brief summaries (Infocards) and CDs, are also prepared.

Underlying all the reports and publications is a data repository, maintained by NCSES, which houses the results of the contractor-collected survey and administrative data and several online databases, which are available to both NCSES staff and the public. There are four major databases, three of which have existed for some time:

Scientists and Engineers Statistical Data System (SESTAT) This database tool captures 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). The data can be downloaded from the web or through the SESTAT Data Tool, which allows users to generate custom data tables.

Integrated Science and Engineering Resources Data System (WebCASPAR) This database system contains information about academic S&E resources and is newly available on the web. Included in the database is information from several of NCSES's academic surveys, as well as information from a variety of other sources, including the National Center for Education Statistics. The system is designed to retrieve multiyear information about individual fields of S&E at individual academic institutions. The system provides a user with opportunities to select variables of interest and to specify whether and how information should be aggregated. Output can be requested in hard-copy form or in Lotus, Excel, or SAS (Statistical Analysis System) formats for additional manipulation by the researcher.

Industrial Research and Development Information System (IRIS) This system 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. These tables are 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. IRIS resembles a databank more than a traditional database system: rather than firm-specific microdata, it is the most comprehensive collection of historical national industrial R&D statistics currently available. The tables in the database are in Excel spreadsheet format, which are accessible either by defining various measures (e.g., total R&D) and dimensions (e.g., size of company) of specific research topics or by querying the report in which the tables were first published.

The most recent addition to NCSES's dissemination offerings are public-use files and restricted (licensed) data-sets:

Public-Use Microdata Files NCSES takes pains to protect the data that are provided by individuals from disclosure. Thus, NCSES develops microdata files for release to the public in a format that will not permit identification of respondents. In some cases, respondents can be protected by ensuring that the files contain records from which identifying information (such as name and address) has been deleted from the records. In most cases, however, it is necessary to suppress selected fields or recode variables in order to provide researchers with as much microdata as feasible.

Licensed Datasets In some cases, the protection of respondent confidentiality would require such extensive recoding that the resulting file would have little, if any, research utility. In these cases, NCSES does not issue a public-use file. Instead, NCSES has developed a licensing procedure that permits a researcher to use the data files at NSF's offices in Arlington, Virginia, or at the researcher's academic institution under carefully controlled circumstances (for details, see http://www.nsf.gov/statistics/database.cfm [December 2010]).

REQUIREMENTS OF DATA USERS

At the workshop, the panel heard from key data users from the federal government, the agencies that support federal government S&E analysis, academia, and the private sector. These presenters were asked to address, from their perspective, the current practices of NSF for communicating and disseminating information in hard-copy publication format as well as on the Internet through the NCSES website, and the WebCASPAR and SESTAT database retrieval systems. The panel had asked these data users to comment on the means used to obtain information on NSF S&E expenditure and human resource data; the comprehensiveness, quality, timeliness, or other aspects of the NSF publications and data release program; use of the website; use of the WebCASPAR and SESTAT databases; and the content, presentation, accessibility, and tools available for these databases.

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 NSF 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.[6]

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.,

[6] National Research Council. (2009). *Federal Spending Data for Research and Development: 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.

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 and 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 [NAICS] 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—including datasets underlying special tabulations related to publications, patents, and other complex data—would better enable assessment of emerging fields. 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, that 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 (see above), and about 40 to 50 times a year, she uses restricted files pertaining to SDR, SED and SESTAT. She also uses *InfoBriefs* and the *Science and Engineering Indicators* appendix tables, and she accesses data through WebCASPAR. Graduate students use WebCASPAR to build tables and such create 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 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. There were data workshops for 3 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 of 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 recessions on salaries, 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 the recession—for July 1, 2008–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 race and ethnic data). So it is "early" recession data, though 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–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."[7]

Private Sector 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 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 geographic metadata 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 over 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 finds that 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.

FEDERAL DISSEMINATION INITIATIVES

Over the years, several significant programs have been initiated to increase the availability of federal government information in electronic format. For the NCSES, as part of

[7] Recommendation 5 in National Research Council. (2000). *Measuring the Science and Engineering Enterprise: Priorities for the Division of Science Resource Studies.* Committee to Assess the Portfolio of the Division of Science Resources Studies of NSF, Office of Scientific and Engineering Personnel and Committee on National Statistics. Washington, DC: National Academy Press.

the statistical community, the FedStats gateway (www.fedstats.gov) has, for several years, incorporated pointers to the S&E data in both the list of topics (education, R&D, and workforce) and the agency listings portion of the FedStats home page. The pointers in FedStats guide the user to the NCSES home page at which the subject matter appears.

Data.gov

NCSES was an early member of the federal government's open government initiative, Data.gov, when it was initiated in May 2009. Workshop presenter Alan Vander Mallie**,** the 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 and 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 the data, including databases, data feeds, graphics, and other data visualizations.

Vander Mallie reported that, at its inception in 2009, Data.gov consisted of 47 raw datasets and 27 tools to assist in accessing the data in some of the complex data stores. Currently (at the time of the workshop), the program supports 2,895 raw datasets and 638 tools, which are accessed through raw data and tool catalogues. Raw data are defined as machine-readable data at the lowest level of aggregation in structured datasets 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.

In the future, Vander Mallie said, Data.gov is slated to continue to expand its coverage of datasets and tools and to continue to support communities of interest by building community pages that collect related datasets and other information to help users find data on a single topic in one location. One objective is to make data available through the application programming interface (API), permitting the public and developers to directly source their data from Data.gov. Expansion into the Semantic Web (sometimes called Web 3.0) 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 enscription framework documents that are available and is working with the World Wide Web Consortium (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" (to 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 the data.

Suzanne Acar, senior information architect for the U.S. Department of the Interior and co-chair 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 by discussing an agency perspective on the lessons learned from this program to improve access to the federal government's data. 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.

Statistical Community of Practice and Engagement

The federal statistical agencies, as a group, have begun to organize to enhance dissemination of their data in a project called the Statistical Community of Practice and Engagement (SCOPE). As described by Ron Bianchi of the Economic Research Service of the U.S. Department of Agriculture, who heads the planning committee, SCOPE represents a recognition on the part of the leadership of statistical agencies that 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.

SCOPE envisions that the dissemination platform for federal statistical data will be a modernized FEDSTATS—which has name recognition but is technologically outdated. It will use the Data.gov statistical community as SCOPE's public interface and dissemination platform, store datasets on the Data.gov dataset hosting platform that is currently being developed, and harness Data.gov cloud computing power.

In recognition of the fact that statistical data collected by federal statistical agencies often raise issues of complexity as well as of confidentiality and privacy, Bianchi said that SCOPE will aim to develop user-friendly data delivery and data display tools to address 508 compliant alternatives to tabular displays, develop displays of complex sample survey data while protecting confidential micro-data, and develop visualization tools for multifaceted statistical designs.

The statistical agencies that are part of SCOPE, which will include NCSES, will participate in promoting data harmonization and integration through the development of metadata and data exchange. Specifically, the project 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.

American FactFinder

NCSES's three major dissemination tools—SESTAT, WebCASPAR, and IRIS—have been in place without major modification for some time, and some workshop participants commented that it needs a retooling. In thinking about an approach to retooling, one approach would be to consider what other government agencies have done or are doing to improve their dissemination tools. One such tool is the Census Bureau's primary web-based data dissemination vehicle, the American FactFinder. 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 now in the process of being redesigned with many goals: increase the effectiveness of user data access; guide users to their data without forcing them to become experts; improve turnaround time; increase the efficiency and flexibility of dissemination operations; address growing usage and data volume needs; and 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. This information is presented in Table B-1.

Sisson reported that the new American FactFinder system is scheduled to launch in January 2011. In the meantime, the Census Bureau has provided email updates on the status of the project, developed a virtual tour of the new system, and, on a flow basis, will be issuing tutorials on the new system.

DataWeb

In his introduction to the discussion of the Census Bureau's DataWeb network, Cavan Capps, the 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 the DataWeb network was one of three parts to the Census Bureau's approach to these challenges The other two are *HotReports* and DataFerrett.

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 nonprofits institutions. The software provides an open source, 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 statistically 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 one time, but using it many times. DataFerrett and *HotReports* both use the DataWeb framework.

HotReports are 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.

DataFerrett is a data web browser that is targeted at sophisticated data users and integrates multiple datasets. It speeds analytical tasks by allowing data manipulation and incorporating advanced tabulation and descriptive statistics. Its mapping and business graphics use statistical rules. It has the capability of adding regressions and other advanced statistics.

Summary of Data Dissemination Practices

Panel member Micah Altman began his summary by noting that this is an exciting, fast-changing time for electronic data dissemination in the public sector. He summarized the state of current practice in terms of publicly available systems for online numeric data sharing, publishing, and visualization. The systems can be further classified as open and closed source (summarized in Table B-2).

DataVerse Network is an open-source and standards-based virtual data archive tool. It handles moderately sized data and supports long term access and preservation and supports Data Documentation Initiative (DDI) metadata. It is a leading example of standard-based open systems.

ProtoViz is a toolkit for dynamic visualization of complex data. This open source tool is in JavaScript and handles small-sized databases. It supports a partial grammar of graphics in high-level abstractions. It is a leading example of dynamic data visualization.

Factual is a data manipulation tool used in the commercial sector. It is closed source and handles moderately 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 API's for extraction and interrogation of datasets. It also integrates with Google charts and maps. It is a very interesting example of collaborative data editing.

Tableau is an extraction tool that produces linked dynamic tables and graphics using raw data as the input. It handles moderately sized data, and it supports downloads in CSV and PDF formats, as well as HTML tables.

Google has a number of offerings, including **Google Sheets**, which is an Excel-type tool that has APIs for integration and handles small datasets; **Fusion Tables**, which focuses on data sharing, linking, and merging; and **Google Public Data Explorer**, which searches across data elements and has some visualization capability.

Altman observed that dissemination is a dynamic field in the private sector, and that many of the start-up dissemination and data-sharing services have closed. In view of this uncertainty, he said users would be well advised to mitigate risk of adopting any of these systems by using open source software when possible, to retain preservation copies of files in other institutions, to limit use to dissemination only (not for managing data), and to leverage metadata and APIs to create one data source that is then disseminated through multiple sources.

Altman identified research challenges and gaps between the state of the art and the state of the practice. Research challenges in this area include petascale 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.

INTERNATIONAL INITIATIVES

The computer revolution and the development of the web have generated a large number of initiatives to improve statistical data dissemination not only in the United States, but also internationally. The workshop included presentations from panel members Christiaan Laevaert and Diane Fournier on the dissemination initiatives of their institutions, the European Community and Statistics Canada.

EUROSTAT: Focus on Data Usability

The Statistical Office of the European Union (EUROSTAT) compiles statistical data that are, for the most part, collected by member states. Christiaan Laevaert said that EUROSTAT adds value by providing statistics at the European level that enable comparisons between countries and regions and by disseminating these data, free of charge, in consolidated format through publications and online databases. Disseminating this information is a large undertaking. In January 2010, for example, the EUROSTAT website had 2.8 million visitors and 3 million page views. There were 400,000 dataset extractions, 400,000 table consultations, and more than 400,000 publication downloads in PDF format. The site is among the top five visited websites of the European Commission.

In the past, hard-copy publications constituted a very large, cumbersome workload for EUROSTAT in a process that has meant that data at the moment of publishing are already 6–9 months older than what can be found on the website. The agency has published annually 5–15 printed statistical books, 15 pocket books, and 120 “statistics in focus” and various methodological publications. Large amounts of content have been created and are disseminated in paper and PDF format. Similar content is often created repetitively in different forms for different purposes. In order to improve the situation, Laevaert said, the Eurostat Board of Directors asked to examine the possibilities to disseminate the content of publications more effectively. The subsequent discussions led to the creation of a wiki-type system for the dissemination of statistical articles and related textual content. This new system does not only change the manner of dissemination, but also the way of collaboration in the preparation of publications.

The wiki approach was selected because EUROSTAT databases have a wide variety of users: one group is a particular challenge to serve—the nonexpert data user. The agency has taken on the task of making the data more useful to this group by developing a dissemination aid called Statistics Explained. Information in Statistics Explained is disseminated using a wiki technology with electronic publication. The wiki explains what the statistics really mean, what is behind the figures, and how they can be of use, in an easily understandable language. Numerous hyperlinks allow for easy navigation.

More than 250 articles are contained in this system, with some 800 glossary items, and the wiki-type website assures a high ranking in search engines. In September 2010, 60,000 unique visitors per month were counted. EUROSTAT has realized a synergy between its main website and Statistics Explained by mutual deep links that assure visitors a coherent navigation between the two websites.

Laevaert said that this dissemination strategy has changed the way EUROSTAT does business. It has introduced a new production process for publications and a paradigm shift in the way users are served—by focusing first on Statistics Explained first and later on publication.

However, unlike Wikipedia, information can only be updated by EUROSTAT staff, thus ensuring the authenticity and reliability of the content, so updating the system involves a large part of the staff: more than 200 of the 800-plus EUROSTAT staff contribute to content updates.

While Statistics Explained is addressed to nonexpert users, the retrieval system of EUROSTAT focuses on all types of users, including more sophisticated ones. It allows users to reuse data with tools they prefer. Thus, users are provided with various data formats and visualization tools, including xls, html, txt, xml, spss, and pc-axis. Each of the more than 5,000 datasets disseminated on the website are also available in several raw formats in the bulk download facility—tsv, sdmx-ml, and dft. A table of contents file in XML format facilitates machine-to-machine interactions.

Laevaert noted that EUROSTAT is also rethinking its approach to visualization tools, adapting procedures to take advantage of cloud computing and being able to supply data in formats required by emerging tools. Working with Google, the data featured on Google's Public Data Explorer are being integrated into Google search with Onebox. The Google search integration makes datasets searchable in 34 languages and assures the highest ranking in search results. Currently, four EUROSTAT datasets have been integrated, which has significantly improved the overall visibility of its data.

Statistics Canada User Outreach

Although Statistics Canada uses many of the same dissemination practices and tools as are used by NCSES and other U.S. statistical agencies, Diane Fournier said it has contributed significantly to the enhancement of the dissemination experience by its sharp focus on data users and usability. The program of usability assessment includes a website evaluation survey and ongoing consultations with users through focus group discussions and usability testing in a lab environment. The agency has recently been a partner in testing new governmentwide design standards. The lessons learned as a result of these activities have resulted in a home page redesign, search improvements, and the development of a "statistics by variable" tool to permit direct retrieval of certain variables (not yet available).

Website evaluation studies were conducted from 1997 until 2007, usually on an annual basis. In 2010, an online questionnaire that reached all website visitors was live for 15 days and, with almost 10,000 respondents, the response rate (responses as a percent of visitors) was 3 percent. The representation of the online respondents was similar to that of earlier surveys: 27 percent were students; 21 percent were from the government and public sector; and 20 percent represented the business and private sector. About 15 percent of users accessed the Internet through mobile devices, but more traditional computers were used more often: desktops PC by 72 percent and laptops or notebooks by 61 percent. In an interesting comparison, in a visitor pattern analysis during a 6-month period from April to September 2009, 0.4 percent of all website visitors used mobile devices.

The 2010 survey focused on the subject of task completion. Some 65 percent of respondents accomplished what they set out to accomplish, and, not surprisingly, satisfaction also registered at 65 percent. When asked for three priority suggestions for improvement of the website experience, the top suggestions were to "make it easier to access to data/information," "improve the search engine/search results," "offer additional free data/information," "simplify the site layout/design," and "use clearer/plain language."

Usability testing is also used by Statistics Canada, Fournier reported. It involves observing "real users" complete specific tasks to see if the experience meets user needs, functions as expected, and is intuitive for its intended audience.

As a result of these user outreach initiatives, Statistics Canada has taken steps to provide better access to the latest content and paid more attention to serving dissemination through mobile devices. In addition, the agency is continuing to improve search technology and has begun to archive content. Fournier said that Statistics Canada sees real benefit in continued consultations with users, continued collaborative efforts with national and international agencies and departments, and continued participation in Canada's governmentwide design standards that focus on brand recognition.

ACCESSIBILITY ISSUES

In order for NCSES science and engineering 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, NCSES 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 also includes people who are disabled and for whom, by law and right, special accommodations need to be made.

The panel benefitted from a presentation by Judy Brewer, who directs the Web Accessibility Initiative (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.

As a federal government agency, NSF is governed by the so-called 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 WAI has responsibility. Brewer also quoted Statistical Policy Directive Number 4 (March 2008), which directs statistical agencies to make information available to all in forms that are readily accessible.

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 you can submit them, images that flash and may cause seizures, text that moves or refreshes before you 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 National Institute of Standards and Technology workshop on complex tabular information that resulted in several recommendations.

Brewer argued for publishing existing science and engineering data in compliance with Section 508 requirements, while continuing research and development on accessibility techniques for new technologies, improved accessibility supports for cognitive disabilities, and more affordable assistive technologies. She said WAI would partner with agencies to ensure that dissemination tools are accessible.

TABLE B-1 Areas for Improvement for American FactFinder Identified from Usability Studies

Area for Improvement	Goal	Measure
Visual Elements	Homepage should be more visual to improve visitor expectations, reduce perceived complexity, and improve the look and feel of the site	Easier to absorb and assess through better balance of text and visuals Use of images, color and negative space can help convey what to expect and make the page easier to digest
Conventional Layout	Search should be presented in the best practice format as a text box directly on the homepage	Prevents visitors from going through a multi-step process to perform a query Search is less likely to be overlooked or to blend in with other links/options surrounding it
Consistent Structure	Navigation options should appear in the same location throughout the site	Visitors don’t have to search for them or use their browser’s back button
Layering Information	Should improve content management and reduce scrolling by more effectively layering page information	Less scrolling gives visitors the impression that the information is much easier to absorb Layering ensures more information is presented higher up on the page—and avoids overwhelming visitors with too much content at once

SOURCE: Data from presentation by Jeffrey Sisson, U.S. Census Bureau, at the Workshop on Communicating National Science Foundation Science and Engineering Information to Data Users, October 28, 2010.

TABLE B-2 Publicly Available Systems for Online Numeric Data Sharing, Publishing, and Visualization

Source	Data Sharing	Publishing	Visualization Only
Open	Dataverse Network		Prefuse Flare Processing ProtoVis
Closed	Data 360 Factual Google fusion tables Google sheets Many Eyes Swivel Statplot	Beyond 20/20 Collectica Nesstar SDA Tableau Track-n-Graph	Google Vis AP VisiFire

SOURCE: Micah Altman, *Data Dissemination: State of the Private Sector Practice*, Workshop on Communicating National Science Foundation Science and Engineering Data to Users, October 28, 2010. Permission granted by author.

Appendix C
Workshop Agenda

Goals for the Workshop:

1. Obtain information from NSF management on the current status of dissemination of science and engineering data, and their plans for the future
2. Obtain information from data users on their requirements
3. Hear from experts in visualization, next generation tools, and accessibility about the current state of the science in data dissemination

Wednesday, October 27, 2010
Room 204, Keck Center, 500 Fifth Street NW, Washington, DC
Open Session

9:00–9:15 AM	**Welcome, Discussion of Agenda for the Workshop** (continental breakfast served) Kevin Novak, Chair
9:15–10:15	**Current Status and Plans for Dissemination of S&E Information** Lynda Carlson, Director, Division of Science Resources Statistics National Science Foundation John Gawalt, Program Director, Information and Technology Services Program, Division of Science Resources Statistics, National Science Foundation
10:15–10:30	**Break**
10:30AM–12:00 PM	**Panel Discussion: User Evaluation of NSF S&E Information Dissemination** Moderator: Patrick Clemins, Panel Member Paula Stephan, Georgia State University Jeff Alexander, SRI International Kei Koizumi, OSTP Bhavya Lal, IDA STPI
12:00–1:00	**Working Lunch (Atrium)** **(**Continued discussion of dissemination issues)
1:00–2:00	**U.S. Government Dissemination Initiatives** Moderator: Elana Broch, Panel Member

Open Government Initiative (Data.Gov)
Alan Vander Mallie, Program Manager, Data.Gov, Program Management Office, Office of Innovative Technologies, Office of Citizen Services and Innovative Technologies, General Services Administration

2:00–3:00 **Statistical Agency Initiatives**
Moderator: Andrew Reamer, Panel Member

Community of Practice Initiative
Ron Bianchi, Economic Research Service

American Fact Finder
Jeffrey Sisson, American Fact Finder, Census Bureau

3:00–3:15 **Break**

3:15–4:30 **International Initiatives**
Moderator: Diane Fournier, Panel Member

Eurostat—Data Usability: Public Data Explorer and Statistics Explained
Christiaan Laevaert, Panel Member

Statistics Canada—Web Redesign Initiatives
Diane Fournier, Panel Member

4:30–5:00 **Private Sector Initiatives**
Moderator: Micah Altman, Panel Member

5:00 PM **Panel in Recess**

6:00–7:30 **Working Dinner (by invitation)**
(Panel will discuss Day 1 presentations and prepare for Day 2)

Thursday, October 28, 2010
Room 110, Keck Center, 500 Fifth Street NW, Washington, DC
Open Session

8:30–8:45 AM **Review of First Day Workshop and Discussion of Agenda**
(continental breakfast served)
Kevin Novak, Chair

8:45–10:00 **W3C Initiatives, Data.gov, and Web 2.0**

Suzanne Acar, U.S. Department of the Interior and Federal Data Architecture Subcommittee

10:00–10:15 **Break**

10:15–11:00 **Accessibility of NSF Internet Data (508 Compliance)**
Judy Brewer, Web Accessibility Initiative, World Wide Web Consortium

11:00–12:00 PM **Open Discussion**
Kevin Novak, Chair

12:00–1:00 **Working Lunch**
(Continued discussion of accessibility issues)

Thursday, October 28, 2010
Room 110, Keck Center, 500 Fifth Street NW, Washington, DC
Closed Session

1:00–2:30 **Committee Discussion of Workshop Lessons Learned**

2:30–3:30 **Plans for Interim Report**

3:30 PM **Adjourn**