Data on Federal Research and Development Investments: A Pathway to Modernization

Panel on Modernizing the Infrastructure of the National Science Foundation Federal Funds Survey; National Research Council

ISBN: 0-309-14524-4, 120 pages, 6 x 9, (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

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.

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 SES-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.

International Standard Book Number 13: 978-0-309-14523-7
International Standard Book Number 10: 0-309-14523-6

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.

Printed in the United States of America

Suggested citation: National Research Council. (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.

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 MODERNIZING THE INFRASTRUCTURE OF THE NATIONAL SCIENCE FOUNDATION FEDERAL FUNDS SURVEY

CHRISTOPHER T. HILL (*Chair*), School of Public Policy, George Mason University
WILLIAM B. BONVILLIAN, Washington, DC, Office, Massachusetts Institute of Technology
KATY BÖRNER, School of Library and Information Science, Indiana University
MARY K. FEENEY, College of Planning and Public Affairs, University of Illinois at Chicago
DAVID GOLDSTON, Center for the Environment, Harvard University
NANCY J. KIRKENDALL, Independent Consultant, Washington, DC
JULIE THOMPSON KLEIN, Department of English, Wayne State University
KEI KOIZUMI, Office of Science and Technology Policy, Executive Office of the President
J. DAVID ROESSNER, SRI International, Prescott, AZ
MARTHA M. TAYLOR, Office of Sponsored Programs, Auburn University

THOMAS J. PLEWES, *Study Director*
MICHAEL J. SIRI, *Program Associate*

COMMITTEE ON NATIONAL STATISTICS
2009

Acknowledgments

Many people contributed time and expertise to the information-gathering efforts of the panel, which, together with its own deliberations, form the basis for this report. The panel appreciates their cooperation and assistance.

The staff of the Division of Science Resources Statistics of the National Science Foundation (NSF) was exceptionally cooperative and forthcoming with information necessary for the panel to conduct its business. Under the leadership of Lynda Carlson, who addressed the panel at its first meeting and helped to establish the framework for our inquiries, and her deputy, Mary Frase, the division staff went to great lengths to assemble information and present it to the panel in a concise and useful manner. John Jankowski, who manages these survey operations, gave three informative orientation presentations to the panel in the first meeting and the workshop the panel held September 5-6, 2008. These presentations summarized the issues with the surveys and the taxonomy of fields of science and engineering, permitting a frank and fruitful discussion of those issues. From her perspective as a staff mathematical statistician, Jeri Mulrow provided a helpful discussion of her work in assessing the implementation of the fields of science and engineering and the findings of her investigation into the use of the fields in the work of NSF. Melissa Pollack coordinated between the panel and NSF in her service as the person responsible for the grant that supported this activity.

The presentations in the workshop provided much of the basis for the analysis and recommendations in this report. The panel expresses its appreciation to Diane DiEuliis of the Office of Science and Technology Policy;

James Wilson, then majority staff director of the Research and Science Education Subcommittee of the U.S. House of Representatives' Committee on Science and Technology; and John Sargent of the Congressional Research Service for representing data users in a session that expanded awareness of important uses of the data from the surveys and offered some very solid suggestions for making the data more relevant and useful. Robert Gropp of the Institute of Biological Sciences provided his perspective on the use of these data and suggested improvements on the spur of the moment when a scheduled speaker was unable to participate. Although not able to participate in the workshop, Louis Lanzerotti, chairman of the National Science Board's Subcommittee on Science and Engineering Indicators, provided letter input to the panel that outlined conclusions of the National Science Board on data resources for federal research and development (R&D) allocation decisions.

Two individuals represented the data providers in the federal agencies with responsibility for conducting research and development programs: Israel Lederhendler, a director in the Office of Extramural Research at the National Institutes of Health (NIH), and Tom Russell, director of aerospace, chemistry, and materials in the Air Force Office of Scientific Research. Subsequent to the workshop, Timothy Hayes of NIH provided additional information on its Research, Condition, and Disease Categorization System. Julia Lane of NSF provided a valuable discussion of the important Science of Science and Innovation Policy initiative and, on the basis of her own extensive work in using administrative data for research and analysis, opened the way to a subsequent discussion of opportunities to capitalize on government-wide initiatives to use administrative data for assessing R&D investments.

To assist the panel in focusing on long-term opportunities to use federal government administrative data and other data sources for measuring R&D investments, Mark Bussow represented the agency with responsibility for overall coordination of these efforts, the Office of Management and Budget. Andrew Reamer of the Brookings Institution provided a discussion and assessment of the various sources of information that represent the federal government's inventory of administrative data on grants and contracts, and Jeff Alexander, then of New Economy Strategies (now associated with SRI International), summarized the utility of those data sources from the prospective of a user of the information. On the second day of the workshop, Gretchen Gano of New York University discussed the state of the science with regard to classification of the fields of science and engineering and summarized several new initiatives that use the power of computing to classify items and domains.

The panel also acknowledges the excellent work of the staff of the Committee on National Statistics (CNSTAT) and the National Research

Council (NRC) for support in developing and organizing the workshop and this report. Under the direction of Constance Citro, director of CNSTAT, Tom Plewes, the study director, provided valuable assistance to the panel in organizing the meetings and preparing this report. He was ably assisted by Michael Siri, also on the staff of CNSTAT.

This report has 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 NRC. The purpose of this independent review is to provide candid and critical comments that assist the institution in making its report as sound as possible, and to ensure that the report meets 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 this report: Rita R. Colwell, Center for Bioinformatics and Computational Biology, University of Maryland; Richard B. Freeman, Department of Economics, National Bureau of Economic Research, Harvard University; Monica Gaughan, Department of Health Policy and Management, University of Georgia; James Hendler, Tetherless World Constellation, Rensselaer Polytechnic Institute; Julia Melkers, School of Public Policy, Georgia Institute of Technology; Roger Pielke, Jr., Center for Science and Technology Policy Research, University of Colorado; and Hal S. Stern, Department of Statistics, University of California, Irvine.

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 this report was overseen by Lawrence D. Brown, Department of Statistics, The Wharton School, University of Pennsylvania. Appointed by the NRC, 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 NRC.

Christopher Hill, *Chair*
Panel on Modernizing the Infrastructure
of the National Science Foundation
Federal Funds Survey

Contents

Summary		1
1	Introduction	7
2	The Current Surveys: Challenges and Opportunities	13
3	Improving the Current System in the Short Term	31
4	Toward a New Data Reporting System Based on Administrative Records	48
5	Toward a Comprehensive Assessment of Federal Research and Development Investment	61
6	The Path Ahead	69
References		72

Appendixes

A	Acronyms and Abbreviations	77
B	Guide to Federal Grants and Contracts Databases	80
C	Modernizing the Infrastructure of the National Science Foundation Federal Funds Survey: Summary of a Workshop	84
D	Biographical Sketches of Panel Members and Staff	101

Summary

Two surveys of the National Science Foundation's Division of Science Resources Statistics (SRS)—the Survey of Federal Funds for Research and Development (the federal funds survey) and the Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions (the federal support survey)—provide some of the most significant data available to understand research and development (R&D) spending and policy in the United States. Building blocks for virtually every analysis of U.S. scientific activity, they help reach conclusions about fundamental policy questions, such as whether a given field of research is being adequately funded, whether funding is balanced among fields, whether deficiencies in funding may be contributing to a loss of U.S. scientific or economic competitiveness, and which agencies are most important for the health of a scientific discipline. Budget officials at science agencies, Congress, and interest groups representing scientists, engineers, and high-technology industries, among others, constantly cite the survey results—or studies based on those results—in making public policy arguments.

However, the survey data are of insufficient quality and timeliness to support many of the demands put on them. For example, reporting agencies sometimes do not assign enough attention to proper recording and timely transmittal of the data. The surveys ask for information in categories that not all agencies use for their own internal purposes, so the information provided to SRS is often a rough estimate, frequently based on unexamined assumptions that originated years earlier.

Although the data from these two surveys have very important uses, the surveys are increasingly difficult to conduct in times of constrained resources, and their technological, procedural, and conceptual infrastructure has not been modernized for procedure or content, in contrast with other surveys in the portfolio of SRS. SRS has recognized the need to upgrade these surveys and to implement recommendations from two previous National Research Council (NRC) studies—*Measuring the Science and Engineering Enterprise: Priorities for the Division of Science Resources Studies* (2000) and *Measuring Research and Development Expenditures in the U.S. Economy* (2004)—which reviewed the federal funds and the federal support surveys as part of the broader SRS portfolio.

With these issues in mind and at the request of the SRS, the Committee on National Statistics of the National Research Council convened this panel to review the uses and collection of data on federal funds and federal support for science and technology and to recommend future directions for the program based on an assessment of these uses and the adequacy of the surveys. The panel was also asked to consider the classification structure, or taxonomy, for the fields of science and engineering, which provides the framework for the federal funds survey as well as other SRS surveys.

The panel has engaged in a variety of activities as part of its responsibilities. We have reached out to senior officials of federal agencies that provide the federal funds data and key data users and solicited advice from providers of complementary and competing data sources. The panel also reviewed past studies on federal funds data, identified common requirements, and considered new data elements and fields that could be useful to collect. As part of our information-gathering activities, the panel conducted a workshop on September 5-6, 2008, at which SRS and outside experts reviewed the uses and collection of data on federal funds for research and development, and assessed the adequacy of the surveys based on the uses. In the workshop, presenters addressed new and emerging methods of data access and retrieval, and recent federal government initiatives to increase the reliability and transparency of contract and grant databases. The workshop concluded with presentations on the issue of an appropriate fields of science classification structure. This report, with recommendations on modernizing the infrastructure of the survey, is the primary product of the study.

The purpose of this report is to provide a pathway for SRS to follow, with the support of the Office of Management and Budget (OMB) and other federal agencies, in order to achieve some modest short-term improvements in the surveys while beginning to build a foundation for a much fuller, more useful R&D data system in the long term. In this report, we define the short term as the next 1 to 4 years; medium-term improvement actions are laid out for a period of 4 to 10 years; and long-term actions are understood to extend beyond the 10-year window. The timing of the pathway for change is outlined in the final chapter.

In recognition of the constrained resources available for making changes in these surveys, the panel's overarching conclusion is that it would be prudent for SRS to make a few short-term improvements to the current system of surveys and then to spend most of the available professional staff time and financial resources pursuing a solution in the medium and long term that involves making use of the new technology and automated databases that will soon be available. One issue to which resources should be devoted in the short run is to reconcile differences in the taxonomies used in the Survey of Federal Funds for Research and Development and its companion survey, the Survey of Research and Development Expenditures at Universities and Colleges (also known as the academic R&D expenditures survey).

Recommendation 3-1: The Division of Science Resources Statistics, in the near term, should make the changes necessary to improve the comparability of the federal funds taxonomy and the taxonomy for the academic research and development expenditures survey and should focus on the medium- and long-term changes the panel recommends.

The panel is convinced that high-level SRS staff involvement with responding agencies should go a long way toward demonstrating to them that SRS considers the data to be important and values their input. Much of the direct contact with reporting agencies has been relegated to the contractors who manage the data collection. The panel therefore recommends that SRS find the resources to establish formal linkages between its own staff and the individuals responsible for data collection and reporting in the various reporting agencies.

Recommendation 3-2: The Division of Science Resources Statistics should devote staff and resources to managing relationships with responding agencies directly, relying less on contractors to maintain those relationships.

The outreach effort would be assisted by the establishment of a more formal mechanism to achieve feedback on an ongoing basis and to provide a forum for guidance as demonstration and evaluation projects are mounted. The panel notes that, in contrast to the companion federal support survey and other SRS surveys that go out to the public, the federal funds survey does not provide respondents with any background on the law under which the data are collected, nor does it outline the important uses of the information. It is essential for SRS to regularly remind agencies about the authority for and importance of the survey. To accomplish this end, information about authority and uses could profitably be placed on the survey form and the associated website.

Recommendation 3-3: The Division of Science Resources Statistics should ensure that all questionnaires and email solicitations sent to respondents provide information on its data collection authority and on the important uses of the data.

The timeliness of reporting is an issue affecting the quality of the R&D investment data. The data have diminished utility owing to the lags in their publication. One way of improving timeliness in this situation would be to impute or estimate the data for late respondents before all data have been received and to publish an estimate. However, some of the late reporters are among the largest supporters of R&D, and including estimates for their data in totals could lead to misleading results. It would be better to continue to work with the late reporters to improve the timeliness of their submissions.

Recommendation 3-4: The Division of Science Resources Statistics (SRS) should maintain its current approach to data reporting, which is to wait for receipt of reports from all respondents before publishing the data. SRS should continue to report complete data without imputation for missing reports and data elements. The agency should focus on working directly with respondents to find ways to improve the timeliness of their response to the surveys.

Although SRS has developed web-based applications for reporting the data, the reporting formats do not make full use of the capabilities of the Internet, nor do they automate many of the functions that would make reporting easier for the responding agencies.

Recommendation 3-5: The Division of Science Resources Statistics should invest in creating more user-friendly web surveys, possibly tailored to each agency, to replace current web versions of the paper surveys.

Several government-wide initiatives hold promise of increasing the quality and availability of administrative records on government expenditures: the E-Government Act of 2002, the Federal Funding Accounting and Transparency Act of 2006 (Public Law 109-282), and recent initiatives on the part of OMB to put in place a significant program to standardize, enhance and validate the data that reside in the federal government's contract and grant databases. These efforts also provide a means of meeting the need for data on R&D expenditures and other kinds of information that are now obtained by means of surveys and data calls. In addition, further improvements can be expected over time—for example, enhancing administrative records with identifiers or "tags" to assist in data retrieval will improve data access.

Recommendation 4-1: The Division of Science Resources Statistics, in cooperation with the Office of Management and Budget and the Office of Science and Technology Policy, should seek to have all federal agencies that fund or conduct research and development (R&D) to incorporate R&D descriptors (tags) into administrative databases. Ideally, in order to enable identification of the R&D components of agency or program budgets, tags should identify: the specific field of science and engineering; whether a record applies to R&D or R&D plant; and whether the record activity is basic research, applied research, or development.

Other data enhancements are possible as the databases are improved in the long term. One necessary enhancement is to enrich the administrative databases with information on intramural R&D activities—those R&D activities conducted in agency-operated laboratories and other facilities that are not likely to be reflected in the contract and grant data systems.

Recommendation 4-2: The Division of Science Resources Statistics should work with the Office of Management and Budget to seek endorsement to work with other research and development funding agencies to incorporate intramural data into existing and future databases or to directly access intramural spending information from performer databases.

Demonstration projects offer the opportunity for SRS to work with reporting agencies and other stakeholders to achieve their collective needs.

Recommendation 4-3: The Division of Science Resources Statistics should initiate work with other federal agencies to develop several demonstration projects to test for the best methods to move to a system based at least partly on administrative records.

The policy context will define the future demand for information on federal R&D expenditures. Novel approaches, such as data federation, automatic text, and linkage analysis, promise to contribute to the development of federal R&D databases rich with detail and increasingly transparent and usable. The panel has identified some promising possibilities for the future of collecting information on federal R&D spending.

This report lays out the pathway for SRS to follow in order to move the collection of data on federal R&D spending from today's survey-centric model to an administrative data–based system. This could be accomplished in a series of overlapping steps, beginning with a series of modest improvements. At the same time as these modest improvements are under way, SRS should begin serious coordination with OMB, most likely through the Office of Science and Technology Policy, to initiate a process that will lead to additional data items being incorporated into the administrative databases. Taking into account lessons learned in the development of the National Institutes of Health's comprehensive database system (the Research, Condition, and Disease Categorization system), SRS has the opportunity to promulgate similar comprehensive systems in other agencies—systems that incorporate taxonomic elements and permit cross-walks between programs, projects, and fields.

1

Introduction

Two surveys of the National Science Foundation (NSF)—the Survey of Federal Funds for Research and Development (the federal funds survey) and the Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions (the federal support survey)—provide some of the most significant data available to understand federal research and development (R&D) investment trends and patterns, illuminating science policy in the United States. Indeed, theses two surveys are building blocks for virtually every analysis of publicly sponsored U.S. scientific and technical activity. They are used by government, academia, industry, and a host of nonprofit analytical and advocacy groups as the primary source of information about federal spending on research and development. For example, they are used by the National Science Board as a basis for its statutorily mandated biannual report on science and engineering (S&E) indicators. In addition, they are one of the primary sources for the analyses of the federal R&D budget prepared regularly by the American Association for the Advancement of Science and other organizations (National Research Council, 2005b, p. 102). Indeed, only the federal budget documents issued by the Office of Management and Budget are as important a source of information on federal R&D spending.

The surveys are used to help reach conclusions about important and fundamental policy questions, such as whether a given field of research is being adequately funded, whether funding is balanced among fields, whether deficiencies in funding may be contributing to a loss of U.S. scientific or economic competitiveness, and which agencies are most important for the health of a scientific discipline.

Users of the survey results typically consider the information to be straightforward, accurate, and complete. However, none of these descriptors is quite the case. Federal agencies that report their R&D spending to NSF treat the surveys with differing degrees of attention to timeliness and accuracy. Some agencies periodically change their internal classifications and the ways in which they account for R&D spending, so the data have problems at their source.[1] Even more problematically, the surveys ask for information in categories that are not used by all agencies for their own internal purposes, so the information provided to NSF is often a rough estimate, frequently based on unexamined assumptions that originated years earlier.

A key component of the reporting of federal R&D spending is the fields of S&E taxonomy, and Macro International recently conducted a study for NSF of its use by federal agencies (Macro International, 2008, p. 5). The study found that some of the major R&D agencies estimate spending by fields of science using staff judgment calls, rules of thumb, percentage distributions, or mapping of the codes to the agency's plans or organizational structure. Some of the agency decisions made in these ways are rather arbitrary. For example, for purposes of federal funds survey reporting, the Centers for Disease Control and Prevention reports its entire research portfolio under one category, Life Sciences–Medical Sciences (Macro International, 2008, p. 7).

With these issues and others in mind, NSF asked the Committee on National Statistics of the National Research Council to review the two main surveys that are used to collect data on federal R&D spending and to consider ways to improve their accuracy and timeliness. Accordingly, the Panel on Modernizing the Infrastructure of the National Science Foundation Federal Funds Survey was established to consider the uses of the NSF federal R&D spending data and, in view of those uses, the quality of the data on federal funds for research and development and to recommend future directions for the program. The panel was asked to include the fields of science classification structure underlying the Survey of Federal Funds for Research and Development in its review.

In approaching this task, the panel has reached out to both senior officials of federal agencies that provide the federal funds data and key data users, solicited advice from providers of complementary and competing data sources, and reviewed past studies on federal funds data. This report, with recommendations on modernizing the infrastructure of the survey, is the primary product of the study.

[1]For example, in fiscal year (FY) 2000, the National Aeronautics and Space Administration reclassified as research activities that had previously been classified as development, and in FY 2004 the agency implemented a new budget approach. Both actions introduced major discontinuities in the R&D data series (National Science Foundation, 2008a, p. 6).

CONTEXT FOR THIS REVIEW

The work of the panel took place during a dynamic time for the U.S. government's S&E policy. In recent years, there has been increasing concern about the role that federal investment in research and development plays in generating innovation and concomitant growth in the U.S. economy. In view of the perceived interconnection between investment in R&D and innovation, there is strong interest in the adequacy of that investment. Furthermore, the focus of top-level decision makers has shifted in recent years from concern about the supply of new knowledge in particular fields and disciplines to a concern about meeting the demands for knowledge to help resolve critical societal challenges, such as infectious diseases, climate change, energy, and food safety. The data now collected by NSF do not readily allow analyses to illuminate these sorts of questions.

Moreover, these concerns are prompting research managers and analysts alike to refocus attention on the metrics that describe the S&E enterprise. The users of R&D data are now raising larger, longer term questions about how to develop a suite of data that would better inform policy debates without losing the information and historical record encapsulated in the two NSF surveys. More specifically, the changing research environment is leading to questions about whether current measures adequately capture the increasingly multidisciplinary nature of research, whether the current taxonomy of fields of S&E accurately describes the research landscape, whether the old division of S&E activities into basic and applied research and development makes for useful categories, and whether the data as now collected permit a comprehensive analysis of the role of the federal government in innovation and growth.

The ultimate goal of data collection on R&D funding should be to enable science policy researchers to draw a much richer picture of federally funded research and its connections to economic growth and other societal goals. That would include being better able to connect research inputs with outcomes, slice the spending data in many different ways, and understand the links among researchers in academia, government, and industry.

Growing Interdisciplinary Research

One concern, for example, that has both immediate and long-term consequences, is the inability of the surveys to account for the growing trend toward interdisciplinary (or cross-disciplinary or transdisciplinary) research. There is a growing belief that demand-driven, problem-solving R&D is often interdisciplinary in character, and there has been a growing discussion in the scientific and science policy communities about whether such work is being sufficiently promoted or possibly even discouraged (National Research Council, 2005a).

It is also important to understand the trends in interdisciplinarity: That is, whether research is most fertile at the boundaries of disciplines, which appears to be the basis for many federal programs seeking to encourage such research, or whether interdisciplinary change is occurring within the core of contemporary disciplines in ways that fundamentally change them. The federal R&D funding and support surveys would be an obvious place to try to get data to help understand these questions, yet the current surveys are likely to obscure the matter by forcing investments in interdisciplinary research to be reported either within a single field or in a miscellaneous ("not elsewhere classified") category. The difficulty of portraying the growing interdisciplinary nature of federal R&D is exacerbated by the fact that much of the interdisciplinary work takes place across agencies as well as across disciplines. Climate change research is an example. It is important that, to the extent possible, these cross-agency R&D initiatives be described in the same way, so they can be identified and aggregated so as to give a view of the totality of the investment throughout the federal government.

Failure of the Fields of Science and Engineering Taxonomy to Describe Research and Development in Useful Ways

Another issue at the heart of the problem is that the surveys provide little help in drawing connections between the research agenda and either public goods or industrial innovation and competitiveness. Most federal agencies manage programs that are defined by categories related to topic areas. For example, several agencies have common breakouts for cross-cutting programs as defined by the Office of Science and Technology Policy and the Office of Management and Budget, including programs categorized by topic areas, such as nanotechnology, climate change, and homeland security. The federal funds survey, however, classifies data by disciplinary fields (such as chemistry, physics, and life sciences); there is no collection of data by topic area. Furthermore, as new fields emerge, old fields are joined in new combinations or decline in importance altogether, and therefore a taxonomy developed around the dominant fields of an earlier era may not provide for an adequate depiction of the relevant data on current federal R&D spending. Beyond that, the currently used taxonomy is now quite uneven with respect to the level of detail reported in various clusters of disciplines, such as the social sciences, the life sciences, and information sciences, in comparison to the physical sciences and engineering.

Antiquated Characterization of Research and Development

The very building blocks of the model of R&D investment are increasingly questioned. Characterizing activities to understand and affect the natural and human environment as basic research, applied research, and development enshrines a linear model that has never been more than a rough approximation of the way R&D actually works, and it is increasingly inadequate as a representation of contemporary reality. In the period following World War II, the U.S. government S&E investment policies increasingly favored basic research. This emphasis was said to have created an organizational disconnect between the federal government's technologically inspired systems for basic research and industry's use-inspired systems for development (Stokes, 1997). A major subsequent preoccupation of science and technology policy since the 1980s has been to bridge the so-called valley of death between research and development and commercialization[2] (Branscomb and Auerswald, 2001, 2002). The data on categories of R&D now collected are useful and still necessary to understand the federal R&D enterprise, but they are not fully adequate to portray that enterprise in today's environment.

Likewise, better metrics are needed to identify and describe the impact of R&D as a source of economic growth. The U.S. economy has significantly altered since NSF first began to compile R&D data. For example, the role of the central industrial R&D laboratory focused on fundamental research has faded and the federal role in basic research has expanded as private basic research has contracted. Studies indicate that federally funded research is now cited in a majority of industrial patent applications, and it underlies many innovations that become successfully commercialized (Broad, 1997; Block and Keller, 2008).

The innovation wave in information technology, for example, was largely underpinned by federal R&D support (National Research Council, 1999, pp. 85-157; Ruttan, 2006, pp. 91-127). Thus, while there is still a strong justification for the conduct of basic research to provide the knowledge basis for improved health, security, and prosperity—the theme of Vannevar Bush's report, *Science the Endless Frontier* (Bush, 1945)—the nation has become more focused on a dual rationale for research: both knowledge and innovation. To the extent that the federal role in the innovation process is growing, more information about the relationship between federal research and subsequent innovation is appropriate. The R&D enterprise is increasingly a matter of interdependence among government, academia, and industry. The role of the federal government in fostering

[2]The "valley of death" refers to the gap between basic research, which is largely federally funded, and applied research and development, which is often industry-funded.

breakthrough innovation needs to be better understood, and more useful data on federal investment in R&D would help.

Increasing Utilization of Administrative Data

While sympathetic to the need for more information to illuminate the trends in research and development and the federal role in that enterprise, the panel is also sensitive to the concern that federal agencies not be overburdened with requirements for additional data reporting. A principal reason the current data are less reliable than desirable is that, in the view of many agencies, filling out NSF reports is labor-intensive and difficult and the benefits barely justify the high cost in labor and other resources (Touhy, 1998). Rather than simply increasing the reporting burden, it would be preferable to consider new data search and analysis technologies tied to expanding efforts to make government data accessible (such as http://www.data.gov), which might gradually make it easier to obtain the raw data for the R&D surveys in ways less burdensome than the current individually conducted compilation processes undertaken separately at each agency.

ORGANIZATION OF THE REPORT

With these user needs and challenges in mind, the panel set about to identify a step-by-step process that would lead to improvement in the federal data on R&D spending. Following this introduction, Chapter 2 provides a description and critique of the current status of the two surveys that now provide the information used to portray federal R&D spending. Chapter 3 focuses on current problems and makes suggestions for a few relatively modest improvements that could be made in the short term (the next four years or so) to the current system of surveys. We then urge NSF to focus attention on a medium-term solution (over roughly 4 to 10 years) that would make use of new technologies and maturing automated databases and set the stage for long-term changes in the collection system—beyond 10 years. Chapter 4 describes the potential and limitations of the use of administrative data for collecting and compiling information on federal R&D spending and identifies opportunities for transitioning to a new system of data collection.

In Chapter 5, we explore some cutting-edge possibilities for long-term changes in the way in which R&D is viewed and the manner in which information about it is collected. In Chapter 6, these threads are gathered into a recommended course of action, which would take NSF through the process of making small short-term improvements in the surveys, undertaking an initiative to build a much fuller, more useful administrative records–based system, and laying the basis for even more revolutionary changes in the long term.

2

The Current Surveys: Challenges and Opportunities

The National Science Foundation Act of 1950 charges the National Science Foundation (NSF) with providing "a central clearinghouse for the collection, interpretation, and analysis of data on scientific and engineering resources, and to provide a source of information for policy formulation by other agencies of the Federal Government." NSF has carried out that task for many years through its Division of Science Resources Statistics (SRS) and its predecessor, the Division of Science Resources Studies.

SRS manages a family of five surveys bundled into the Research and Development Statistics Program, two of which are specifically designed to elicit information about the scope of spending on research and development (R&D) by federal agencies: the Survey of Federal Funds for Research and Development (the federal funds survey) and the Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions (the federal support survey). Table 2-1 provides details on the five surveys in the Research and Development Statistics Program.

The federal funds survey asks all federal departments and agencies that conduct or support R&D for information about their financial investments in R&D, and the federal support survey asks those same agencies to report on their financial investments for R&D in individual institutions as well as their spending on a range of educational and student assistance programs related to science and engineering (S&E). Both surveys are administered annually.

The National Research Council (NRC) has reviewed these surveys in two prior studies (National Research Council, 2000, 2005b). These reviews

TABLE 2-1 The SRS Research and Development Statistics Program

Survey	Universe	Description
Survey of Federal Funds for Research and Development (federal funds survey)	Federal departments and agencies that conduct and/or support R&D programs	An annual census survey of the 15 federal departments, their 70 subagencies, and 15 independent agencies that conduct R&D programs. It collects information about the federal funding of R&D in the United States for three fiscal years. The survey measures federal support of, and participation in, national scientific activities in terms of obligations and outlays. Conducted since the early 1950s.
Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions (federal support survey)	Federal departments and agencies that conduct and/or support R&D programs	An annual census survey of the same agencies that respond to the federal funds survey, which report only if they had obligations for science and engineering to universities, colleges, or nonprofit institutions during the past fiscal year. Agencies report obligations separately for each academic institution and nonprofit organization they fund.
Survey of Research and Development Expenditures at Universities and Colleges (academic R&D expenditures survey)	Academic institutions with greater than $150,000 in R&D expenditures in the previous year or doctoral programs in R&D	An annual census survey that collects information on separately budgeted R&D expenditures by academic institutions in the United States and outlying areas. Conducted since FY 1972, it collects information on R&D expenditures by source of funds and by academic field using a taxonomy similar (but not identical) to the one used in the federal funds survey.
Survey of Industrial Research and Development (industrial R&D survey)	For-profit R&D-performing companies, whether publicly or privately held	An annual sample survey that intends to include or represent all for-profit R&D-performing companies, either publicly or privately held. It is the primary source of information on R&D performed by industry in the 50 states and the District of Columbia. Data are collected in the technology categories of biotechnology, software, materials synthesis and processing, and others. Research in nanotechnology is separately identified and asked for in the 4 categories above. Respondents are asked to report energy R&D in areas of fossil fuels, geothermal and solar, nuclear, and all other energy sources.
Survey of State Research and Development Expenditures	State organizations and agencies that sponsor research and development	An annual census survey of 423 state organizations and state agencies that sponsor research and development. Begun in 2006, the survey measures the extent of R&D activity performed and funded by the 50 states, the District of Columbia, and Puerto Rico. Data are not collected by fields of science and engineering.

focused broadly on the entire portfolio of SRS R&D surveys, devoting only minimal attention to the federal funds and federal support surveys. The 2005 report, however, did examine whether the data collected in the federal funds survey were relevant and adequate for their intended uses, and it recommended that SRS reconsider several aspects of the survey operation in order to modernize it. The report concluded that SRS could improve its operations by collecting the information needed to complete the federal science and technology budget framework recommended in an earlier NRC report, *Allocating Federal Funds for Science and Technology* (National Research Council, 1995). The report also urged SRS to begin to work with the U.S. Office of Management and Budget (OMB), under the auspices of the E-Government Act of 2002, which had been recently enacted to develop guidance for standardizing the development and dissemination of R&D project-level data as part of an upgraded administrative records–based system.

The 2005 NRC report also recommended improvements in the federal support survey. It found the survey to be a useful supplement to the federal funds survey—but also concluded that data collection was cumbersome and time-consuming and needed significant modernization. One option offered for modernizing the survey is to more intensively use microdata in administrative records that are part of the standardized, automated reporting systems in the key federal agencies that provide the bulk of federal support to academic and nonprofit institutions. The 2005 report also briefly considered issues regarding the taxonomy of fields of S&E used by these surveys, recommending that OMB initiate a review of its Classification of Fields of Science and Engineering, last published in 1978 as Directive 16, and that OMB appoint the SRS division of NSF as the lead agency for the effort.

Since these earlier studies, the environment of the surveys on federal R&D spending has changed dramatically. Under the guidance of the E-Government Act and subsequent legislation, OMB has made progress in developing administrative databases for contracts and grants, establishing new standards for the quality of the data, and establishing guidance to making the new databases available to the public in a timely manner. In addition, SRS commissioned a major study by Macro International to assess the quality of the information provided in the surveys, particularly with regard to the classification and reporting of fields of S&E (Macro International, 2008). OMB also assigned responsibility to SRS for reviewing and upgrading the fields of S&E classification, as recommended by the 2005 NRC panel.

In this chapter, our panel examines the federal funds survey and the federal support survey in order to provide the basis for updated and more focused advice on collecting more accurate, timely, and useful data of the type now collected by these surveys.[1]

SURVEY OF FEDERAL FUNDS FOR RESEARCH AND DEVELOPMENT

Since the early 1950s, SRS has conducted the federal funds survey as an annual census survey of federal science and technology agencies. There are now 15 federal departments, 70 subagencies, and 15 independent agencies that conduct R&D programs.[2] The survey is conducted for SRS by a private-sector contractor (Macro International).

The survey has a response rate of 100 percent for both reporting units and survey items. This high rate of survey completion comes at a high price, as discussed later in this chapter and addressed more fully in Chapter 3, because the release of the survey results is often delayed for a year or more awaiting the receipt of reports from all of the reporting departments and agencies.

Years Covered

Data collected on the federal funds survey cover three federal fiscal years: actual (final) expenditures for the most recently completed fiscal year, preliminary estimates of actual expenditures for the current fiscal year, and projected expenditures for the forthcoming fiscal year based on the president's budget submission to Congress. The best-documented data are those for the completed fiscal year, which are based on actual R&D outlays made or R&D obligations entered into by federal entities. The data for the current fiscal year are a stew of amounts in the congressional appropriations, obligation actions to date, and apportionment and reprogramming decisions

---

[1]Information in this chapter is based on SRS's online survey descriptions, documentation provided to the panel by SRS, the findings of a quality profile on the Research and Development Statistics Program (Bailar, 2004), presentations made at the panel workshop (see Appendix C), and additional information provided by SRS at the panel's request.

[2]Departments: Agriculture (10 subagencies); Commerce (5 subagencies); Defense (15 subagencies); Education; Energy; Health and Human Services (11 subagencies); Homeland Security (4 subagencies); Housing and Urban Development; Interior (4 subagencies); Justice (3 subagencies); Labor (6 subagencies); State; Transportation (9 subagencies); Treasury (3 subagencies); and Veterans Affairs. Independent agencies: Agency for International Development; Appalachian Regional Commission; Broadcasting Board of Governors; Consumer Product Safety Commission; Environmental Protection Agency; Federal Communications Commission; Federal Trade Commission; General Services Administration; Library of Congress; National Aeronautics and Space Administration; National Archives and Records Administration; National Science Foundation; Nuclear Regulatory Commission; Smithsonian Institution; Social Security Administration.

made and documented during the year. The least reliable data are those for the president's budget year. These projections represent the amounts in the administration's budget proposals not yet passed by Congress. The overall amounts in the preliminary and projected estimates are subject to revision as the president's budget is executed and authorization, appropriation, deferral, and apportionment actions completed after these data were collected are realized over time.

There is little hard information on which to base estimates of expenditures by fields of S&E for the preliminary and projected years. The Congressional Research Service (CRS) reported that these estimates are computed as a percentage of current-year spending. Agency representatives told CRS that projecting by broad fields of S&E is, at best, an "educated guess" (Congressional Research Service, 2000, p. 10).

These uncertainties in reporting and the inevitable discrepancies among budgets, plans, and actual expenditures contribute to fairly large revisions in the data as the agencies accumulate information on their actual expenditures over time. As Table 2-2 indicates, some of these revisions can be quite significant. During periods of budget uncertainty, as when there is a change in administration, these revisions are even more marked. In 2002, for example, the projected level of federal R&D spending was about 6 percent lower than finally reported; the year after, the total of agency projections was about 5 percent greater than the total of final estimates made when actual spending data became available. Therefore, generally the projected data are thought to be useful only in suggesting broad trends.

Variables Collected

Data collected for each year include outlays and obligations. The key variables collected and their definitions (taken from the survey questionnaire) are found in Table 2-3. Only the 10 largest agencies report obligations for plant by performer and state. They account for about 97 percent of total R&D and R&D plant obligations each year. Only the six largest agencies report the data for obligations to colleges and universities by field of science.

Reporting agencies are challenged by the fact that their internal records do not use the categories requested by SRS. For example, internal records do not separate expenditures on R&D plant from expenditures for the conduct of R&D. This is one example of the kind of measurement error that arises because of the difficulty agencies have in translating the data from the categories in which they are maintained on agency records into the categories that are requested by SRS.

TABLE 2-2 Federal Obligations for R&D and R&D Plant: Projected, Preliminary, and Final, 1990-2008 (in millions of dollars)

Year	Projected	Preliminary	Final	Percentage Change, Projected to Preliminary	Percentage Change, Projected to Final
1990	68,524	63,353	65,831	–7.5	–3.9
1991	66,690	66,227	64,148	–0.6	–3.8
1992	73,484	75,586	68,577	+2.9	–6.7
1993	75,045	75,303	70,415	+0.3	–6.2
1994	74,436	72,818	69,451	–2.2	–6.7
1995	71,746	73,029	70,443	+1.8	–1.8
1996	70,906	71,048	69,399	+0.2	–2.1
1997	70,149	71,996	71,753	+2.6	+2.3
1998	71,593	74,202	73,914	+3.6	+3.2
1999	75,330	77,650	77,386	+3.1	+2.7
2000	77,186	81,772	77,356	+5.9	+0.2
2001	83,609	85,452	84,003	+2.2	+0.5
2002	84,938	97,465	90,158	+14.7	+6.1
2003	103,114	101,008	97,928	–2.0	–5.0
2004	105,220	106,488	105,371	+1.2	+0.1
2005	110,193	113,118	112,995	+2.7	+2.5
2006	112,076	116,873	112,271	+4.3	+0.1
2007	116,417	116,700		+0.2	
2008	113,213				

SOURCE: National Science Foundation (2009a).

TABLE 2-3 Variables and Definitions of Items Collected on the Survey of Federal Funds for Research and Development

Variable	Definition
Outlays for R&D and R&D plant by year	Outlays (expenditures) represent the amounts for checks issued and cash payments made for research and development activities and plant (facilities and fixed equipment) during the fiscal year.
Obligations for R&D and R&D plant by year	Obligations are the amounts of orders placed, contracts and sub-grants awarded, goods and services received, and similar transactions during a given period that will require payment by the grantee during the same or a future period for research and development activities and plant (facilities and fixed equipment) during the fiscal year.

TABLE 2-3 Continued

Variable	Definition
Obligations for basic, applied, and total research by field of S&E (detailed field for past fiscal year, broad field for current and next years)	Basic research is systematic study directed toward fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts without specific applications towards processes or products in mind. Applied research is systematic study to gain knowledge or understanding necessary to determine the means by which a recognized or specific need may be met. Development is systematic application of knowledge or understanding directed toward the production of useful materials, devices, and systems or methods, including design, development, and improvement of prototypes and new processes to meet specific requirements. A field of S&E is a recognized category of specialized expertise within S&E as defined by OMB Directive No. 16.
Obligations for basic and applied R&D by performer by year	A performer is either an intramural group or organization carrying out an operation or function or an extramural organization or person receiving support or providing services under a contract or grant.
Obligations to individual federally funded R&D centers for past year	Federally funded R&D centers are S&E performing organizations that are exclusively or substantially funded by the federal government, administered by an industrial firm, a college or university, or another nonprofit institution.
Obligations for R&D for foreign performers for past year	Foreign performers include foreign citizens, organizations, universities and colleges, or governments as well as international organizations (such as United Nations organizations).
Obligations for R&D plant by performer by year	See above.
Obligations for R&D and R&D plant by state for the past year	States include state and local government agencies, excluding colleges and universities. State data are reported only by the 10 agencies with largest R&D obligations.
Obligations for R&D to universities and colleges for field of S&E (detailed field for past fiscal year, broad field for current and next years)	Universities and colleges are institutions engaged in providing resident and/or accredited instruction for at least a 2-year program above the secondary school level. These data are reported only by the six agencies with the largest R&D obligations.

SOURCE: NSF, Survey of Federal Funds for Research and Development, Questionnaire. Available: http://www.nsf.gov/statistics/srvyfedfunds/surveys/srvynondod_fy03-05.pdf.

For example, actual data on outlays and obligations are available for only the past year, and agencies are asked to make their best estimates for both outlays and obligations for the current and future years. R&D plant data are underreported because of difficulties encountered by the Department of Defense (DoD), the National Aeronautics and Space Administration (NASA), and others in identifying and reporting them. DoD reports obligations for R&D plant funded under the agency's appropriation for construction, but it is able to identify only a small portion of the R&D plant support from R&D contracts funded from DoD's appropriation for research, development, testing, and evaluation. Similarly, NASA cannot separately identify the portions of industrial R&D contracts that apply to R&D plant, since these data are subsumed in the R&D data covering industrial performance. NASA R&D plant data for other performing sectors are reported separately.

Categorization of Research and Development Activities

A major issue affecting data accuracy is discrepancies in how federal agency staff assign their research spending into the basic research, applied research, and development categories, which may result in a lack of comparability in reporting the categories among federal agencies. In its 2000 study, the Congressional Research Service observed that "while SRS provides agencies and survey respondents with definitions of each data item collected on its surveys (e.g., basic research, applied research, and development, etc.), agencies, as well as individuals within agencies, interpret the definitions differently" (Congressional Research Service, 2000, p. 9).

The CRS study further observed that some agencies found it difficult to determine whether certain activities, such as training, program evaluation, and construction of facilities, met the criteria for inclusion in the categories (p. 9). There is little doubt that individual staff members sometimes make arbitrary judgments in order to report in the requested categories. Quite often, these staff members work in agency budget offices that have been assigned the task of completing the SRS reports, rather than in the scientific and technical program offices, where staff might more readily distinguish among the categories.

The panel is aware that the practice of characterizing scientific and technical inquiry as basic or applied research or development has long been controversial. The so-called linear model underlying these categories (basic research → applied research → development) has been met with some skepticism (Godin, 2005), and is widely understood to be a gross oversimplification. This linear model often fails to describe the actual processes of moving from inquiry to application. Furthermore, whether a particular project or program falls into one or another of the categories depends, to some degree, on the intentions and perceptions of those making the categorization. The panel chose not to delve into this subject or to make recommendations about it, but we recognize that it is one of the issues that will need to be addressed in order for SRS to act on our recommendations to move ahead with deriving the federal funding data reports from administrative databases, as this report recommends.

Timeliness of Release

Although the federal funds survey enjoys 100 percent response rates, this is achieved, in part, because of the SRS policy that the data are not released until all agencies have reported. This policy puts SRS at the mercy of the slowest respondent and can severely delay publication of the data.

The delay in publication can be measured by the lag from the period of reference for the data to the time of publication. For example, the data for fiscal years 2006-2008 were solicited in mid-February 2007, with a due date in mid-April 2007. The last of the agency reports was not received until some eight months later, in mid-December 2007. Adding in the time for the lengthy process at SRS to aggregate and analyze the data as well as the agency inputs for release, publication of the data for these years did not occur until January 2009 (National Science Foundation, 2009a, p. 2). In this illustration, the “most recent” year data were actually being reported in January 2009 for outlays that had been completed by September 2006, some 28 months earlier. This was actually an improvement from prior years, when the final agency submissions were received as late as June of the year following the year of solicitation.

Clearly, time lags of this duration, resulting from the failure of agencies to submit timely reports as well as from the lengthy procedures at SRS, reduce the usability of the data for real-time planning and policy-making purposes. Although SRS is able to publish the federal funding data without introducing error-prone imputation for missing agency reports, it does so at a high price: the data are usually well out of date by the time they are finally published.

Collection Technology

On the positive side, SRS has made significant enhancements in recent years to the procedures and technology used for collecting these data from federal agencies. FEDWeb, a web-based data collection system, is used to collect and manage data for the federal funds survey. Most data collection efforts, data imports, data editing, and trend checking are accomplished using FEDWeb. This web-based system helps improve survey reporting and reduce data collection and processing costs by offering respondents direct online reporting and editing. The web-based option has proven popular with most agency reporters. In the 2006-2008 survey cycle, 89 of the 100 reporting agencies submitted their data via FEDWeb. Three others submitted at least part of their data in Microsoft Excel files, which were then entered into FEDWeb by the SRS contractor. However, while modernization of the data collection method might have been expected to assist in improving the consistency and completeness of the data, the transition to web-based collection has not dramatically reduced the delays in data collection and reporting.

TAXONOMY OF FIELDS OF SCIENCE AND ENGINEERING

A key component of the federal funds survey is the taxonomy it uses to classify spending by fields of S&E. The taxonomy is anchored in a classification system that was last updated by OMB in 1978 in its Directive 16 (U.S. Office of Management and Budget, 1978). A specific part of the panel's charge is to assess this classification structure.

The Challenge of Interdisciplinarity

Several challenges are associated with the current taxonomy of fields of S&E used in the federal funds survey. A significant one is that it does not account for new and emerging science or the more recent phenomenal growth in interdisciplinary, multidisciplinary, and transdisciplinary research.

In this report, *multidisciplinary* approaches are defined as when researchers maintain their disciplinary and professional perspectives but add breadth and available knowledge, information, and methods from other disciplines. *Interdisciplinary* approaches integrate separate disciplinary data, methods, tools, concepts, and theories in order to create a holistic view or common understanding of a complex issue, question, or problem. In *transdisciplinary* approaches, researchers develop comprehensive frameworks through an overarching synthesis, such as general systems, policy sciences, feminism, ecology, and sociobiology, in which the disciplines lose their identity.

The importance of updating the taxonomy to better incorporate interdisciplinary research is widely recognized by policy makers, funding agencies, professional organizations, and across academia. The growing role of research involving more than one discipline is a serious challenge to any taxonomy of fields and therefore to gathering, analyzing, and using federal funds data based on a single-field taxonomy.

There are four primary drivers of interdisciplinary research—the inherent complexity of nature and society, the desire to explore problems and questions that are not confined to a single discipline, the need to solve societal problems, and the power of new technologies (National Research Council, 2005a, p. 40).

The heightened momentum for interdisciplinary research dates from the late 1970s, roughly the same time as when the current taxonomy used in the federal funds survey was developed. International competition in science-based fields of high technology propelled greater involvement and investment in interdisciplinary research in engineering and manufacturing, computers, biotechnology, and biomedicine. Breakthroughs in scientific research were also occurring increasingly at the interstices of established fields of inquiry. In 1986, a report from the National Research Council assessing major fields of physics noted "rapidly emerging interdisciplinary advances—which are enriching all of science" (National Research Council, 1986a, p. 15).

Subsequent reports have continued to document these trends. A 1990 NRC report on interdisciplinary research called attention to new intellectual understandings of biological systems, problem complexity, the costs of instrumentation and facilities, and increased collaborations between the life sciences and medicine as well as the physical sciences and engineering (National Research Council, 1990). More recently, an NRC report on the NSF Materials Research Science and Engineering Centers Program affirmed that "frequently, the most exciting and important advances in materials science and engineering occur at the interfaces between, or by unconventional combinations of, traditional disciplines" (National Research Council, 2007a, p. 8).

Funding agencies have responded to these developments with new structures and strategies. The National Institutes of Health Roadmap, for example, recognizes that new combinations of analytical skills and disciplines and new technologies are needed to deal with complex challenges of human health and well-being. New technologies of molecular imaging, nanomedicine, and bioinformatics are prominent in the life sciences. New tools of quantitative and computer-assisted mathematical analysis and advanced computing power are also facilitating the sharing of large quantities data across disciplinary boundaries in areas as diverse as medicine and the geosciences (National Institutes of Health, 2009). The implications may not merely be technical. In the journal *Science,* Alan Leshner observed that "new technologies are driving scientific advances as much as the other way around," allowing new approaches to older questions and posing new ones (Leshner, 2004, p. 729).

The Fluid Boundaries of Traditional Disciplines

In addition to the challenge of interdisciplinary research, the changing nature of disciplines has contributed to making the 1978-era taxonomy less relevant to today's world. The current interface between physics and chemistry, for example, has been crossed so often in both directions that authors of *Scientific Interfaces and Technological Applications* reported "its exact location is obscure" and "its passage is signaled more by gradual changes in language and approach than by any sharp demarcation in content" (National Research Council, 1986b, p. 53). Interactions and cross-fertilizations that characterize the interface have been sources of continual advances in concepts and applications across the science of molecules and atoms, surfaces and interfaces, and fluids and solids. "Thirty years ago," Norman Burkhard observed, "the difference between a physicist and a chemist was obvious. Now we have chemists who are doing quantum-level, fundamental studies of material properties, just like solid-state physicists. There's almost no difference" (National Research Council, 2005a, p. 54).

Inadequacies of the Current Taxonomy

The taxonomy of fields of S&E used in the federal funds survey takes little notice of the trends in the conduct and organization of R&D discussed above. The current taxonomy includes eight broad field categories, each including a number of detailed fields. The broad fields are life sciences, psychology, physical sciences, environmental sciences, mathematics and computer sciences, engineering, social sciences, and other sciences. Each broad field includes a subfield of "not elsewhere classified" (n.e.c.). SRS provides illustrative disciplines for each detailed field in its guidance documentation for survey respondents.

The n.e.c. category is used for both multidisciplinary projects involving more than one field and single-discipline projects for which a separate field is not part of the taxonomy. The lack of up-to-date categories has resulted in overuse of the n.e.c. category. "Not elsewhere classified" has become an amorphous category that lumps together many unrelated types of research, including work in new subfields, emergent fields, established interdisciplinary fields, cross-cutting initiatives, "problem-focus" areas of research, and miscellaneous "other." Moreover, it fails to discriminate multidisciplinary and genuinely integrated interdisciplinary activities. Furthermore, there is no specific way to report interdisciplinary research.

A further problem with the R&D taxonomy is that, at the detailed level, SRS uses somewhat different classifications structures for the federal funds survey and its other surveys. Table 2-4 compares the various taxonomies used by SRS for the federal funds survey with another of its surveys—the Survey of Research and Development Expenditures at Colleges and Universities (known as the academic R&D expenditures survey)—and the "official" taxonomy from OMB Directive 16. There are potentially confusing differences among the three.

TABLE 2-4 Comparison of Taxonomies

Federal Funds Survey	Academic R&D Survey	OMB Directive 16
Life Sciences	Life Sciences	Life Sciences
Biological sciences (excluding environmental)	Biological sciences	Biological
Environmental biology		
Agricultural science	Agricultural sciences	
Medical science	Medical	Clinical medical
		Other medical
Life sciences, n.e.c.	Other	
Psychology	Psychology (total)	Psychology
Biological aspects		Biological aspects
Social aspects		Social aspects
Psychological sciences, n.e.c.	Other	Psychological sciences, n.e.c.
Physical sciences	Physical sciences (total)	Physical sciences
Astronomy		Astronomy
Chemistry		Chemistry
Physics		Physics
Physical sciences, n.e.c.	Other	Physical sciences, n.e.c.
Environmental sciences	Environmental sciences (total)	Environmental sciences—terrestrial and extraterrestrial
Atmospheric sciences	Atmospheric	Atmospheric sciences
Geological sciences	Earth	Geological sciences
Oceanography		Oceanography
Environmental sciences, n.e.c.	Other	Environmental sciences, n.e.c.
Mathematics	Mathematical sciences (total)	Mathematics
Computer sciences	Computer sciences (total)	
Mathematics and computer sciences, n.e.c.		
Engineering	Engineering (total)	Engineering
Aeronautical	Aeronautical and astronautical	Aeronautical
Astronautical		Astronautical
	Bioengineering/biomedical engineering	
Chemical	Chemical	Chemical
Civil	Civil	Civil

continued

TABLE 2-4 Continued

Federal Funds Survey	Academic R&D Survey	OMB Directive 16
Electrical	Electrical	Electrical
Mechanical	Mechanical	Mechanical
Metallurgical and material	Metallurgical and material	Metallurgy and materials
Engineering, n.e.c.	Other	Engineering, n.e.c.
Social sciences	Social sciences (total)	Social sciences
Anthropology		Anthropology
Economics	Economics	Economics
		History
		Linguistics
Political science	Political science	Political science
Sociology	Sociology	Sociology
Social sciences, n.e.c.	Other	Social sciences, n.e.c.
Other sciences, n.e.c.	Other sciences (total)	Other sciences, n.e.c.

NOTE: n.e.c. = not elsewhere classified.
SOURCE: National Science Foundation, Survey of Federal Funds for Research and Development, Questionnaire. Available: http://www.nsf.gov/statistics/srvyfedfunds/surveys/srvynondod_fy03-05.pdf; National Science Foundation, Survey of Research and Development Expenditures at Universities and Colleges, FY 2008 Questionnaire. Available: http://www.nsf.gov/statistics/srvyrdexpenditures/surveys/srvyrdexpenditures_2008.pdf; U.S. Office of Management and Budget (1978).

Recent SRS Efforts to Address Problems

The SRS is well aware of the challenges faced by its current taxonomy of fields of S&E. As noted above, at the direction of OMB it has undertaken a project that is to lead to recommendations on how to revise and update OMB Directive 16, on which the taxonomy is based. SRS has commissioned or conducted several recent studies and workshops to address these issues. SRS convened an interagency working panel meeting (hosted at SRI International on October 21-22, 2004), sponsored a study of agency reporting practices, and commissioned a detailed internal review of a proposed social sciences taxonomy. In NSF, a process was initiated to add a data element to its internal R&D project reporting system (called the "E-jacket") to capture fields of science codes. In this section, we summarize the findings of the interagency working meeting and the study of agency reporting practices.

The October 2004 SRI workshop focused on means of updating the fields of S&E taxonomy. This workshop included discussions about the needs of the users of SRS data; the nature of S&E disciplines and multidisciplinary, interdisciplinary, and transdisciplinary work; criteria for good classification systems; data mapping and mining; international issues; and approaches to improving classification systems (Cheney and Park, 2005, p. 2). The workshop participants included the government officials who provide the data and two major groups of users: representatives of policy communities, who use the information to make policy decisions about S&E investments, and representatives of the research community, who use the data to study the characteristics of the S&E enterprise.

The workshop identified three options for updating the taxonomy. One is to make only minor revisions to the taxonomy in order to maintain continuity in the data, particularly at the higher levels of aggregation. A second is to expand the taxonomy to increase the level of detail, although that might compound the difficulty of reporting the fields. A third is to restructure reporting to permit respondents to report more than one field, enabling analysis of multidisciplinary, interdisciplinary, and transdisciplinary research. The workshop participants emphasized the importance of consulting with users, representatives of the disciplines, survey respondents, and others as SRS moves forward with any revision of the fields of science and engineering, as well as to consider international, educational, and employment taxonomies in the process (Cheney and Park, 2005, p. 7).

Reaching out to other federal agencies, SRS organized and hosted an interagency working group on June 16, 2006, consisting of representatives from OMB and the six largest research-supporting federal agencies: NSF, the National Institutes of Health (NIH), DoD, the Department of Energy (DoE), NASA, and the U.S. Department of Agriculture (USDA). Its aim was to lay the groundwork for the review and revision of OMB Directive 16, identify the primary uses of R&D data organized using this classification, and acknowledge the problems associated with implementing new classification schemes across agencies. Importantly, SRS commissioned Macro International to undertake a study to document (1) the taxonomy each agency uses for analyzing or reporting its research funds (or combined R&D totals, if appropriate); (2) the purposes for which the taxonomy is used; (3) how each agency actually goes about classifying its research/R&D totals into this taxonomy; and (4) how each agency uses the fields of science and engineering taxonomy reported to the federal funds survey (Macro International, 2008, p. 4).

The study of agency reporting practices by Macro International documented interviews with representatives of the 16 major agencies that report on the federal funds survey (Macro International, 2008). The top five sponsors of R&D (NIH, DoD, DoE, NSF, and NASA), accounting for 86 percent of all federal research obligations in FY 2005, stated that they do not use the federal funds taxonomy for their own program management or budgeting. As a result, they consider the data they submit to the federal funds survey to be of questionable value (to them) because of the methods they must use to classify and report them (Macro International, 2008, p. 5).

For example, NIH respondents reported that the agency has no use for the taxonomy because it does not reflect the current state of science and does not include reporting for interdisciplinary research. The DoD respondents reported similar concerns, noting that much of their current cutting-edge research overlaps two or more of the disciplines in the current taxonomy. The DoE respondents reported that the current taxonomy may limit how agencies report their funding. Furthermore, DoE reported confusion about how to classify its own programs, noting that it could more accurately report research activities by theme (e.g., energy, environment, national security). For internal operations, the NSF research program staff (not to be confused with the SRS staff) reported that the current taxonomy is adequate but recognized the need for a taxonomy that reflects new scientific areas and interdisciplinary research. Although the National Institute of Standards and Technology, NSF, DoE, and DoD reported that the resulting data are of some benefit to them, they concurred that the data have limited value for the reasons noted above.

Some interesting uses of the taxonomy were reported. For example, the DoD participant reported using the federal funds taxonomy to compare DoD research funding with funding for the education of graduate students by the U.S. Department of Education. The Macro International report recommended revising the taxonomy to be useful to the reporting agencies, warning that methods used to collect and report data, as well as the quality of the resulting information, are otherwise unlikely to improve.

SURVEY OF FEDERAL SCIENCE AND ENGINEERING SUPPORT TO UNIVERSITIES, COLLEGES, AND NONPROFIT INSTITUTIONS

The federal support survey is an annual census survey of the same agencies that respond to the federal funds survey, but they report only if they had obligations for S&E to universities, colleges, or nonprofit institutions during the past fiscal year (the time frame of this survey). The data that are collected from the 19 agencies (in FY 2006) that made such obligations include federal obligations for R&D; R&D plant; facilities and equipment for S&E instruction; S&E fellowships, traineeships, and training grants; general support for S&E; and other S&E activities.

Like the federal funds survey, this survey has a response rate of 100 percent (both unit and item) because of the SRS policy to withhold release of the data until all agencies have reported. Also like the federal funds survey, the policy of withholding release of the federal support data until all agencies have reported is a contributing factor to excessive lags in publication of the survey results. For example, the data for FY 2006, which ended on September 30, 2006, were solicited in mid-February 2007 with a requested due date of mid-April 2007. The data from the last 2 of the 19 reporting agencies were not received by SRS until November 2007. The brief analysis and abbreviated set of summary tables for FY 2006 were finally released by SRS in October 2008 (National Science Foundation, 2008b), and the full set of detailed information by agency with institutional rankings was not published until March 2009 (National Science Foundation, 2009b).

In this survey, on behalf of SRS, a contractor asks departments and agencies to complete a web-based survey instrument for each university and college for which they obligated R&D funding during the previous fiscal year.[3] There are approximately 1,200 such colleges and universities. In FY 2006, the reporting agencies obligated over $28 billion to those institutions, most of it going to the top 100 universities and colleges (National Science Foundation, 2009b, Table 17).

Agencies also complete a survey instrument for each of the approximately 1,300 nonprofit institutions to which they obligated funds. The list of nonprofit institutions receiving federal R&D funds has been growing. Obligations to universities and colleges are reported in six categories, whereas the obligations to nonprofit institutions are reported only for R&D and R&D plant (see Table 2-5). In FY 2006, over $6.5 billion was obligated to nonprofit institutions, about 80 percent of those funds going to the top 100 recipient nonprofits (National Science Foundation, 2009b, Table 29).

Respondents are instructed that totals for R&D transfers to universities should be similar to those reported by them for R&D to academic institutions in the federal funds survey. If differences exist, respondents are asked to explain. One reason for differences in totals is that methods differ for reporting funds that are transferred from one agency to another before being sent to an institution. For example, the federal support survey asks the agency that distributes the funds to a university to report the transaction, whereas the federal funds survey asks the agency transferring the funds to the agency that ultimately sent them on to report the transaction.

This survey also uses a web-based collection tool akin to the FEDWeb system described above. The FSSWeb system, like the FEDWeb system, is not universally used by the agencies. A few agencies submit their data in alternative formats (e.g., ASCII text files, Excel spreadsheets). In the 2006-2008 survey cycle, 47 agencies reported via FSSWeb, 13 agencies or subagencies reported using Excel files, and, ironically, NSF itself provided an ASCII file.

---

[3]Note that, to be included in the survey, an institution must perform R&D, even though data on funding for topics other than R&D are included in the survey and the report.

TABLE 2-5 Categories for Reporting Federal Science and Engineering (S&E) Support to Universities, Colleges, and Nonprofit Institutions

Category of Support	Academic Institutions	Nonprofit Institutions
R&D	X	X
Fellowships, traineeships, and training grants	X	
R&D plant and equipment	X	X
Facilities and equipment for instruction in S&E	X	
General support of S&E	X	
Other activities related to S&E	X	
All other activities	X	

SOURCE: National Science Foundation, Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions, Overview (FY 2007 cycle). Available: http://www.nsf.gov/statistics/srvyfedsupport/.

The contractor that carries out the survey for SRS must convert these multiple formats into the FSSWeb data system.

Since this survey does not ask for a breakdown of R&D funding by fields of S&E, the problems with completing it are different from those for the federal funds survey. Agencies find it difficult to report funding for each college or university in the six categories requested.

SUMMARY

The federal funds survey and the federal support survey are challenged and fragile, but they are not broken. With its limited resources, SRS has been able to implement a patchwork of improvements and to fund a series of studies that together provide a basis for continuing to provide an indication of the size and direction of federal R&D spending while laying the groundwork for making improvements in the future.

The surveys have a number of problems. The panel judges that two of them—delays in the assembly of data from the reporting agencies and in the publication of results and problems with the taxonomy of fields of S&E—are not solvable in the near term.

If conducted using the current methodology into the future, the surveys are likely to remain fragile, subject as they are to the reporting decisions made by agencies that do not always accord them a high priority and in an environment in which good practices are not always guaranteed and good documentation is not always available. In the next chapter, the panel turns to a discussion of short-term possibilities for improvement. We lay out a series of recommendations for the short term that will serve to keep the surveys viable while establishing the basis for major improvements in both the process of gathering data and the way they are reported.

3

Improving the Current System in the Short Term

Improving statistical practices takes time and money, but the Division of Science Resources Statistics (SRS) of the National Science Foundation (NSF), like other federal agencies, has a limited budget for its statistical activities, most of which must be devoted to accomplishing the day-to-day tasks associated with collecting, processing, and disseminating data to users. The panel therefore advises SRS to make only a few modest improvements to the current system of surveys in the short term (over the next four years) and to spend most of its resources for improvement following a longer term strategy of adopting new technology to extract data automatically from administrative databases that we expect to become available in the next few years.

Our short-term recommendations and the rationales for them are detailed in this chapter. Longer term recommendations follow in Chapter 4. In Chapter 6, the pathway for leading to these changes is described, beginning with the incremental improvements outlined here and progressing to the stage at which the current surveys can be replaced with a system based on administrative records.

Our recommendations for short-term improvements in the two surveys address five major areas: (1) reform of the taxonomies of fields of science and engineering (S&E), (2) SRS's relationships with the responding agencies, (3) the adequacy of SRS's data collection authorities, (4) the timeliness of data collection and reporting, and (5) survey technology. Each is discussed below.

REFORM OF THE TAXONOMIES OF FIELDS OF SCIENCE AND ENGINEERING

Practical Aspects of the Selection of a Taxonomy

For a taxonomy to be useful for research and development (R&D) data, it must not only include categories that are meaningful to users of the data but also be suitable for data collection purposes. Users should experience the categories as being reflective of their reality. At the same time, the taxonomy should reflect the way organizations providing the data are organized and staffed for the support and the performance of R&D.

The panel is aware that not all supporters and performers of R&D classify their activities in the same way and that this fact introduces a certain tension in deciding whether a taxonomy is "suitable." For example, a mission agency might classify a bioinformatics research activity as life science, whereas a funded performer might report the same activity as computer science. For this reason, and others, no single taxonomy will satisfy all. However, for purposes of collecting data on research and development statistics in a consistent manner across federal government agencies, it is necessary to establish a common taxonomy that will be useful to the largest number of data providers and users. In the longer term, a provision can be made for tailoring structures that meet the specific needs of providers and users by flexibly categorizing administrative records as outlined in Chapter 5. Box 3-1 lays out some general issues in creating a taxonomy of activity in research and development.

SRS has made a good start on identifying and analyzing issues with the taxonomy now used. As described in Chapter 2, SRS has recently held workshops and conducted interviews with federal agency respondents to delve into issues surrounding the taxonomy used in the federal funds survey. In these workshops, the agencies have provided SRS with practical comments from their perspective as respondents. For example, some agencies have reported that their databases do not collect or store data in a way that would permit easy reporting by fields of S&E, and that the current taxonomy is not relevant to the way in which they manage and track their R&D programs (Macro International, 2008, p. 8).

Users of the federal funds data who participated in various discussions over the past few years, including the workshop held by our panel, seem nearly unanimous in their view that the current taxonomy has shortcomings. Among other criticisms, users point to the fact that new fields of science are not included, there is no way to report interdisciplinary research, and more detailed breakdowns are needed. These criticisms focus on the need for additional categories in the current taxonomy, rather than a wholesale change in the taxonomy itself. The same users caution against

BOX 3-1
Issues in Creating a Taxonomy of Activity in Research and Development

A taxonomy groups together entities according to their common characteristics. The English word derives from the Greek *τάξις*, *taxis* (meaning order or arrangement), and *νόμος*, *nomos* (meaning law or science). Used originally for biological classification of living and extinct beings, the term now refers to any classification scheme. There is no perfect or ideal classification, Lenoir and Beghtol (2004) caution, only better or worse schemes for particular purposes, contents, users, and contexts. Units and levels of analysis also vary by specificity or "granularity," from shallow analysis dividing entities into large aggregates to deep analysis subdividing them into smaller units. Moreover, taxonomies are never finished and so become outdated.

Since the late 19th and early 20th centuries, taxonomies of knowledge in the West have been dominated by a system of disciplinarity that distinguishes domains of specialized inquiry by their particular subjects and objects of study. The federal funds survey, for example, categorizes fields of science and engineering into eight major types: (1) life sciences, (2) psychology, (3) physical sciences, (4) environmental sciences, (5) mathematics and computer sciences, (6) engineering, (7) social sciences, and (8) other sciences "not elsewhere classified" (n.e.c.). Each major type, in turn, contains subtypes. Physical sciences, for example, encompass astronomy, chemistry, and physics. Over the course of the 20th century, the scope and size of disciplinary domains has expanded with the growing number of specialties and subfields, turning many of them into large groups of disciplines that encompass a broad range of identifiable and in some cases autonomous specialties.

The social sciences illustrate the challenge of classification. The mainstream disciplines of social sciences are anthropology, economics, political science, psychology, and sociology. Yet, Neil Smelser advises, describing social sciences solely with reference to the "big five" disciplines distorts reality in two ways (Smelser, 2003). First, under those headings, subareas of investigation rely on variables and explanations outside the commonly understood scope of social sciences. Geopolitics, sociobiology, behavioral genetics, and behavioral neuroscience all appeal to nonsocial and nonpsychological explanatory variables and explanations. Second, another range of disciplines could be labeled behavioral and social scientific, although not entirely so. Demography might be considered a separate social science or part of sociology, economics, and anthropology. Archaeology might be classed as part of anthropology or

continued

BOX 3-1 Continued

an independent social science. Geography, history, psychiatry, law, and linguistics present similar complications for taxonomy. So do relations with the intersecting fields of genetics, behavior, and society; behavioral and cognitive neurosciences; psychiatry; health; gender studies; religious studies; expressive forms; environmental/ecological sciences and technology studies; area and international studies; and urban studies, planning, and public policy. Strict assignment to one category of inquiry or another would vary according to the criteria used (Smelser, 2003).

The growth of multidisciplinary and interdisciplinary modes of research has further complicated classification, a challenge amply evident in the n.e.c. category of the federal funds survey. The category of "not elsewhere classified" is large and amorphous and lumps together a plurality of developments, including new subfields, single-discipline projects for which a separate field has not been assigned, emergent fields, established interdisciplinary fields, cross-cutting initiatives, "problem-focus" areas of research, and miscellaneous "other." It also fails to discriminate multidisciplinary juxtapositions of different disciplinary approaches from interdisciplinary approaches that integrate separate disciplinary data, methods, tools, concepts, and theories, as well as comprehensive transdisciplinary frameworks that posit a new conceptual synthesis or theoretical framework (Klein, 2009). Authors and users of taxonomies also have different views of how they should be constructed. Birger Hjørland distinguishes four fundamental methods of classification: *empiricism* (based on observations and inductions), *rationalism* (principles of pure reason, deductions), *historicism* (context and development), and *pragmatism* (analysis of goals, values, and consequences in both subject and object) (Hjørland, 2008). The greater plurality and complexity of knowledge today has resulted in three major views of current taxonomy. A first group continues to use standardized classification schemes based on a limited number of broadly aggregated categories, lumping together diverse practices. A second group advocates widening the broad categories, with the possibilities of adding a few more major categories and using a "hierarchy of preference" approach that allows splitting into highly aggregated (2-3 digit) levels and distributing percentages of emphasis and time into more than one discipline. A third group supports a more open, flexible, dynamic, and transactional approach, depicting research in a network representation that allows for greater granularity and employs techniques of semantic mapping, web- and text-mining, controlled thesauri, and tag clouds, and Internet-based, user generated taxonomies ("folksonomies"). The work of the second and third groups inform this study and its recommendations.

making significant changes to the taxonomy, for fear of introducing discontinuity in the historical data it provides.

Associating a government-wide taxonomy for R&D funding with the disciplinary structure of academia is no easy matter. The academic community is most often organized around departments for the purpose of instruction, but research activities are often multidisciplinary, engaging more than one discipline and department. Federal mission agencies tend to organize their applied R&D activities around broad national challenges, such as energy efficiency or space exploration, or around technology areas, such as nanotechnology, biotechnology, or information technology. Thus, updating the taxonomy to make it more relevant for communicating with the academic community might work for basic research, but it will not contribute much when classifying applied R&D programs and projects.

One promising approach is to use federal agency records to quantify the extent of interdisciplinary research. Under the auspices of the National Academies' Keck Futures Initiative, a team from the Georgia Institute of Technology performed keyword searches of awards databases at NSF, the National Institutes of Health, and the Department of Energy to estimate the amount of interdisciplinary research being supported. Over the time period covered by the research (1999-2003), the team found apparent growth in interdisciplinary research (Yasitis et al., 2004). Using agency contract and award information as a source of information on fields of S&E is explored at greater length in Chapter 4.

SRS currently uses slightly different taxonomies for the federal funds survey and the academic R&D expenditures survey, and the taxonomies utilized in these surveys have differences from the standard taxonomy in OMB Directive 16 (see Table 2-1). Ideally, classification of R&D funding by disciplines as reported by the supporters of that R&D should be the same as reported by the recipients of those funds. Uniformity of definitions and classifications between surveys is one means of achieving that goal.

Practically, some differences in taxonomies are likely to persist because it may be difficult to collect the same level of detail from different respondent groups, or there may be differences in the uses to which the taxonomic information is put. That said, because of the need to compare the results of the two surveys, SRS would benefit from using the same taxonomy for the federal funds survey and the academic R&D expenditures survey.

Alternative Taxonomies

The previous National Research Council (NRC) studies discussed in Chapter 2 have laid out the problems associated with an aging taxonomy of R&D spending. Our panel was specifically asked to make a recommendation about updating Directive 16 issued by the U.S. Office of Management and Budget (OMB) in 1978—the taxonomy that was the original basis of the classification of fields of S&E for the federal funds survey.

Several alternative taxonomies could replace or supplement the OMB Directive 16 taxonomy. In considering options for a new taxonomy, the panel looked not only at OMB Directive 16, but also at taxonomies used internationally. Two relevant international standards are "Recommendations Concerning the International Standardization of Statistics on Science and Technology" of the United Nations Educational, Scientific, and Cultural Organization (UNESCO) (United Nations Educational, Scientific, and Cultural Organization, 1978) and "Proposed Standard Practice for Surveys on Research and Experimental Development," called the Frascati manual, of the Organisation for Economic Co-operation and Development (OECD) (Organisation for Economic Co-operation and Development, 2002).

The UNESCO-recommended fields of S&E are shown in Box 3-2. Like OMB Directive 16, the UNESCO fields have not been updated since 1978. However, these old taxonomies were considered when the international community developed the more recent OECD Frascati manual classification, which uses the most current categories in its taxonomy of fields of S&E.

As shown in Box 3-3, the Frascati manual classification provides a useful system for organizing subactivities under the major fields. The Frascati manual recognizes that need for disaggregation within the major fields may differ from country to country, stating in the instructions, "While the major fields of science and technology are clearly defined, the level of disaggregation within each component field is left to each country" (Organisation for Economic Co-operation and Development, 2002, p. 66). Adoption of the Frascati taxonomy is not a full solution, however. Although newer than Directive 16, the Frascati manual suffers from some of the same limitations, in that it fails to accommodate multidisciplinary fields and has no procedures for periodic updating.

Two classification systems, one in widespread use and the other in development, have the advantage of permitting portrayal of multidisciplinary and interdisciplinary fields. Both the U.S. Department of Education's Classification of Instructional Programs (CIP) and the NRC's Taxonomy of Fields (see Box 3-4) are designed to support the collection of information from educational institutions, yet they may have wider application and attributes that commend them for consideration as alternative classification structures to OMB Directive 16.

The CIP was originally developed by the National Center for Education Statistics (NCES) in 1980 to provide a taxonomy to support the tracking,

BOX 3-2
UNESCO Fields of Science and Technology

Natural sciences, including astronomy, bacteriology, biochemistry, biology, botany, chemistry, computer sciences, entomology, geology, geophysics, mathematics, meteorology, mineralogy, physical geography, physics, zoology, other allied subjects.

Engineering and technology, including engineering proper, such as chemical, civil, electrical, and mechanical engineering, and specialized subdivisions of these; forest products; applied sciences such as geodesy, industrial chemistry, etc.; architecture; the science and technology of food production; specialized technologies or interdisciplinary fields, e.g., systems analysis, metallurgy, mining, textile technology, other allied subjects.

Medical sciences, including anatomy, dentistry, medicine, nursing, obstetrics, optometry, osteopathy, pharmacy, physiotherapy, public health, other allied subjects.

Agricultural sciences, including agronomy, animal husbandry, fisheries, forestry, horticulture, veterinary medicine, other allied subjects.

Social sciences, anthropology (social and cultural) and ethnology, demography, economics, education and training, geography (human, economic, and social), law, linguistics, management, political science, psychology, sociology, organization and methods, miscellaneous social sciences, and interdisciplinary, methodological, and historical science and technology activities relating to subjects in this group.

Humanities, arts (history of the arts and art criticism, excluding artistic research of any kind), languages (ancient and modern languages and literature), philosophy (including the history of science and technology), prehistory and history, together with auxiliary historical disciplines, such as archaeology, numismatics, paleography, etc., religion, other fields and subjects pertaining to the humanities, and interdisciplinary, methodological, historical, and other science and technology activities relating to the subjects in this group.

SOURCE: Adapted from United Nations Educational, Scientific, and Cultural Organization (1979).

BOX 3-3
Frascati Manual Taxonomy of Fields of Science and Technology

1. Natural Sciences

1.1. Mathematics and computer sciences [mathematics and other allied fields: Computer sciences and other allied subjects (software development only; hardware development should be classified in the engineering fields)]

1.2. Physical sciences (astronomy and space sciences, physics, other allied subjects)

1.3. Chemical sciences (chemistry, other allied subjects)

1.4. Earth and related environmental sciences (geology, geophysics, mineralogy, physical geography and other geosciences, meteorology and other atmospheric sciences, including climatic research, oceanography, vulcanology, paleoecology, other allied sciences)

1.5. Biological sciences (biology, botany, bacteriology, microbiology, zoology, entomology, genetics, biochemistry, biophysics, other allied sciences, excluding clinical and veterinary sciences)

2. Engineering and Technology

2.1. Civil engineering (architecture engineering, building science and engineering, construction engineering, municipal and structural engineering and other allied subjects)

2.2. Electrical engineering, electronics [electrical engineering, electronics, communication engineering and systems, computer engineering (hardware only) and other allied subjects]

2.3. Other engineering sciences (such as chemical, aeronautical and space, mechanical, metallurgical and materials engineering, and their specialized subdivisions; forest products; applied sciences, such as geodesy, industrial chemistry, etc.; the science and technology of food production; specialized technologies of interdisciplinary fields, e.g., systems analysis, metallurgy, mining, textile technology and other allied subjects)

3. Medical Sciences

3.1. Basic medicine (anatomy, cytology, physiology, genetics, pharmacy, pharmacology, toxicology, immunology and immunohematology, clinical chemistry, clinical microbiology, pathology)

3.2. Clinical medicine (anesthesiology, pediatrics, obstetrics and gynecology, internal medicine, surgery, dentistry, neurology, psychiatry, radiology, therapeutics, otorhinolaryngology, ophthalmology)

3.3. Health sciences (public health services, social medicine, hygiene, nursing, epidemiology)

4. Agricultural Sciences

4.1. Agriculture, forestry, fisheries and allied sciences (agronomy, animal husbandry, fisheries, forestry, horticulture, other allied subjects)

4.2. Veterinary medicine

5. Social Sciences

5.1. Psychology

5.2. Economics

5.3. Educational sciences (education and training and other allied subjects)

5.4. Other social sciences [anthropology (social and cultural) and ethnology, demography, geography (human, economic and social), town and country planning, management, law, linguistics, political science, sociology, organization and methods, miscellaneous social sciences and interdisciplinary, methodological, and historical science and technology activities relating to subjects in this group. Physical anthropology, physical geography and psychophysiology should normally be classified with the natural sciences]

6. Humanities

6.1. History (history, prehistory and history, together with auxiliary historical disciplines such as archaeology, numismatics, paleography, genealogy, etc.)

6.2. Languages and literature (ancient and modern)

6.3. Other humanities [philosophy (including the history of science and technology), arts, history of art, art criticism, painting, sculpture, musicology, dramatic art excluding artistic research of any kind, religion, theology, other fields and subjects pertaining to the humanities, methodological, historical and other science and technology activities relating to the subjects in this group]

SOURCE: Organisation for Economic Co-operation and Development (2002, p. 67).

BOX 3-4
National Research Council Taxonomy of Fields

Life Sciences

Animal Sciences
Biochemistry, Biophysics, and Structural Biology
Biology/Integrated Biology/Integrated Biomedical Sciences
(Note: Use this field only if the degree field is not specialized.)
Cell and Developmental Biology
Ecology and Evolutionary Biology
Entomology
Food Science
Forestry and Forest Sciences
Genetics and Genomics
Immunology and Infectious Disease
Kinesiology
Microbiology
Neuroscience and Neurobiology
Nursing
Nutrition
Pharmacology, Toxicology, and Environmental Health
Physiology
Plant Sciences
Public Health

Emerging Fields:
Bioinformatics
Biotechnology
Systems Biology

Physical Sciences, Mathematics, and Engineering

Applied Mathematics
Astrophysics and Astronomy
Chemistry
Computer Sciences
Earth Sciences
Mathematics
Oceanography, Atmospheric Sciences, and Meteorology
Physics
Statistics and Probability
Aerospace Engineering
Biomedical Engineering and Bioengineering
Chemical Engineering

Civil and Environmental Engineering
Computer Engineering
Electrical and Computer Engineering
Engineering Science and Materials (not elsewhere classified)
Materials Science and Engineering
Mechanical Engineering
Operations Research, Systems Engineering, and Industrial Engineering

Emerging Fields:
Computational Engineering
Information Science
Nanoscience and Nanotechnology
Nuclear Engineering

Social and Behavioral Sciences
Agricultural and Resource Economics
Anthropology
Communication
Economics
Geography
Linguistics
Political Science
Psychology
Public Affairs, Public Policy, and Public Administration
Sociology

Emerging Fields:
Criminology and Criminal Justice
Science and Technology Studies
Urban Studies and Planning

Arts and Humanities
American Studies
Classics
Comparative Literature
English Language and Literature
French and Francophone Language and Literature
German Language and Literature

continued

BOX 3-4 Continued

Language, Societies, and Cultures
History
History of Art, Architecture, and Archaeology
Music (except performance)
Philosophy
Religion
Spanish and Portuguese Language and Literature
Theatre and Performance Studies

Emerging Fields:
Feminist, Gender, and Sexuality Studies
Film Studies
Race, Ethnicity, and Post-Colonial Studies
Rhetoric and Composition

SOURCE: Taxonomy of Fields (http://sites.nationalacademies.org/PGA/Resdoc/PGA_044521).

assessment, and reporting of fields of study and program completion.[1] This tracking is accomplished through the Integrated Postsecondary Education Data System (IPEDS) completions survey, which is submitted by postsecondary institutions that receive Title IV federal funding. This survey summarizes the number of completions by field of study across the full spectrum of school offerings. The CIP provides the lists of the field of study.

Although the CIP includes fields other than S&E, one obvious advantage of associating the fields of S&E taxonomy for reporting R&D data to the CIP is that it is frequently updated to stay current with educational offerings. The CIP was updated in 1985, 1990, and 2000. NCES is currently updating the 2000 CIP with the goal of releasing an updated version in June 2009. These updates are based on rigorous procedures.[2] A disad-

[1]Memorandum, Michelle Coon, "Update of the Classification of Instructional Programs (CIP)," National Center for Education Statistics, August 11, 2008.

[2]One process involves examining data from the IPEDS completion survey to identify institutions that produced the largest number of completions for a specific CIP code. The course catalogs for these institutions are then mapped onto the existing CIP and examined to find instructional programs that did not fit into the existing CIP. Stakeholders and coordinators are involved in this process to identify instructional programs that are not currently represented in

vantage of using the IPEDS process to drive updates of the taxonomy of fields of S&E in the SRS surveys is that it could lead to unanticipated and undesirable changes in the reporting of R&D support.

If OMB and SRS were to more closely align the federal funds fields with the CIP fields, some modifications of the latter would be in order. The CIP taxonomy includes a greater number of categories than are needed for the SRS surveys, so the SRS taxonomy might need to select and combine CIP categories. The burden of this task should be minimal, since SRS currently provides respondents to another of its R&D surveys—the academic R&D expenditures survey—with a cross-walk between the CIP taxonomy and the taxonomy of fields of S&E.

The NRC Taxonomy of Fields is the most recently developed taxonomy for classifying fields of S&E. Developed to support collection of data on research doctorate programs, it is based on the classification of fields used in the Doctorate Records File, which is also maintained by SRS. The criteria for inclusion of a field in this taxonomy are tied to doctoral program production—that is, fields are included that have produced a total of at least 500 Ph.Ds in the past 5 years with participation by at least 25 universities.

One advantage of the NRC classification structure is that it attempts, whenever possible, to specifically include multidisciplinary, cross-disciplinary, and transdisciplinary fields and to make provision for emerging fields exhibiting significant growth. Thus, interdisciplinary fields—such as neuroscience, biomedical engineering, and American studies—are included, and emerging fields—such as bioinformatics, biotechnology, systems biology, computational engineering, information science, science and technology studies, feminist studies, race and ethnic studies, and rhetoric and composition—also make the list (National Research Council, 2003, pp. 19-20; see also National Research Council, 2006).

The NRC system was designed to depict academic research programs, so it would fit well with the federal support survey. However, it was not designed to support collection of R&D data, so it would have to be modified to serve as a general-purpose R&D taxonomy.

Need for Historical Continuity

The need to modernize the taxonomy of fields of S&E must be balanced with the need for historical continuity of the data series that are based on the existing taxonomy. Data series based on consistent definitions of fields

---

the CIP, and non-NCES data sources, such as the College Board's Annual Survey of Colleges and SRS's Survey of Earned Doctorates, are reviewed in order to identify new instructional programs. Finally, input from federal agencies that use IPEDS completions data is solicited.

of S&E go back many years: Several of the series published by field in *Science and Engineering Indicators* (National Science Board, 2006) have been consistently published since the early 1980s. Because of this consistency, it is possible to trace, for example, the dramatic increase in federal obligations for life science fields over the past four decades.

The need for historical continuity suggests caution in proposing changes in the fields of S&E taxonomy. Changes should be incremental, and care should be taken to either carry forward the current taxonomy, even as data are published using a new taxonomy, or to develop cross-walks between any new taxonomy and the old one so as to minimize disruption to the historical data series. Abrupt changes to the taxonomy could confuse data users.

Recommendation 3-1: The Division of Science Resources Statistics, in the near term, should make the changes necessary to improve the comparability of the federal funds taxonomy and the taxonomy for the academic research and development expenditures survey and should focus on the medium- and long-term changes the panel recommends.

SRS'S RELATIONSHIPS WITH RESPONDING AGENCIES

SRS collaborates with the responding agencies through regular workshops and has initiated several efforts to study reporting issues. The 2008 report by Macro International is the most recent example. In that report, Macro International staff contacted respondents by email to solicit information and schedule face-to-face meetings for in-depth interviews. Although the information provided in the report is of high quality and ultimately very useful, the contracted investigators faced many challenges with regard to access to the agency respondents. The Macro International staff reported difficulties in identifying appropriate high-level personnel in each agency and in obtaining cooperation in some. Furthermore, the panel observes that SRS would have benefited from the social professional capital that can be built through an ongoing program of having its own staff conduct structured in-depth interviews and discussions with agency staff about the surveys, data, and reporting issues. Building relationships between SRS staff and agency respondents is critical to ensuring goodwill, understanding how the agency internal reporting systems interface with the survey questionnaire, improving reporting, and ultimately obtaining high-quality data.

The panel concludes that SRS would benefit from building better direct relationships with agencies responding to the federal funds and federal support surveys. The core competency and management capacity to maintain relationships with other agencies is an inherent governmental responsibility, and SRS loses an opportunity to improve practices when it relies on contractors to facilitate and manage these relationships. Senior-level SRS staff involvement with responding agencies should go a long way toward demonstrating to agencies that SRS considers the data to be important and that it values their input.

Recommendation 3-2: The Division of Science Resources Statistics should devote staff and resources to managing relationships with responding agencies directly, relying less on contractors to maintain those relationships.

ADEQUACY OF THE DATA COLLECTION AUTHORITIES

We considered whether SRS needs more explicit statutory authority to collect data from federal agencies in a way that would improve reporting. The panel notes that agencies are already required to respond to the federal funds and federal support surveys (Congressional Research Service, 2000, p. 1). The panel concludes that restating SRS's statutory authority would be unlikely to affect how agencies maintain their data, so difficulties in responding to the questionnaire would continue.

Instead, SRS would benefit from pursuing better relationships with responding agencies, reminding them regularly of the importance of the survey and the usefulness of the data and working with them to make the survey forms as easy as possible to complete. This activity could be initiated as one of the major activities on the path toward modernized collection of R&D spending data. A time schedule for this activity is suggested in Chapter 6.

Some cosmetic changes in the collection forms might be helpful. For example, in contrast to its companion federal support survey and surveys that go out to the public, letters to respondents, survey forms, and instructions for the federal funds survey do not provide respondents with any background on the law under which the data are collected, nor do they address the important uses of the information. It is important for SRS to regularly remind agencies about the authority they have and importance of the survey with each call for data. To accomplish this end, information about authorities and uses could be featured prominently on the survey form, in the instructions, and on the associated website.

Recommendation 3-3: The Division of Science Resources Statistics should ensure that all questionnaires and email solicitations sent to respondents provide information on its data collection authority and on the important uses of the data.

TIMELINESS OF DATA COLLECTION AND REPORTING

The inability of SRS to release the data in a timely way limits the usefulness of the federal funds survey and the federal support survey data to policy makers and interest groups. As previously noted, the delays result largely from the current practice of delaying publication of results until reports are received from all reporting agencies and the long delays of some agencies in assembling and forwarding the data. Some of the agencies that have delayed transmittal of their data in the past are quite large, and the publication of the estimates without their contribution could severely distort the data. For example, the two slowest reporters in 2006 were the National Institutes of Health and the National Aeronautics and Space Administration, together accounting for about one-third of federal R&D expenditures.[3]

SRS could consider alternatives to this practice. Alternatives include reporting incomplete information earlier, providing a preliminary report with imputation for late respondents, or designing a simple schedule to the form that could be completed more easily and quickly as the basis for publishing a preliminary report.

The panel reluctantly concludes that SRS should stick with current practice. Each of these alternatives has pros and cons. Incomplete data might be misinterpreted, and preliminary totals would necessarily be on the low side because of missing data. Imputation for late respondents would introduce a new source of error and could make the data less accurate. Some data users hold that the potential for increased error due to omission or imputation is less desirable than more timely publication. Developing a new and simpler schedule to elicit early reporting would be a major new activity for SRS. In consideration of these issues, the best course of action at this time would be to continue the current policy of delaying publication until all reports are received, aggressively pursue better relationships with the agencies to encourage more timely response, and devote scarce resources to the improvements the panel recommends.

Recommendation 3-4: The Division of Science Resources Statistics (SRS) should maintain its current approach to data reporting, which is to wait for receipt of reports from all respondents before publishing the data. SRS should continue to report complete data without imputation for missing reports and data elements. The agency should focus on working directly with respondents to find ways to improve the timeliness of their response to the surveys.

---

[3]Presentation of John Jankowski, SRS, the Workshop on Modernizing the Infrastructure of the National Science Foundation Federal Funds Survey, September 5, 2008.

SURVEY TECHNOLOGY

Both the federal funds survey and the federal support survey have been early adopters of a web-based reporting system to facilitate response. However, not all respondents use the web to report, and the current web forms do not tap the full potential of current web-based survey methods. For example, at present the focus is on filling in the blanks in the reporting instrument. Little attention has been paid to developing an online survey instrument that is user-friendly and reduces respondent burden by using skip patterns and automatically populating responses. Some possibilities for process improvement include the automatic entry of zero values to subquestions when a response to one broad question is "no." For example, if the agency respondent reports that it does not support research in non-profit institutions, the subsequent subquestions should automatically be populated with zero values, instead of requiring the respondent to enter a zero for each item.

At the same time, SRS could consider tailoring the web survey for each agency based on prior knowledge and prior reports and with extensive collaboration with the agency. One goal could be to include only those data items that could be expected to be included in the agency submissions with relative ease and accuracy and that would be familiar to the agencies. This tailoring of the collection of data does not have to be applied to all agencies at once; SRS could begin with the larger ones.

Recommendation 3-5: The Division of Science Resources Statistics should invest in creating more user-friendly web surveys, possibly tailored to each agency, to replace current web versions of the paper surveys.

4

Toward a New Data Reporting System Based on Administrative Records

As this report describes, the federal funds and federal support surveys conducted by the Division of Science Resources Statistics (SRS) of the National Science Foundation (NSF) have several weaknesses, including special challenges with regard to timeliness and quality.[1] These difficulties stem, in part, from the complexity of the surveys. The federal funds survey, for example, asks the responding federal agencies to enter data on outlays for research and development for three years (actual for two years prior, prior year preliminary, and current year preliminary) and obligations for the same three years by categories of research and development (R&D), selected fields, types of research performers, specific federally funded research and development center, country of foreign performer, and state by type of performer. Although SRS has done much over the past few years to modernize the entry and transmission of the data, the surveys are burdensome for reporting agencies and, for many, do not reflect the reality of their R&D spending.

There may be a better way to obtain data in the future. Statistical agencies across the federal government are building the capability to use data from administrative records maintained by program administration agencies. The individual records are typically applications or reports completed by individuals and institutions to meet mandated requirements to compete for

[1]As defined by the Office of Management and Budget, quality is an encompassing term, incorporating utility (usefulness), objectivity (relating to the accuracy, reliability and lack of bias of the data), and integrity (making sure the data are protected from unauthorized access or manipulation) (U.S. Office of Management and Budget, 2002).

awards or receive benefits, or for compliance, credit, tax, or other reporting. The use of administrative records on financial transactions, including grants, contracts, and other awards, is becoming a possible optional source of information on federal R&D spending, mainly as a result of new initiatives to improve data on federal spending across the government. These new initiatives promise to make administrative data much more accessible and to improve their quality. As these data sources are improved, they offer a way for SRS to improve the collection and dissemination of comprehensive and timely data on federal R&D spending.

This chapter discusses the potential and limitations of the use of administrative data for collecting and compiling federal R&D spending data. It discusses past and current efforts to use the data and summarizes some initiatives that could change the way that agencies account for and report R&D spending. The requirements for a successful database and the challenges facing the SRS in developing this new system are then highlighted. Finally, a general plan for implementing a new vision for the federal funds and federal support data is outlined that specifically recommends planning for the transition from an all-survey to an integrated survey–administrative record approach, using demonstration projects to test various aspects of a possible transition to a system at least partly based on administrative data.

THE ROLE OF ADMINISTRATIVE RECORDS AS A SOURCE OF STATISTICAL DATA

The use of administrative records to substitute for or enhance surveys has been a goal for the federal statistical system for several decades. Particularly with regard to micro-level data from business entities, we can point to a number of highly successful examples of the development of information from administrative record sources (National Research Council, 2007b). Indeed, the increased use of administrative records has been recognized and documented since the early 1980s (Federal Committee on Statistical Methodology, 1980). However, the same reports that describe in glowing terms the potential of administrative records to provide detailed information at minimal cost, with an associated reduction in response burden in order to supplement or replace surveys, usually take pains to provide cautionary discussion as well. That is because problems of quality, consistency, and access have often plagued attempts to use administrative records for statistical purposes.

Still, the practice of using administrative records appears to be advancing, significantly aided by advances in information technology. With increasing use of administrative records, there has been greater attention to timeliness and other aspects of the quality of the input data. Examples can be seen in recent initiatives to improve the federal government's administrative records on grants and contracts, reflected in upgrades to the Federal Procurement Data System—Next Generation and the Federal Assistance Award Data System, as well as emerging cross-agency compilations of records that are not yet complete but could become fully fledged databases, such as www.grants.gov. (Each of these government-wide administrative data sources is detailed in Appendix B.)

The E-Government Act of 2002 and the Federal Funding Accountability and Transparency Act of 2006 (FFATA, Public Law 109-282) have the objective of systematically improving the contract and grant databases maintained by the U.S. General Services Administration and the U.S. Census Bureau, as well as to standardize, enhance, and validate the R&D spending data that reside in those databases.

The Federal Funding Accountability and Transparency Act, in particular, is a wide-ranging federal law requiring the full disclosure of all organizations receiving federal funds. It provides legal backing for gaining more information about extramural federal funding, including R&D. The act requires the establishment of a single searchable website providing comprehensive information on all federal awards, to be populated by the Federal Procurement Data System, the Federal Assistance Award Data System, and www.grants.gov. The act also includes, in Section 2(b)1, a provision that the searchable website shall include for each federal award, "any other relevant information specified by the Office of Management and Budget." These improvements have the potential for reporting on federal R&D spending at the project level and associating fiscal year obligations with such attributes as performer, performing institution, and geographic location. These and other attributes could serve as the foundation for SRS data collection efforts, which are also tied, in part, to fiscal year obligations and implicitly require the aggregation of project-level data into agency-wide data.

Under these new laws, a supporting infrastructure is being developed across the government under the leadership of the Office of Management and Budget (OMB) that has the potential of improving the quality and timeliness of administrative records on government expenditures and, by doing so, to provide at least part of the data on R&D expenditures that are now extracted by means of surveys. Before these new initiatives, administrative records did not have the capacity to provide current or reliable information on federal R&D expenditures.

For R&D spending data, the use of administrative records was first tested in the mid-1990s in a project called RaDiUS (Research and Development in the United States), which was developed by the Critical Technologies Institute at RAND for the Office of Science and Technology Policy. Although this project was discontinued in 2006, it illustrates the possibility of developing a data system to collect, store, and disseminate information on R&D expenditures from agency source documents, supplemented by independent, expert judgment. The RaDiUS database captured detailed data on federal R&D from agency records for the 24 agencies with the largest R&D expenditures. It accumulated not only records from the agency systems but also information from the Federal Assistance Awards Data System (FAADS) (Fossum et al., 2000). Although discontinued, RaDiUS taught valuable lessons about the quality of the contract and grant databases and how to approach development of a comprehensive system of information based on administrative data (National Research Council, 2005b, pp. 112-113).

OUTLINE OF A NEW REPORTING SYSTEM BASED ON ADMINISTRATIVE DATA

If SRS could ensure that public administrative records contain the required information fields, are recorded at the project level, are accurate, and are relatively easy to access, it could be confident in developing programs to collect and process administrative data instead of relying on surveys of agencies. This is a big order, and it does not describe the state of the various agencies' administrative records at this time. The challenge for SRS is to ensure that the administrative databases include all relevant research spending (to include intramural spending); have records that are accurate at the source, perhaps as entered by a knowledgeable person, such as a principal investigator; and include all relevant classification variables, particularly field of science and engineering (S&E) and character of work. Moving from the current situation to one in which administrative data can be fully used for purposes of understanding federal R&D spending will not be simple, nor can it occur soon. It will require the development of means for ensuring accuracy, completeness, consistency, and compatibility with the analytical needs now fulfilled by the federal funds and federal support surveys.

Accurate and Complete Data

An administrative record–based data collection system will be of use to SRS only if those records include information in the categories SRS needs to collect. To fully portray R&D spending, the data should be suitable for collection and aggregation, if needed, to recreate the current data for R&D versus R&D plant, the character of work (basic research, applied research and development), and field of S&E. In addition, information on area of application and the identities of recipients of funds is needed for a full understanding of the nature of the R&D investment. Generally, administrative data today fail to meet these requirements. As recently as 2005, for example, the Government Accountability Office reported that users of the

Federal Procurement Data System—Next Generation "lacked confidence in the system's ability to deliver timely and accurate data on contracts" (U.S. Government Accountability Office, 2005).

An important step toward improving the quality of administrative data on contracts and grants was taken with the publication and initial implementation of OMB guidance on data submission under the Federal Funding Accountability and Transparency Act. The guidance has the effect of increasing oversight of the data by establishing standards for a centralized system. This new administrative data system would retrieve data from selected systems in specified file formats, add data elements, specify requirements for timely reporting, and define quality assurance controls (U.S. Office of Management and Budget, 2006).

Consistency Over Time

Although the new system may ultimately offer more detailed and accurate data, it is critical for these new data to be comparable with the data from the current survey-based system. A major purpose of federal funds and federal support data is to enable analysis of trends in R&D spending, so that this spending can be connected with societal goals to help shape future patterns of spending in socially desirable ways. The ability to portray trends must be a feature of an administrative data-driven system for collecting the data necessary to report on R&D funding and expenditures. This ability would be especially critical during the time when shifting from the survey-based system to an administrative records-based system, and it could be enhanced by a period of dual publication of old and new data, widespread publicity, and full discussion of any apparent discrepancies.

Buy-in and Support

A successful administrative database depends on the support of those who are required to input, edit, manage, and evaluate the data. Although many of the current agency and government-wide databases are mandated by law, they rarely gain their success solely by virtue of such mandates. In addition, successful database systems are well understood by the stakeholders who can benefit from the data. For example, budgeting and fiscal accounting data systems are generally accurate and current because agencies have both incentives and requirements to keep them accurate and updated, and they perceive those data as being necessary for the success of their missions. The R&D contracts and grants in agency administrative record systems are beginning to receive that kind of attention, with a growing recognition that transparency is an important agency objective.

The current administration has strongly supported this transparency in regard to R&D spending, mandating in the guidance to the federal agencies that, in preparing their 2011 budgets, agencies "have a responsibility to explain how Federal science and technology investments contribute to increased economic productivity and progress, new energy technologies, improved health outcomes, and other national goals. In order to facilitate these efforts, Federal agencies, in cooperation with the Office of Science and Technology Policy and the Office of Management and Budget, should develop datasets to better document Federal science and technology investments and to make these data open to the public in accessible, useful formats" (Office of Management and Budget and Office of Science and Technology Policy, 2009, pp. 2-3).

ADVANTAGES AND DISADVANTAGES

Advantages of a System Based on Administrative Data

Administrative databases, if designed, managed, and implemented properly, would have some characteristics that can make them preferable to surveys as the source of information on R&D investments—although there would be challenges as well. This section discusses the accuracy, detail and flexibility, currency, and accessibility that are needed if administrative databases are to be useful for this purpose.

- *Accuracy*. The way in which an administrative database is constructed should serve to enhance the accuracy of the data obtained from it. In most cases, the data are entered by people who are in a position to know the topic. Principal investigators, budget officers, accountants, and project managers are more likely to know such information as the project's field of science than a person in an agency budget office who has been assigned responsibility to complete the SRS federal funds or federal support survey forms and who may not know or understand the intellectual and scientific nature of a project. However, some effort will be needed to ensure that the database is free of errors and misreporting that can be caused by careless or quick entries by individuals and institutions.

- *Detail and Flexibility.* A potential benefit of administrative databases is that they typically are composed of the original, raw form of information, such as data related to a single research project. In the current system, budget officers and others who complete the federal funds survey do so by aggregating a variety of inputs to represent the agency's R&D portfolio. Errors in compilation and aggregation can be made when the reports are being prepared. In contrast, project-level administrative information generally permits data users, including potentially the SRS staff, to produce tailored aggregations or link data from other administrative databases. Access to databases that offer data at the project level, with geographic and performer-level detail, enables ready aggregation of data to help answer the performer-type and performing-institution sections of the current surveys. Relying on administrative data instead of survey responses could enhance the ability to classify science in new and interesting ways.

- *Currency.* Administrative databases can be continuously updated. Under legislative guidance, agencies are now working to make the administrative data on grants and contracts more current, offering the possibility of dramatically shortening the time needed to obtain data on R&D spending and thus improving the timeliness of data. For example, the goals of the Federal Funding Accountability and Transparency Act legislation and guidelines are to ensure that grant and contract data are submitted within 3 days after the award and that the public database should be updated not later than 30 days after the award of any federal award requiring posting.

- *Accessibility.* The databases that are being developed under FFATA rules are being designed to ensure full transparency of all award actions by federal agencies in standard formats and thus would be more accessible to all users, including SRS, than the current survey databases. Moreover, because these administrative databases will be composed of an assortment of project-level data, rather than aggregations, users will be able to drill down or query according to their needs. These drill-down capabilities, using data mining and other sophisticated techniques, can greatly enhance the ways in which federal employees, budget experts, R&D specialists, and science and technology policy experts use these data for a variety of purposes. Thus, as it develops such R&D reporting systems, NSF should ensure that the public will have full access to them.

If NSF is to pursue a new system based on administrative records, there are a number of hurdles to clear. Foremost are the myriad technical requirements to create a workable database driven by administrative data. This will require buy-in and support from the reporting federal agencies. NSF will need full agency cooperation in order to modernize the gathering of accurate data in a more organized and timely fashion. The reporting agencies will have work to do in order to more adequately integrate the various internal systems that report R&D, so their reports to NSF can provide a complete picture of R&D throughout the federal government.

NSF does not need to face this task alone, however. The modernization and transparency of contract and award databases is a major federal government initiative, largely administered by OMB. The full weight and authority of OMB should establish an environment for improving administrative data and, through the administrative data, enhancing the transparency of federal spending.

Disadvantages of Administrative Databases

Currently, none of the contract and grant administrative databases discussed in this report provides the categories needed for direct reporting of federal funds or federal support data. These categories include R&D, R&D plant, character of work (basic research, applied research and development) and fields of S&E. In database management language, the administrative data systems do not currently contain the necessary "tags" (record descriptors) to permit extracting these sorts of data items.

The current administrative data systems are defined for agency-related administrative purposes, and not for the statistical purposes of the federal funds and federal support surveys. This could lead to a continuation of the current problems that affect the surveys, such as the fact that definitions of data items with the same name vary among agencies and even within agencies.

The problem of lack of coding for fields of S&E presents particular challenges. None of the contract and grant databases is organized in a way that would readily allow for the reporting of fields of S&E. For SRS to successfully transition from survey-based reporting of R&D activities to reporting based on administrative data, agency and government-wide databases should be able to associate each contract or grant funding record with descriptors of the work done under it, as described above.

It may be possible to obtain fields of S&E information without the burden that would be incurred if a relevant data field were to be added to each record. For example, it may be possible to construct cross-walks between agency-relevant keywords (tags) that are used in project descriptions and the fields of S&E taxonomy. These fields of S&E tags could be drawn from the taxonomy, or they could be based on free text in cases in which no existing tag fits, for example, for newly emerging areas. Field of S&E tags could be automatically derived from the name of the funding agency or program, or they could be provided by the funded entity by means of investigator-supplied keywords on project proposals and descriptions. Text mining techniques might be applied to extract key terms or to group semantically similar funding records to speed up manual determination and assignment of fields of S&E.

Recommendation 4-1: The Division of Science Resources Statistics, in cooperation with the Office of Management and Budget and the Office of Science and Technology Policy, should seek to have all federal agencies that fund or conduct research and development (R&D) to incorporate R&D descriptors (tags) into administrative databases. Ideally, in order to enable identification of the R&D components of agency or program budgets, tags should identify: the specific field of science and engineering; whether a record applies to R&D or R&D plant; and whether the record activity is basic research, applied research, or development.

Most agency contract and grant databases capture only extramural awards, with the notable exception of the Research, Condition, and Disease Categorization (RCDC) system of the National Institutes of Health (NIH), which explicitly captures intramural R&D (see Box 4-1). For existing databases to be useful for SRS's purposes, they would need to account for both extramural and intramural R&D. Intramural R&D is of particular importance to SRS reporting, since current data show that nearly one-quarter of federal R&D dollars are spent in intramural laboratories (National Science Board, 2006, p. 4-23). Intramural spending at the project, laboratory, or portfolio level will need to be incorporated into agency databases, perhaps following the approach used by NIH in populating the RCDC system with intramural as well as extramural project information, or it could be extracted directly from the parts of agencies that manage the intramural projects.

Recommendation 4-2: The Division of Science Resources Statistics should work with the Office of Management and Budget to seek endorsement to work with other research and development funding agencies to incorporate intramural data into existing and future databases or to directly access intramural spending information from performer databases.

Even after taking steps to identify R&D activities with some certainty and include both extramural and intramural projects, the thorny issue of accounting for classified R&D spending will remain. The spending on classified programs is an important part of R&D spending in some agencies, but these projects are not likely to be contained in administrative databases available to SRS or the public. This suggests implementation of a dual system based on administrative records for unclassified R&D supplemented by agency reporting of summary information for classified R&D.

BOX 4-1
The National Institutes of Health Research, Condition, and Disease Categorization System

The Research, Condition, and Disease Categorization (RCDC) system launched by the National Institutes of Health (NIH) is one of the promising administrative databases that could assist SRS in the transition from the current surveys to a new system for collecting federal funds and federal support data. The RCDC uses a computer database to sort NIH-funded projects into categories of research area, disease, or condition and allows these projects to be aggregated into annual reports on funding by category. The RCDC data are primarily used by Congress and the NIH Office of the Director to assess and evaluate NIH R&D spending priorities (Macro International, 2008, p. 58).

The RCDC is noteworthy because it replaces an annual survey of funding by category that was sent to the 27 units that constitute NIH. The annual survey had required respondents in each unit to estimate three years of funding for 360 research (e.g., clinical research, minority health, nanotechnology) and disease categories (e.g., Parkinson's, diabetes, cancer). NIH took the estimates from the 27 unit surveys to aggregate its total funding and spending. In many ways, the NIH annual survey closely resembles the current federal funds and federal support surveys because both methods asked separate units (or agencies) to report their spending behavior and then aggregated those self-reported data to obtain grand totals for R&D.

According to NIH, the RCDC will allow the same budget data to be extracted automatically from the database without relying on decentralized surveys, which increase the threat of data entry errors and inaccurate estimates. The RCDC will allow NIH to consistently report how its research dollars are spent. The RCDC has the added advantage of containing data on intramural research, which is absent from most administrative databases. Thus, the RCDC has the potential of capturing all NIH research, extramural and intramural. The potential for RCDC to enhance federal funds data reporting is magnified by the size of the NIH research portfolio; NIH alone is now responsible for supporting half of all federal basic and applied research.

The RCDC process involves creating category definitions for NIH, and currently there are 360 categories of which 215 are publicly reported in "Estimates of Funding for Various Diseases, Conditions, and Research Areas" (http://report.nih.gov/rcdc/categories). A category definition is a series of terms or concepts chosen from an RCDC thesaurus of more than 350,000 terms or concepts derived from various thesauri (from

continued

BOX 4-1 Continued

the Congressional Research Service, the National Cancer Institute, the Medical Subject Headings system of the National Library of Medicine, and Jablonski's Dictionary of Medical Acronyms and Abbreviations) and in conjunction with the Collexis text mining/matching tool. These terms are then weighted by scientific experts to identify the relative significance of each term or concept to the category. The same scientific experts set a threshold for each category to determine the minimum number of times a term or concept must be mentioned in a project description to make the project eligible for a specific category. Periodically, scientific experts validate these categories. The RCDC system can then search all funded grants and contracts in the NIH database to create a project summary containing terms and concepts that match the RCDC thesaurus; it then compares each summary with the category definitions to determine how closely they match. If the RCDC summary meets the threshold set by scientific experts for a category, RCDC assigns that project to that category, which makes it possible for RCDC to display not only a list of projects in each category but also funding amounts. The system creates, for the first time, NIH-wide category definitions that are consistent across all NIH institutes to solve the problem of different institutes using their own definitions to respond to the current survey—the same problem that affects the SRS surveys. The system allows the category definitions to be applied uniformly to all types of research.

The NIH unveiled the new RCDC system in 2009.

For more on RCDC methodology, see http://report.nih.gov/rcdc/category_process/default.aspx and http://report.nih.gov/rcdc/faqs/default.aspx.

IMPLEMENTATION OF A SYSTEM BASED ON ADMINISTRATIVE DATA

The augmentation of current federal government-wide initiatives to provide basic information to identify R&D spending is a promising avenue for SRS to consider as it moves toward a mixed survey and administrative database system. In order to develop a flexible, administrative data-driven system for tracking federal funds and federal support data, the panel recognizes that SRS needs adequate authority and resources. The plan we outline in Chapter 6 for implementing the transition from a survey-based system to a mixed system of surveys and administrative databases requires that SRS staff work closely with agencies, OMB, and other relevant stakeholders in ongoing efforts to further develop e-government and federal spending database capabilities. Cooperative efforts alone may not lead to the inclusion of key variables, such as R&D identifiers and fields of science, into current and future administrative databases. SRS requires adequate budget resources and the full support of NSF management and OMB to participate in ongoing efforts to build in capabilities for collecting R&D survey data from current and future databases.

The panel notes that it would be helpful to have congressional endorsement for a modernization of the federal funds and federal support program, even though no new legislative authority is required. NIH, for example, was assisted in building its RCDC system by an explicit requirement in the National Institutes of Health Reform Act of 2006 to build such a tool to categorize the agency's research (Section 104 of Public Law 109-482). Outside organizations can play an important role as well. For example, two reports from the National Academies are claimed to have assisted in laying the groundwork for the RCDC system.[2]

OMB is the lead executive branch agency for collecting, organizing, and providing information on federal spending, and recent legislation mandates much of this data collection. In issuing the guidance for data submission under the Federal Funding Accountability and Transparency Act, OMB has shown a willingness to use its authority under the legislation to specify data fields. If the OMB guidance were extended to mandate identification of R&D awards, R&D versus R&D plant, character of work and field of science, these databases would be much more useful for understanding federal R&D spending.

The FFATA databases already offer some promise in obtaining some of the data relevant to understanding R&D spending, although much work remains to make these data useful for SRS's purposes. For example, www.USAspending.gov, the portal for the public to access the FFATA databases, allows users to generate detailed reports on external federal spending by performing institution, performer type, and geographic location. However, the website does not enable users to generate reports of federal spending by character of work or field of S&E, and it lacks information on intramural R&D. Furthermore, it does not distinguish between spending on R&D and on R&D plant. However, these databases appear to be the only existing cross-government databases that can meet both FFATA requirements and, potentially, SRS's data needs.

TRANSITION STRATEGY

The new vision for the federal funds and federal support data outlined in this chapter will not be implemented overnight. Many of the precondi-

[2] Available: http://report.nih.gov/rcdc/faqs/Default.aspx.

tions for a successful conversion of the program from a survey-only to an integrated survey-administrative record approach are not yet in place. For example, the FFATA-enhanced administrative databases on contracts and grants are still maturing, and little work has yet been done with the major reporting agencies to set the basis for direct SRS exploitation of their administrative records.

Several initiatives in the short term, however, would position SRS to effectively seize the moment when the preconditions for conversion of the program are in place. One approach would be to set up a series of demonstration projects to help determine good ways to transition to a system based at least in part on administrative data.

The initial demonstration projects could be based on lessons learned by NIH in developing the RCDC system. With selected large reporting agencies, SRS could explore what would be necessary to develop agency-appropriate approaches to a more comprehensive system—in one set of demonstrations, using the current agency administrative databases to test mining for terms that could yield field of S&E taxonomic elements and, in another, perhaps testing the development of cross-walks between program/projects and fields.

Such demonstration projects, conducted by the reporting agencies in conjunction with the implementation of government-wide administrative record improvement programs (and, one hopes, partially funded by those initiatives) could help illuminate the way to identify fields of S&E in data records at the program and project level, using the text-based technologies described in Chapter 5.

Recommendation 4-3: The Division of Science Resources Statistics should initiate work with other federal agencies to develop several demonstration projects to test for the best methods to move to a system based at least partly on administrative records.

5

Toward a Comprehensive Assessment of Federal Research and Development Investment

The preceding chapters have defined the need for this study, identified user issues, discussed methods for improving the current system in the short term, and laid out a plan for the design and implementation of a new system. This chapter begins by considering the future of collection of federal research and development (R&D) spending data in the context of a "science of science" analysis framework, which is evolving to address the needs of science policy analysts and decision makers for R&D spending information. These needs were identified in the workshop the panel held as part of its data-gathering activities (see Appendix C). Many of these needs require bringing together federal funding information from across the federal government, as is now the task of the federal funds and federal support surveys. In addition, they require data on R&D spending outcomes, the role of federal R&D spending in fostering innovation, and the "capacity of the science enterprise to contribute to the wide array of social goals that justifies society's investment in science" (Sarewitz, 2007, p. 1).

This chapter also considers novel, cutting-edge approaches, such as data federation, automatic text, and linkage analysis that could help enrich the information attainable from administrative sources. This chapter suggests several medium- and long-term initiatives that the Division of Science Resources Statistics (SRS) of the National Science Foundation (NSF) should consider as a basis for modernizing the federal R&D data system.

SCIENCE OF SCIENCE METRICS

The needs and opportunities for improved data and methods for making and analyzing science and technology policies are the principal focus of the NSF Science of Science and Innovation Policy (SciSIP) Program. SciSIP aims to foster the development of relevant knowledge, theories, data, tools, and human capital. 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. 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).

Metrics from such sources as the federal funds and federal support surveys are essential inputs to such analysis. To further develop the vision of a new science of science policy, which was originally articulated by John H. Marburger III, the former director of the Office of Science and Technology Policy and presidential science adviser, the National Science and Technology Council Interagency Task Group developed "The Science of Science Policy: A Federal Research Roadmap" (National Science and Technology Council and Office of Science and Technology Policy, 2008). The roadmap document points out the importance of public investments in science, technology, and innovation but notes that a rationale for scientific investment decisions has insufficient theoretical and empirical bases. The roadmap calls for the development of more rigorous tools, methods, and data to help arrive at sound and cost-effective investment strategies.

The portfolio of statistics prepared by SRS is central to these science and innovation policy initiatives. However, the development of an infrastructure for the science of science and innovation policy cannot be accomplished by SRS alone. It will require contributions from academic research and from a multitude of other federal agencies and departments. The workshop summary in Appendix C describes the kind of information that users need to support an assessment of science and innovation policy. In addition to the current input indicators that are offered by the federal funding data, these needs include output indicators (e.g., publications, graduate students, citations, patents) that support "return on investment" studies and other science policy analyses. To best take advantage of the dynamic nature of these investments, the data would need to be retrievable in new ways. Ideally, it should be possible to select a bar graph, a geospatial region, or a field of S&E, and to delve deeply into the data to see the specific projects funded by different agencies or to see the papers, patents, or products that resulted.

A new data system to support monitoring and analyzing science and innovation policy would require the ability to bring together data at the contract, project, program, and activity levels. The data should be supported by standards of transparency, accountability, and comparability, as well as an infrastructure that permits the linking of data records across agencies.

A FEDERATED SYSTEM OF SCIENCE INVESTMENT POLICY-RELEVANT DATA

This federated database of the future would require that major units of analysis (projects) be supported by unique and persistent data identifiers. The unique identifiers for major projects would include not only grants and contracts, but also papers, patents, people, authors, institutions, and countries, as well as geographic locations, R&D or R&D plant, character of work, and field of science. Plus, records will have to be interlinked—authors would be linked to their respective institutions as well as to all of their papers and funding; papers would be linked to other citing or cited papers; and contracts, grants, and papers would be linked to the fields of science that they represent. This would require the development of a data federation system that would be able to link data records across agency boundaries and would be supported by tools that enable searching and integrating very large databases.

The federated system would have new ways of accessing the data. A management system for the persistent identification of the content of digital networks has been developed by a group of international registration agencies in the Digital Object Identifier (DOI) System. The DOI identifiers, or names, are unique and interoperable from system to system. The DOI names are strings of information about an object backed up by descriptive metadata (Paskin, 2009).

The federated system would incorporate the Semantic Web (Berners-Lee, Hendler, and Lassila, 2001), which is based on standards developed by the World Wide Web Consortium (W3C). The Semantic Web uses uniform resource identifiers (URIs) as globally unique, persistent identifiers. URIs are defined in the resource description framework (RDF[1]), as is the representation of URI relationships and attributes, using "triples." A triple describes the data in a sentence-like fashion. Some common links are *rdf:seeAlso* to refer to related records, *owl:sameAs* to indicate record identity, or *foaf:knows* to interlink records of people who know each other.

[1]Available: http://www.w3.org/RDF/.

Sets of triples can be stored in a single file or distributed across the entire web. SPARQL (Sparql Protocol and RDF Query Language),[2] another W3C standard, makes it possible to query Semantic Web data using SQL-like syntax. RDF relationships can also be embedded into standard web pages using RDFA (Resource Description Framework in Attributes).[3] In this way, browsers or search engines can extract structured data.

Semantic Web standards such as RDF Schema (RDFS)[4] and the Web Ontology Language (OWL)[5] make it possible to exchange ontologies, which specify the semantics of the terminology and relationships used in RDF descriptions. Ontologies also enable reasoning, or inference of new triples based on existing data. Another W3C language that should be considered is the Simple Knowledge Organization System (SKOS). This language is particularly tailored to the development of taxonomies and thesauri, being based on a KOS thesaurus standard. It is now being used by many of the U.S. national libraries, including the Library of Congress. SKOS may be especially appropriate for some of the administrative data reporting issues faced in developing a modernized retrieval system for federal grant and contract actions.[6]

eXtensible Markup Language (XML) is another tool that adds structure, although it does not add meaning to records. It permits data providers and users to create their own "tags" or labels that annotate web pages or content on web pages. In the case of R&D investment data, tags could be developed through XML technology to locate words or phrases that identify a contract or grant as pertaining to a class of research or development, as well as other information, such as a field associated with the object identified as research and development.

The Linking Open Data (LOD) community project led by the World Wide Web Consortium Semantic Web Education and Outreach Group (W3C SWEO)[7] shows the power of exposing, sharing, and connecting data via dereferenceable uniform resource identifiers (URIs) on the web.[8] Data sets federated via LOD vary from Wikipedia data (DBPedia, wikicompany), to geographical data (e.g., Geonames, World Factbook, Eurostat), to news data (e.g., BBC News), and recently to government data (e.g., U.S. Census Data, GovTrack). The number of interlinked data sets is growing rapidly—from over 500 million RDF triples in May 2007 to around 20 billion RDF

[2] Available: http://www.w3.org/TR/rdf-sparql-query/.
[3] Available: http://www.w3.org/TR/xhtml-rdfa-primer/.
[4] Available: http://www.w3.org/TR/rdf-schema/.
[5] Available: http://www.w3.org/TR/owl-features/.
[6] Available: http://www.w3.org/2004/02/skos/.
[7] Available: http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.
[8] URIs are used in the World Wide Web to identify resources. Using a standardized protocol such as HTTP, they can "dereferenced" to obtain information about the resource.

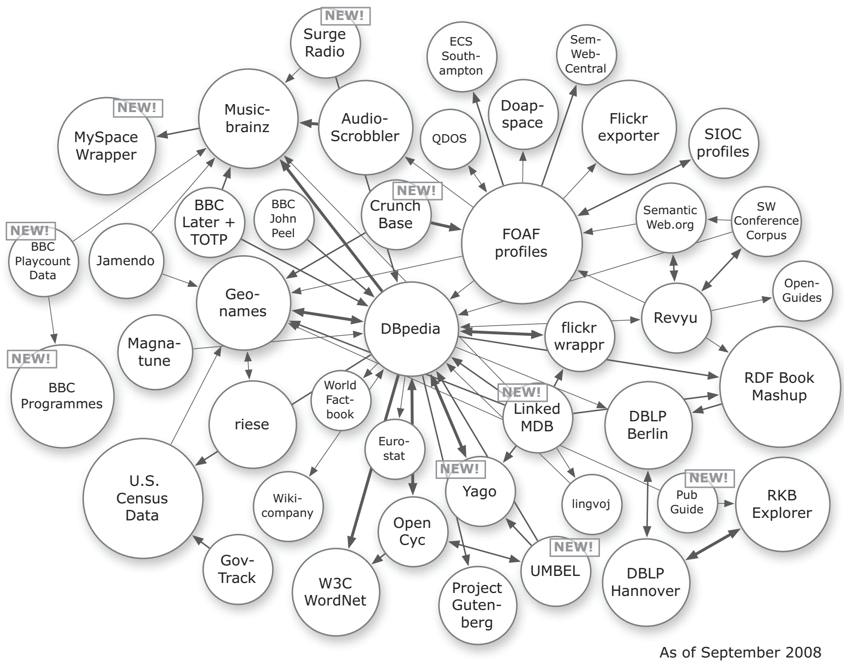

FIGURE 5-1 Linking open data sources.
SOURCE: Linking Open Data W3C SWEO community project. Available: http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

triples in September 2008. Figure 5-1 illustrates the diversity and number of data sets federated by the LOD project.

Several of the government data sets that have already been converted to RDF format and interlinked within the LOD project are relevant for monitoring the health of S&E, such as U.S. Census Data, GovTrack, Eurostat, and the World Factbook. Other existing data sets, such as SRS funding data, the www.data.gov, and the www.USAspending.gov files, have already been or could be easily converted into RDF format and interlinked to other LOD data sets via RDF links, improving the coverage and utility of the data for analysis of R&D investments.

One important aspect of this work is that about 80-90 percent of the required data unification and data interlinkage can be done automatically. The remaining 10-20 percent of data correction needs to be done by the creators of the objects—the funding agencies that changed award-naming conventions or the principal investigators who best know the subject matter of their grants and contracts. Hence, it will be necessary to support and encourage data acquisition, editing, and annotation by the applicants and recipients of the grants and contracts and the agencies that fund them. A system used in the Air Force Research Laboratories, using JIFFY R&D program management software, has provided such an automated tool for managing program execution. Importantly, the JIFFY system provides a common repository for data elements related to several different functions of the laboratories.[9]

A major data federation project that assembles data on academic research and researchers in the United States and other countries is the Community of Science (COS), a product of ProQuest. The COS Scholar Universe has access to information on more than 2 million researchers in over 200 disciplines and 9 countries. The information about scholars is linked to their publications in other databases.[10]

An international example is the free admission Lattes Database compiled and served by the National Council for Scientific and Technological Development in Brazil.[11] The site provides access to around 1,100,000 researcher curricula and about 4,000 institutions in Brazil, including education, business, nonprofit private, and government organizations. Researchers in Brazil were asked to log in to Lattes to ensure that their data are complete and correct with the incentive that these data would soon be used in funding decision making. The result is acclaimed to be one of the cleanest researcher databases in existence today.

The Lattes Database was further interlinked with data from other institutions, such as SciELO, LILACS, SCOPUS, Crossref, and university databases, to increase its coverage and quality. Many institutions in Brazil use the Lattes Database to retrieve data about their teachers, researchers, students, and employees. They interlink the data with their own information systems, generating internal indicators of scientific and technological production or using the data in support of the implementation of management policies.[12]

A number of other countries have created their own databases. Among them are the Directory Database of Research and Development Activities (ReaD) in Japan[13] and the Italian Network for Innovation and Technology Transfer to Small and Medium Sized Enterprises (RIDITT).[14] ReaD is a database service designed to promote cooperation among industry, academia, and government by collecting and providing scientific information on research institutes, researchers, research projects, and research resources

[9]Available: http://www.stormingmedia.us/32/3280/A328024.html.
[10]Available: http://www.refworks-cos.com/GlobalTemplates/RefworksCos/cosschuniv.shtml.
[11]Available: http://lattes.cnpq.br/english.
[12]Available: http://lattes.cnpq.br/english/conteudo/acordos.htm.
[13]Available: http://read.jst.go.jp/index_e.html.
[14]Available: http://www.riditt.it/page.asp?page=faq&action=detail&IDObject=122.

in Japan. RIDITT is an initiative aimed at improving the competitiveness of small- and medium-sized enterprises by strengthening the supply of services for innovation and technology transfer and the creation of new high-tech enterprises. It is promoted by the Italian Ministry for Economic Development and managed by the Italian Institute for Industrial Promotion.

If the Office of Management and Budget (OMB) were to oversee implementation of a similar system, over time, the federated system of science and innovation policy–relevant data could encompass a majority of U.S. scholarly data. Given the value and importance of these data, the system should not be owned by a private entity but should be developed using federal funds and hosted by a governmental institution, such as the National Library of Medicine or the National Institute of Standards and Technology. If such a system existed, the job of SRS to provide innovative data on S&E would be significantly enhanced.

Obtaining Fields of Science and Engineering Information in the Future

Many of the currently used science taxonomies are manually compiled for specific domains of science. Examples are the National Library of Medicine's Medical Subject Headings (MeSH) thesaurus, the Computer Retrieval of Information on Scientific Projects (CRISP) thesaurus, and the National Cancer Institute's thesaurus. As the amount, complexity, and diversity of relevant data grow, it becomes more and more difficult to ensure that manually compiled structures truly match the stream of data they aim to organize.

Text analysis techniques, such as the Topic Model by Griffith and Steyvers (2004), can be applied here. The techniques read a large volume of text, for example, all NIH awards for a certain year, and set a parameter that states the number of desired topics (typically around 500). An algorithm then compiles a list of unique words that occur in the award texts. Using Latent Dirichlet Allocation by Blei, Ng, and Jordan (2003), a topic model then computes and outputs two probability matrices: "awards × topics" and "unique words × topics." The topic model has been successfully applied to all 2007 NIH awards and to data sets as large as Medline (about 18 million papers).[15]

Recent work on mapping knowledge domains (Börner, Chen, and Boyack, 2003; Shiffrin and Börner, 2004) uses citation links to study and communicate the structure and dynamics of science at the local and global levels. Wagner and Leydesdorff have used new tools emerging from network science to better understand international collaborations at the subfield level (Wagner and Leydesdorff, 2005). Klavans and Boyack recently

[15] The 2007 mapping is shown as a visual browser at http://scimaps.org/maps/nih/2007.

compared 20 existing maps of science (Klavans and Boyack, 2009). Some of the maps were compiled by hand, others automatically, using very different data sets and approaches. They had three basic visual forms: hierarchical, centric, and noncentric (or circular). The authors found that a circular "consensus map" generated from consensus edges occurs in at least half of the input maps. The ordering of areas in the consensus map is as follows: Mathematics is (arbitrarily) placed at the top of the circle, followed clockwise by physics, physical chemistry, engineering, chemistry, earth sciences, biology, biochemistry, infectious diseases, medicine, health services, brain research, psychology, humanities, social sciences, and computer science. The link between computer science and mathematics completes the circle. If the lowest weighted edges are pruned from this consensus circular map, the result is a hierarchical map stretching from mathematics to social sciences. This result is valuable, as it supports the argument that a general structure of science can be derived from very different data sources using different approaches.

6

The Path Ahead

The times are right for the federal research and development (R&D) statistics programs at the National Science Foundation (NSF). There is growing interest in building metrics to help science policy decision makers understand the role of government R&D spending in scientific advances and economic growth, and there are likewise strong forces propelling government agencies to enhance the content and accessibility of their administrative databases. Because of these forces, NSF has a prime opportunity to modernize its system of obtaining information on federal spending for research and development. And a third force is also at play. Pioneering work in the academic and computing worlds has led to protocols and solutions that promise to enable the federal government to structure, manage, and extract the massive amount of data residing in agency databases from which information on R&D spending is drawn. These conditions open new possibilities for the short- and long-term future, as this report discusses, and can, with effort, time, and resources, lead to a revolution in the way in which NSF obtains and presents the federal spending data.

Decisive action is required to capture these opportunities. One imperative is to obtain the cooperation of the Office of Management and Budget (OMB)—the agency that is driving the movement toward improving the databases. Laying a new foundation of trust and cooperation with the agencies that now report on their R&D spending to NSF is another imperative. The development of a strategic, long-term approach is yet another.

Simply trying to do things in the same way but better will not yield the necessary major long-term improvements in data quality when dealing with the myriad of relatively intractable agency management and reporting systems. NSF at present lacks the resources to significantly affect the way in which agencies manage and report their expenditures. Therefore, as NSF undertakes to work with agencies to make incremental improvements in reporting the survey information and continues to upgrade and to simplify its collection mechanisms as recommended in this report, the agency is advised to devote attention and resources to harnessing public policy interest in improving the reporting of federal spending, including R&D expenditures.

The E-Government Act and the Federal Funding Accountability and Transparency Act can be harnessed to provide powerful support in this effort, enabling NSF to advance in a manner that imposes little or no additional reporting burden on the reporting agencies. The strategic process that will lead to modernized collection should begin with some modest steps to shore up the current system while laying the basis for far-reaching changes later. To assist in visualizing this process, the panel suggests some broad, time-phased actions that, in essence, lay out the pathway to a modernized federal R&D spending data system (see Figure 6-1).

To move the collection of data on federal R&D spending from today's survey-centric model to a modernized system could be accomplished in a series of overlapping steps. It is important, for the reliability and credibility of the current system, to make as many of the modest improvements suggested in Chapter 3 as early as possible. These changes will not only shore up the current system, but also will set the stage for coming major improvements.

At the same time as these modest improvements are under way, NSF should begin more systematic coordination with OMB, which has the responsibility for enhancing the quality and accessibility of agency contract and award data, and the Office of Science and Technology Policy, which has a key responsibility for assessing the contribution of federal R&D spending to economic growth and innovation. The immediate objective of this coordinated effort is to initiate a process that will lead to additional data items being incorporated into the administrative databases.

Taking lessons from the Research, Condition, and Disease Categorization (RCDC) system of the National Institutes of Health (NIH), NSF could take steps to promulgate agency-appropriate approaches to a more comprehensive system, one that includes taxonomic elements and permits cross-walks between program/projects and fields. It should be a priority for NSF to find ways to identify fields of science and engineering in data records at the program and project levels, perhaps through use of text-based technologies as described in Chapter 5. Finally, in graduated measure, NSF should develop the capacity for mining the standard and newly enriched government-wide contracts and awards databases to extract comprehensive information on R&D spending.

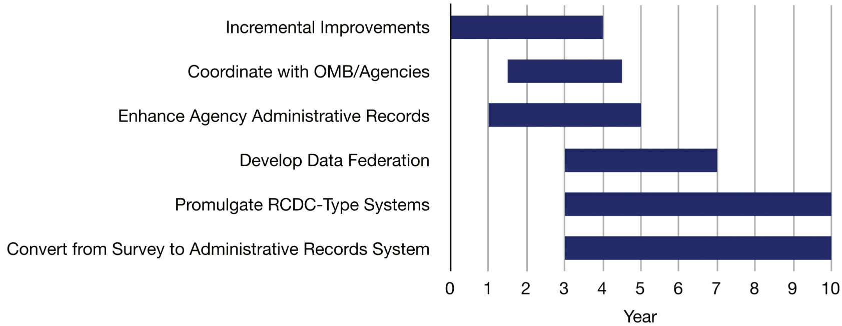

FIGURE 6-1 Pathway to a modernized federal R&D spending data system.

Bumps are expected along the path. For example, agencies may resist taking transparency down to the project level, where it could eventually be used to judge the worth of individual projects. Thus, an implementation strategy needs to incorporate a way to communicate and discuss the benefits of proposed solutions to the reporting agencies, to policy makers, and to the public.

The improvements recommended in this report will not be easy to implement. They would constitute a major coordination and information technology initiative. But the development of a cross-agency data federation and analysis framework building on administrative records and advanced retrieval technologies has great promise. As is becoming apparent in work going on in pioneering agencies, such as with the NIH RCDC project, and in the academic community where analysis has gained from advanced retrieval processes, there will be demonstrable gains in the efficiency, transparency, and analytical capacity for collecting and assessing federal investments in research and development.

References

Bailar, B. (2004). A quality profile of the federal research and development surveys. In National Research Council, *Measuring Research and Development Expenditures in the U.S. Economy: Interim Report* (Appendix A, pp. 24-67). L. Brown and T. Plewes, Eds. Panel on Research and Development Statistics at the National Science Foundation, Committee on National Statistics. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Berners-Lee, T., J. Hendler, and O. Lassila (2001). The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, May 17.

Blei, D.M, A.Y. Ng, and M.I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993-1022.

Block, F., and M.R. Keller (2008). *Where Do Innovations Come From? Transformations in the U.S. National Innovation System 1970-2006*. Washington, DC: Information Technology and Innovation Foundation. Available: http://www.itif.org/files/Where_do_innovations_come_from.pdf [accessed November 2009].

Börner, K., C. Chen, and K. Boyack (2003). Visualizing knowledge domains. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (vol. 37, pp. 179-255). Medford, NJ: American Society for Information Science and Technology.

Branscomb, L., and P.E. Auerswald (2001). *Taking Technical Risks: How Innovators, Executives, and Investors Manage High-Tech Risks*. Cambridge, MA: MIT Press.

Branscomb, L., and P.E. Auerswald (2002). *Between Invention and Innovation: An Analysis of Funding for Early-Stage Technology Development*. NIST CR 02-841, November. Gaithersburg, MD: National Institute for Standards and Technology. Available: http://www.atp.nist.gov/eao/gcr02-841/contents.htm [accessed November 2009].

Broad, W.J. (1997). Study finds public science is pillar of industry. *New York Times*, May 13.

Bush, V. (1945). *Science The Endless Frontier, A Report to the President*. Washington, DC: U.S. Government Printing Office. Available: http://www.nsf.gov/od/lpa/nsf50/vbush1945.htm [accessed November 2009].

Cheney, D., and J. Park (2005). *Task Order 07: Panel Meeting to Provide Guidance on SRS's Evaluation of Its Science and Engineering Taxonomies. Final Summary Report.* Arlington, VA: SRI International.

Congressional Research Service (2000). *Challenges in Collecting and Reporting Federal Research and Development Data.* M. Davey and R. Rowberg, Eds. Washington, DC: Congressional Research Service.

Executive Office of the President (2005). FY 2009 Administration Research and Development Priorities. Memorandum for the Heads of Executive Departments and Agencies. August 14. Office of Management and Budget and Office of Science and Technology Policy, Washington, DC.

Federal Committee on Statistical Methodology (1980). *Report on Statistical Uses of Administrative Records.* Working Paper No. 6, December. Washington, DC: U.S. Department of Commerce.

Fossum, D., L. Painter, V. Williams, A. Yezril, E. Newton, and D. Trinkle (2000). *Discovery and Innovation: Federal Research and Development Activities in the Fifty States, District of Columbia, and Puerto Rico.* Santa Monica, CA: Rand.

Godin, B. (2005). *The Linear Model of Innovation: The Historical Construction of an Analytical Framework. Project on the History and Sociology of S&T Statistics.* Working Paper No. 30. Available: http://www.csiic.ca/PDF/Godin_30.pdf [accessed November 2009].

Griffiths, T.L., and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(Suppl. 1):5228-5235.

Hjørland, B. (2008). Core classification theory: A reply to Szostak. *Journal of Documentation,* 64(3):333-342.

Klavans, R., and K. Boyack (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology* 60(3):455-476.

Klein, J.T. (1996). *Crossing Boundaries: Knowledge, Disciplinarities, and Interdisciplinarities.* Charlottesville: University Press of Virginia.

Klein, J.T. (2009). *The Taxonomy of Interdisciplinarity.* Available: http://www.ndsciencehumanitiespolicy.org/oup2/itoc/KleinTaxonomy.pdf [accessed November 2009].

Lattuca, L. (2001). *Creating Interdisciplinarity: Interdisciplinary Research and Teaching Among College and University Faculty.* Nashville, TN: Vanderbilt University Press.

Lenoir, T., and C. Beghtol (2004). *Guiding Principles for Classification in Practice.* Presentation at panel meeting to provide guidance on SRS's evaluation of its field of science and engineering taxonomies, National Science Foundation, October 12, Arlington, VA.

Leshner, A. (2004). Science at the leading edge. *Science* 303(5659):729.

Macro International (2008). *Fields of Science and Engineering Taxonomy Study.* Calverton, MD: Macro International.

Marburger, J. (2005). Wanted: Better benchmarks. *Science* 308(5725, May):1087.

Mervis, J. (1997). Databases: RaDiUS draws a bead on U.S. R&D. *Science* 278(5342):1392.

National Institutes of Health (2009). *NIH Roadmap for Medical Research.* Available: http://nihroadmap.nih.gov/ [accessed November 2009].

National Research Council (1986a). *An Overview: Physics Through the 1990's.* Physics Survey Committee, Board on Physics and Astronomy. Washington, DC: National Academy Press.

National Research Council (1986b). *Scientific Interfaces and Technological Applications.* Physics Survey Committee, Board on Physics and Astronomy. Washington, DC: National Academy Press.

National Research Council (1990). *Interdisciplinary Research: Promoting Collaboration Between the Life Sciences and Medicine and the Physical Sciences and Engineering.* Washington, DC: National Academy Press.

National Research Council (1995). *Allocating Federal Funds for Science and Technology.* Washington, DC: National Academy Press.

National Research Council (1999). *Funding a Revolution: Government Support for Computing Research*. Science and Telecommunications Board. Washington, DC: National Academy Press.

National Research Council (2000). *Measuring the Science and Engineering Enterprise: Priorities for the Division of Science Resources Studies*. Committee to Assess the Portfolio of the Division of Science Resources Studies of NSF. Washington, DC: National Academy Press.

National Research Council (2003). *Assessing Research-Doctorate Programs: A Methodology Study*. Policy and Global Affairs Division. Washington, DC: The National Academies Press.

National Research Council (2005a). *Facilitating Interdisciplinary Research*. Committee on Facilitating Interdisciplinary Research, Committee on Science, Engineering, and Public Policy. Washington, DC: The National Academies Press.

National Research Council (2005b). *Measuring Research and Development Expenditures in the U.S. Economy*. Panel on Research and Development Statistics at the National Science Foundation, Committee on National Statistics. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council (2006). *Taxonomy of Fields and Their Subfields*. Research Doctorate Programs, Board on Higher Education and Workforce. Available: http://sites.nationalacademies.org/PGA/Resdoc/PGA_044522 [accessed November 2009].

National Research Council (2007a). *The National Science Foundation's Materials Research Science and Engineering Centers Program: Looking Back, Moving Forward*. Washington, DC: The National Academies Press.

National Research Council (2007b). *Understanding Business Dynamics*. Panel on Measuring Business Formation, Dynamics, and Performance, Committee on National Statistics. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Science Board (2001). *Federal Research Resources: A Process for Setting Priorities.* NSB 01-156. Arlington, VA: National Science Foundation.

National Science Board (2006). *Science and Engineering Indicators 2006*. (Volume 1, NSB 06-01; Volume 2, NSB 06-02). Arlington, VA: National Science Foundation.

National Science Foundation (1999). *Report on the NSF Agency Workshop on Federal R&D*. Arlington, VA: National Science Foundation.

National Science Foundation (2008a). Federal R&D funding down in FY 2007. *InfoBrief*, Division of Science Resources Statistics, February.

National Science Foundation (2008b). *Federal S&E Obligations to Academic Institutions Reach New Highs in FY 2006 but Fail to Keep Up with Inflation.* NSF 08-316, Division of Science Resources Statistics, October. Arlington, VA: National Science Foundation.

National Science Foundation (2008c). *Science of Science and Innovation Policy, FY 2009 Program Solicitation*. NSF 08-586. Arlington, VA: National Science Foundation. Available: http://www.nsf.gov/pubs/2008/nsf08586/nsf08586.htm [accessed November 2009].

National Science Foundation (2009a). FY 2008 data show downward trend in federal R&D funding. *Info Brief*, Table 1, NSF 09-309, January.

National Science Foundation (2009b). *Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions: FY 2006*. NSF 09-210, March. Arlington, VA: National Science Foundation.

National Science and Technology Council and Office of Science and Technology Policy (2008). *The Science of Science Policy: A Federal Research Roadmap*. Report on the Science of Science Policy to the Subcommittee on Social, Behavioral and Economic Sciences, Committee on Science. November. Available: http://www.ostp.gov/galleries/NSTC%20Reports/39924_PDF%20Proof.pdf [accessed November 2009].

Office of Science and Technology Policy (2006). *The Science of Science Policy: A Federal Research Roadmap*. Washington, DC: National Science and Technology Council.

Organisation for Economic Co-operation and Development (2002). *Frascati Manual 2002: Proposed Standard Practice for Surveys on Research and Development*. Paris: Organisation for Economic Co-operation and Development.

Paskin, N. (2009). Digital Object Identifier system. *Encyclopedia of Library and Information Sciences*. London: Taylor & Francis Group.

Petrie, H.G. (1976). Do you see what I see? The epistemology of interdisciplinary inquiry. *Educational Researcher* 5(2):9-15.

Porter, A.L., J.D. Roessner, A.S. Cohen, and M. Perreault (2006). Interdisciplinary research: Meaning, metrics, nurture. *Research Evaluation* 15(3):187-195.

Rhoten, D. (2003). *Final Report, National Science Foundation BCS-0129573: A Multi-Method Analysis of the Social and Technical Conditions for Interdisciplinary Collaboration*. San Francisco: Hybrid Vigor Institute.

Rossini, F.A., and A.L. Porter (1978). Frameworks for integrating interdisciplinary research. *Research Policy* 8:70-79.

Ruttan, V. (2006). *Is War Necessary for Economic Growth? Military Procurement and Technology Development*. New York: Oxford University Press.

Sarewitz, D. (2007). Does science policy matter? *Issues in Science and Technology*, Summer. Available: http://findarticles.com/p/articles/mi_qa3622/is_200707/ai_n19511079/pg_6/?tag=content;col1 [accessed November 2009].

Shiffrin, R., and K. Börner (2004). Mapping knowledge domains. *Proceedings of the National Academy of Science* 101(Suppl 1, April).

Smelser, N. (2003). Interdisciplinarity in theory and practice. In C. Camic and H. Joas (Eds.), *The Dialogical Turn: New Roles for Sociology in the Postdisciplinary Age*. Lanham, MD: Rowman and Littlefield.

Solow, R. (1987). *Growth theory: An exposition*, 2nd ed. New York: Oxford University Press.

Stokes, D.E. (1997). *Pasteur's Quadrant: Basic Science and Technological Innovation.* Washington, DC: Brookings Institution Press.

Touhy. R.V. (1998). How the Department of Defense utilizes federal funds survey data. In National Science Foundation, *Report on the NSF Agency Workshop on Federal R&D*. Washington, DC: National Science Foundation.

United Nations Educational, Scientific, and Cultural Organization (1979). Recommendation concerning the international standardization of statistics on science and technology. In *Records of the General Conference, 20th Session, Paris, 24 October to 28 November 1978, Volume 1* (pp. 27-28). Paris: UNESCO. Available: http://portal.unesco.org/en/ev.php-URL_ID=13135&URL_DO=DO_TOPIC&URL_SECTION=201.html [accessed September 2009].

U.S. Department of Commerce (2008). *Innovation Measurement—Tracking the State of Innovation in the American Economy.* A report to the Secretary of Commerce by the Advisory Committee on Measuring Innovation in the 21st Century Economy. Available: http://www.kauffman.org/uploadedFiles/innovation_measurement_11808.pdf [accessed November 2009].

U.S. Government Accountability Office (2005). *Improvements Needed to the Federal Procurement Data System—Next Generation*. GAO-05-960R, September 27. Available: http://www.gao.gov/new.items/d05960r.pdf [accessed November 2009].

U.S. Office of Management and Budget (1978). *Directive No. 16, Standard Classification of Fields of Science and Engineering*. Washington, DC: Executive Office of the President.

U.S. Office of Management and Budget (2002). *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies*. Notice: Re-publication. February 22. Washington, DC: Executive Office of the President.

U.S. Office of Management and Budget (2006). *OMB Guidance on Data Submission under the Federal Funding Accountability and Transparency Act (Transparency Act)*. M-08-04, November 9. Washington, DC: Executive Office of the President.

U.S. Office of Management and Budget and Office of Science and Technology Policy (2009). Memorandum for the heads of executive departments and agencies. *Science and Technology Priorities for the FY 2011 Budget*. M-09-27, August 4. Available: http://www.whitehouse.gov/omb/assets/memoranda_fy2009/m09-27.pdf [accessed November 2009].

Van Raan, A.F.J. (1996). The interdisciplinary nature of science: Theoretical framework and ibliometric-empirical approach. In P. Weingart and N. Stehr, Eds., *Practising Interdisciplinarity* (pp. 66-78). Toronto: University of Toronto Press.

Wagner, C., and L. Leydesforff (2005). Network structure, self-organization and the growth of international collaboration in science. *Research Policy*, 34(10):1608-1618.

Yasitis, A., S. Zecher, S. Shackleford, and E. Kim (2004). *Interdisciplinary Research Funding Patterns: An Analysis of the National Science Foundation, the National Institutes of Health, and the Department of Energy*. Policy Analysis Workshop, Georgia Institute of Technology, November 14.

Appendix A

Acronyms and Abbreviations

AIBS	American Institute of Biological Sciences
CDC	Centers for Disease Control and Prevention
CFDA	Catalog of Federal Domestic Assistance
CFFR	Consolidated Federal Funds Report
CIP	Classification of Instructional Programs
CRISP	Computer Retrieval of Information on Scientific Projects
CRS	Congressional Research Service
DoD	U.S. Department of Defense
DoE	U.S. Department of Energy
DOI	Digital Object Identifier system
DTIC	Defense Technical Information Center
FAADS	Federal Assistance Award Data System
FAC	Federal Audit Clearinghouse
FFATA	Federal Funding Accountability and Transparency Act of 2006
FFRDC	federally funded research and development center
FICE	Federal Interagency Committee on Education
FPDS	Federal Procurement Data System
FPDS—NG	Federal Procurement Data System—Next Generation
HHS	U.S. Department of Health and Human Services

IMPAC-II	Information for Management, Planning, Analysis, and Coordination
IPEDS	Integrated Postsecondary Education Data System
LOD	Linking Open Data community project
NAICS	North American Industrial Classification System
NASA	U.S. National Aeronautics and Space Administration
NCES	National Center for Education Statistics
NEC	not elsewhere classified
NIEHS	National Institute of Environmental Health Sciences
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NITRD	National Coordinating Office for Networking and Information Technology Research and Development
NRC	National Research Council
NSF	National Science Foundation
NSTC	U.S. National Science and Technology Council
OECD	Organisation for Economic Co-operation and Development
OMB	U.S. Office of Management and Budget
OSTP	U.S. Office of Science and Technology Policy
OWL	Web Ontology Language
PCAST	President's Council of Advisors on Science and Technology
PRIMUS	PRIsm MUlti-object Survey
RaDiUS	Research and Development in the United States
RCDC	Research, Condition, and Disease Categorization system
R&D	research and development
RDF	resource description framework
RDFA	resource description framework in attributes
RFFS	resource description framework schema
SBE	Directorate for Social, Behavioral and Economic Sciences, NSF
SciSIP	Science of Science and Innovation Policy Program
S&E	science and engineering
SKOS	Simple Knowledge and Organization System
SOC	standard occupational classification
SPARQL	SPARQL Protocol and RDF query language
SPIRES	Scientific Publication Information Retrieval System
SRS	Division of Science Resources Statistics, NSF

SWEO	Semantic Web Education and Outreach Interest Group
UNESCO	United Nations Educational, Scientific, and Cultural Organization
URI	uniform resource identifiers
USDA	U.S. Department of Agriculture
W3C	World Wide Web Consortium
W3C SWEO	World Wide Web Consortium Semantic Web Education and Outreach Group
XML	eXtensible Markup Language

Appendix B

Guide to Federal Grants and Contracts Databases

This appendix describes the main sources of information about federal grants and contracts—grants.gov and USAspending.gov; the databases that support them—the Federal Assistance Awards Data System (FAADS) and the Federal Procurement Data System (FPDS); and two specialized databases—the Office of Management and Budget MAX Information System and the National Institutes of Health IMPAC II system. The descriptions summarize their usefulness in terms of the discussion in this report.

GRANTS.GOV

Grants.gov is the main portal for grant applicants. Managed by the U.S. Department of Health and Human Services (HHS), grants.gov is used by 26 federal agencies that award grants, including almost all of the agencies that respond to the federal funds surveys from the Division of Science Resources Statistics (SRS) of the National Science Foundation (NSF). Grants.gov offers a standardized interface, which has simplified the research grant application process. As of fall 2008, grants.gov did not offer an accessible system for aggregating data on either grant applications or grant awards, but the portal does offer some promise for enabling a common reporting system for data on research grant awards across multiple federal agencies. Because grant applicants complete the information entered into grants.gov, this system offers an ideal mechanism for allowing individual scientists to note their field of science (for example, Ph.D. field, field of employing department, and the project's field of science). In addition, using grants.gov to collect administrative data about awarded grants would enable cross-references to reported agency spending and university acquisition of federal funds.

USASPENDING.GOV

USAspending.gov was created in response to the requirement of the Federal Funding Accountability and Transparency Act (Transparency Act). USAspending.gov aims to provide the public with information about how their tax dollars are spent. The ability to look at contracts, grants, loans, and other types of spending across many agencies, in greater detail, is a key ingredient to building public trust in government.

USAspending.gov collects data about the various types of U.S. government contracts, grants, loans, and other types of spending. For the convenience of users, USAspending.gov presents data in a different way than in many transactional databases, making it more easily understood and accessed. The original data fields and information are also available unmodified in USAspending.gov. The data can be seen if the user chooses the "Complete (all information)" level of detail, which is available on all data searches. In some cases USAspending.gov has modified data from the Federal Assistance Awards Data System (FAADS) and the Federal Procurement Data System (FPDS).

FEDERAL ASSISTANCE AWARD DATA SYSTEM (FAADS)

The Federal Assistance Award Data System (FAADS) was established by Title 31 Section 6102(a) of the U.S. Code, which mandates a uniform system for reporting information on federal government financial assistance transactions. Since 1982 the Census Bureau has served as the executive agent for FAADS. The Census Bureau receives data files from federal awarding agencies and disseminates them electronically.

Although FAADS is neither an accounting system nor a searchable database, it provides detailed listings of federal awards to specific institutions, which permits detailed reports on federal funding by performing institutions, geographic locations, and type of performer. There is currently no simple method for separating research and development (R&D) spending from other forms of federal spending, and FAADS does not capture all forms of federal extramural spending. However, FAADS does include reports from nearly all of the NSF SRS respondent units. Thus, FAADS does allow for detailed listings of extramural research grants from sponsoring agencies (such as NSF), and as such offers great promise in being able to populate federal support survey data fields.

FEDERAL PROCUREMENT DATA SYSTEM (FPDS)

The Federal Procurement Data System—Next Generation (FPDS-NG) is the central repository of information on federal contracting and contains detailed information on contract actions over $3,000 (FY 2004 and subsequent data). The FPDS data system aims to identify who buys what, from whom, for how much, when and where. It is managed by the Federal Procurement Data Center (FPDC), part of the U.S. General Services Administration. FPDS inputs into the FFATA-mandated database, but appears to have little relevance to NSF SRS because its reports and data focus on procurement actions, a category of federal spending mostly distinct from R&D. Related to the above databases is the Catalog of Federal Domestic Assistance (CFDA), a comprehensive source of federal assistance opportunities that is updated continually; it does not contain a data reporting function, however, meaning it is unlikely to be relevant for obtaining funding data. According to the FDPS website, "The ability to look at all contracts across many agencies, in greater detail, is a key component in establishing trust in our government and credibility in the professionals who use these contracts."[1]

MAX INFORMATION SYSTEM (MAX)

The Office of Management and Budget uses the MAX Budget Information System to collect, validate, analyze, model, and publish budget information. OMB's MAX Budget Systems provide an integrated platform for the collection, retrieval, manipulation, presentation, and publication of budget data. The system is used extensively throughout the year. The database includes hundreds of budget "versions" encompassing current and past policy, baseline, program and financing, object class, character class, credit, federal employment, financial management, program assessment, congressional action, and special budget exercise data. Research and development contacts are classified in object class 25.4 (advisory and assistance services), 25.4 (operation and maintenance of facilities) or 25.5 (research and development contracts) as appropriate.

IMPAC II

IMPAC II (Information for Management, Planning, Analysis, and Coordination) is an internally focused National Institute of Health system that supports management system of research grants and maintains a data repository for reporting tools and functions for the public. IMPAC II is

[1] Available: http://www.fpdsng.com/questions.html.

integrated with the NIH eRA Commons—an online interface where grant applicants, grantees and federal staff at NIH and grantor agencies can access and share administrative information relating to research grants—to permit a two-way flow of information between NIH and the external research community.

Appendix C

Modernizing the Infrastructure of the National Science Foundation Federal Funds Survey: Summary of a Workshop

As a key data-gathering activity, the Panel on Modernizing the Infrastructure of the National Science Foundation Federal Funds Survey hosted a workshop in Washington, DC, in September 2008. The first day of the workshop included presentations from four perspectives: (1) users of data from the federal funds and federal support surveys for research and development (R&D); (2) agencies that provide the data to the National Science Foundation (NSF) with responsibility for the surveys and other activities; (3) representatives of the U.S. Office of Management and Budget (OMB) with responsibility for overseeing government-wide implementation of the E-Government Act and other laws that are designed to improve administrative data; and (4) users of administrative data on grants and contacts, who focused on long-term opportunities to use federal government administrative data and other sources for measuring federal R&D spending. On the second day of the workshop, attention was directed toward issues associated with the classification of fields of science and engineering used in these and other NSF surveys. The workshop included presentations on emerging classification systems and a NSF staff presentation on the interface of the classification system used by NSF with other systems. (See the end of this appendix for the workshop agenda.)

This appendix summarizes the presentations and discussion at the workshop. Several of the presenters made suggestions and recommendations during the workshop; the panel considered them in the course of its work, but they are not included in this summary.

USER NEEDS

Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP), which was represented at the workshop by Dianne DiEuliis, is a major consumer of data from the federal funds survey. OSTP's mission is to advise the President and others in the Executive Office of the President on the effects of science and technology on domestic and international affairs. An important aspect of that mission is to lead interagency efforts to develop and implement sound science and technology policies and budgets. Data on federal expenditures for research and development play a central role for OSTP in its oversight and program coordination functions.

In 2005, the OSTP director, John H. Marburger, III, explained the specific interest of OSTP in the data from the federal funds and other NSF surveys of research and development spending at the 30th Annual AAAS Forum on Science and Technology Policy in Washington, DC. He pointed out that indicators of the health of U.S. science are based on "indicators that are based on a data taxonomy that is nearly three decades old" and that "methods for defining data in both public and private sectors are not well adapted to how R&D is actually conducted today." He referred to the previous National Research Council (2005b) study, which concluded that NSF R&D expenditure data are often ill-suited for the purposes for which they have been used and urged that the report's recommendations for improving various components of the data and enhancing their usefulness "should receive high priority in future planning within NSF."[1]

Jointly with OMB, OSTP prepares the administration's statement of R&D priorities, reflecting input from the President's Council of Advisors on Science and Technology (PCAST) and the National Science and Technology Council (NSTC). That statement provides general guidance for setting priorities for agency R&D programs: Significantly, it defines a set of initiatives for which budget and expenditure data should be collected. The most recent guidance defined six areas of highest priority for R&D investments: (1) homeland security and national defense, (2) energy and climate change technology, (3) advanced networking and information technology, (4) a national nanotechnology initiative, (5) complex biological systems, and (6) the environment.[2]

The identification of priority areas leads to a demand for data that will measure the status of implementation of the investment priorities. In terms

[1]Available: http://www.aaas.org/news/releases/2005/0421marburgerText.shtml [accessed February 2009].

[2]Available: http://www.whitehouse.gov/omb/memoranda/fy2007/m07-22.pdf [accessed March 2009].

of specific requirements for data from the federal funds survey, OSTP made a strong case for the maintenance of historical data in order to preserve information on spending trends in detail.

Congress

Congress plays an active role in generating a demand for information on federal funds and support, and, in its oversight role, pays close attention to the management and direction of the NSF data collection efforts. For example, the federal support survey has been mandated by Congress since 1950.[3] James Wilson, then majority staff director of the Research and Science Education Subcommittee of the House Committee on Science and Technology, spoke on the interests of that committee, which has key authorization jurisdiction over the NSF portfolio of R&D surveys. Although there is an interest in R&D spending data by field of science and engineering with consistency over time in order to understand trends in funding, the interest of Congress also often has to do with program categories—such as those defined by the administration's investment priorities—rather than specific fields. For example, Wilson said the committee would like to be able to see data on the projects that support the nanotechnology initiatives, but cannot do so because the field's information is not sufficiently granular. In addition, much of the R&D activity of interest is hidden in the "not elsewhere classified" classification, and the combination of mathematics with computer sciences is also too broad for policy makers. Wilson said the committee would like more information on collaborative research and the nature of the performers of the research.

Wilson also expressed a need for data that are be compiled and published in a timely manner. Due to the legislative calendar and the budget cycle, congressional committees need fiscal year expenditure information within 6 months of the end of the fiscal year. Currently, the relevant data are not available in time to have meaningful input into the authorization and appropriation processes except in retrospect.

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

[3]The National Science Foundation Act of 1950, as amended, requires that the National Science Foundation "initiate and maintain a program for the determination of the total amount of money for scientific and engineering research, including money allocated for the construction of the facilities wherein such research is conducted, received by each educational institution and appropriate nonprofit organization in the United States, by grant, contract, or other arrangement from agencies of the Federal Government, and to report annually thereon to the President and the Congress."

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 publications such as *Science and Engineering Indicators*.

John Sargent of the CRS, who has written several studies with emphasis on categories of R&D expenditures, talked about his recent study of nanotechnology as an example of how the data are used. By definition, nanotechnology crosses several fields of science and engineering and represents considerable complexity. The multidisciplinary nature of nanotechnology spending is not depicted in the regular NSF data, so CRS relies on special data calls for its information. CRS analysis is also limited by the large "not elsewhere classified" classification: It is believed that it includes many of the growing and emerging research areas of considerable interest. To obtain a full picture of R&D spending, CRS needs data from the support survey on facilities and infrastructure maintenance spending in addition to the investment data.

Professional Societies, Associations, and Public Interest Groups

Users in the community of professional societies, associations, and public interest groups were represented by Robert E. Gropp, senior public policy representative of the American Institute of Biological Sciences (AIBS). AIBS is an umbrella society for 87 professional biological science societies whose 240,000 members study every sub-discipline of the biological sciences, including botany, ecology, taxonomy, evolution, and agricultural sciences. Gropp expressed concern with the treatment of the biological sciences in the NSF taxonomy of fields of science and engineering. There is also a tendency to lump basic and applied sciences together, making it difficult to identify the evolution of R&D from basic to applied research to development.

Gropp also discussed the need for data to shed light on the growing multidisciplinary category of service science (also known as service science management and engineering). Service science is an interdisciplinary approach to the study, design, and implementation of service sector systems. This is a growing academic discipline and research area that is characterized by the application of computer science, cognitive science, economics, organizational behavior, human resources management, marketing, and operations research in support of understanding aspects of the service sector. Like nanotechnology and other categories that lump together various disciplines, the emerging service science field is very difficult to measure with current NSF data.

National Science Board

The National Science Board is the body that provides oversight for and establishes the policies of the National Science Foundation, within the framework set by the President and Congress. The board also serves as an independent body of advisers to both the President and Congress on broad national policy issues related to science and engineering research and education. The board is responsible for preparing the biannual *Science and Engineering Indicators* report, which provides a broad base of quantitative information about U.S. science, engineering, and technology for use by public and private policy makers and makes extensive use of the information from the federal funds and federal support surveys.

Louis Lanzerotti, the chair of the board, sent a letter to the panel in connection with its work; the letter was made available to the participants in the workshop and is summarized below:

> As Chairman of the National Science Board's Subcommittee on Science and Engineering Indicators, I am an enthusiastic supporter of improvements in R&D data resources, as are other members of the Board. Although I cannot participate in person, I applaud efforts to improve the quality and utility of the Federal Funds for Research and Development survey, an important data resource for the Board's *Science and Engineering Indicators 2008* report in Chapter 4, "Research and Development: National Trends and International Linkages," and in Chapter 5, "Academic Research and Development."
>
> In order to contribute to the Workshop's discussions, I would like to refer you to the Board's conclusions on data resources for Federal R&D budget allocation decisions as stated in its 2001 study and report, *Federal Research Resources: A Process for Setting Priorities* (NSB 01-156), which may be helpful in this current study. Although the Board's 2001 report addresses data resources for Federal R&D budget allocation decisions, I believe some of the conclusions are relevant to your more focused examination of the Federal Funds survey. These conclusions are paraphrased as follows:
>
> - Improving Federal budget data and data systems requires a long-term commitment and appropriate support from OMB and Congress.
> - Input from potential users and contributors are needed.
> - Data must be made easily accessible to users.
> - Definitions of research activities must be consistently applied across Federal departments, agencies and programs and measured to capture the changing character of research and research needs.
> - Flexibility in defining categories of research for tracking purposes is especially important for monitoring emerging research areas and addressing the range of modes for research—from individual investigator to major center or facility.

ISSUES OF DATA PROVIDERS

Data providers in the federal government were represented by the National Institutes of Health (NIH) and the U.S. Air Force. In addition, the panel heard from a representative of the Centers for Disease Control and Prevention (CDC), a major unit of NIH. The NIH reports the agency's R&D expenditures directly to NSF, while the Air Force reports through the Department of Defense (DoD), and the CDC reports through NIH. Thus, the panel was able to gain an appreciation of the concerns of both direct reporters and those who report through other agencies.

Israel Lederhendler, the director of the Office of Extramural Research at NIH, detailed several concerns with the NSF data collection program. The NIH manages its R&D portfolio at the project level and, for its own management and reporting purposes, aggregates projects into categories centered around research areas, diseases, and conditions that are not easily described in the NSF taxonomy of science and engineering fields. Furthermore, as a matter of policy, NIH classifies its program as "medical science," given the output of the R&D expenditures, and aggregates its reporting to that level even though other fields are clearly represented in the agency R&D program. The agency is focused on outcomes, but the taxonomy is organized around inputs.

Lederhendler said there is also a problem caused by having to force-fit projects with multiple disciplines in a single category. The "fitting" is highly subjective, and so the numbers can change over time because of reporting changes rather than real changes in disciplines. A solution would be to add metadata (information about the data) to project descriptions. The metadata would describe all aspects of a project and serve as the basis for coding to the various reporting requirements. Such content-rich project descriptions could be maintained on a system like research.gov, which could serve as a portal for a federated system.

Large R&D agencies like NIH are further challenged because of the requirement to report on many surveys with different definitions. The federal funds and federal support surveys, for example, have different definitions of fields. Lederhendler suggested the need for a federated system of information and ontology. NIH is now developing the Research, Condition, and Disease Categorization (RCDC) system—a prototype system for common categorizing and reporting for both intramural and extramural research that could be a step toward the needed federated system.

Lederhendler said that several other issues raise questions about the quality of the data that agencies provide to NSF. For example, different grant and contracting practices, including the lag between the award and the payment, affect expenditure data. There is also no good information about international R&D expenditures, so these data in the NSF report may be questionable. As a remedy, he suggested that NSF convene a standing advisory group composed of reporting agencies to provide input to NSF on reporting issues.

His colleague, Robin Wagner, formerly of CDC, raised additional concerns. She stated that there is no agency information on performer by location. Since the data are not automatically available, CDC must issue an internal data call for the information, which delays submission of the reports and possibly adds error to the information. As a practical matter, the coding in CDC is done mainly by budget specialists, not scientists, so the coding may not be informed by the scientific purpose of the work.

The Department of the Air Force was represented by Tom Russell, director of Aerospace, Chemistry, and Materials for the Air Force Office of Scientific Research. He reflected the view of a manager of R&D programs in a large, decentralized system and the consequent difficulties for reporting in the manner prescribed by NSF. Within the DoD, the maintenance of information about R&D projects and outcomes is the responsibility of the Defense Technical Information Center (DTIC), which provides centralized information on DoD scientific, technical, engineering, and business-related work. DoD research agencies, both the policy and program agencies and the procurement system and laboratories that support them, are geared to reporting in the DoD system, while the annual NSF requirement is an outside reporting requirement and only marginally relates to the internal reporting system.

Russell is encouraged by several initiatives toward integrating definitions, classifications, and reporting requirements across the government. The Chief Financial Officer Act is increasingly improving the integrity of government financial data, while the Federal Funding Accountability and Transparency Act (FFATA) and the performance reporting and evaluation system that has evolved in response to the Government Performance and Results Act promises to provide a common language and a common basis for reporting in all federal agencies and should lead to an ability to integrate reporting across the government.

SCIENCE OF SCIENCE POLICY: METRICS

The science of science policy, first proposed by OSTP, has been institutionalized in NSF with a Science of Science Policy Program that is expected to use the data provided by the federal funds and support surveys. This information will be needed in response to new requirements for metrics to assess the progress of 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 research are transformed into social and economic outcomes.

Julia Lane of the NSF's Directorate for Social, Behavioral, and Economic Sciences (SBE) described this new initiative. A key aspect of the initiative is to develop new and improved metrics, datasets, and analytical tools. NSF has solicited input on several questions of substance (how fields of science and engineering are defined and if they are changing), whether the critical input measures (basic research, applied research and development) are appropriate, and the identification of critical output measures. Her presentation stressed that the way to deal with the lack of metrics is tied to more extensive and intensive use of administrative record data, which have provided answers in other research, areas such as understanding business dynamics and the nature of work.

QUALITY AND CONTENT ISSUES

Against this backdrop of unfulfilled user needs and producer concerns, NSF collects, processes, and publishes the only source of information on federal expenditures for R&D based on the federal funds and support surveys. John Jankowski, the program manager for the surveys in NSF's Division of Science Resources Statistics, talked about the administrative and technical aspects of the surveys, underscoring the current strengths and practical limitations of the surveys, as a basis for considering possible survey changes.

The use of the term "survey" may be a bit of a misnomer, Janowski said. The surveys are essentially censuses of federal R&D spending: The federal funds survey, which covers all known federal agencies that fund R&D (both in-house and external); and the federal support survey, which covers the federal agencies that account for almost all federal R&D support to academic institutions. Thus, coverage is clearly a strength of the surveys.

Content is also very robust. The federal funds survey collects aggregate totals to performer sectors (e.g., R&D obligations to all universities and colleges combined), and has both obligations and outlays–the only source of such information. The federal support survey collects totals on all science and engineering obligations, including R&D, by federal agencies to institution-specific academic and nonprofit institutions.

There is also great depth to the published detail. The federal funds survey publishes outlays for total R&D and R&D facilities data for 3 years—the past year (actual), the current year (preliminary), and the next or budget year (projected)—and by funding agency and performing sector—federal intramural, industry, universities, nonprofit institutions, individual federally funded research and development centers (FFRDCs), nonfederal governments, and foreign performers. These data are further classified by R&D work category—basic research, applied research, development or R&D plant—and by detailed science and engineering fields.

The location of the performer is also published at the state and foreign country level.

The federal support survey publishes data for the immediate past year for 19 departments and agencies on obligations to nearly 1,200 individual universities and colleges (as of fiscal 2006), broken down by R&D; R&D plant; fellowships, traineeships, and training grants; facilities and equipment for instruction in science and engineering; general support for science and engineering; and other activities related to science and engineering. Within the academic sector, totals can be derived for historically black colleges and universities, high Hispanic enrollment institutions, minority serving institutions, and tribal colleges and also by public or private academic institutions and for 1,323 individual independent nonprofit institutions (as of fiscal 2006).

Janowski noted that the surveys are neither large nor particularly expensive (as federal government recurring surveys). NSF has collected data from about 60 reporting entities in recent years, yielding published data for about 90 agencies for the federal funds survey and for 19 agencies for the federal support survey. The survey costs for fiscal 2007, the most recent year available, were $450,000 for the federal funds survey and $420,000 for the federal support survey. Collection is a relatively straightforward operation, with the majority of the agencies using the FEDWeb reporting tool, and most of the others reporting by providing electronic data files.

Janowski said that timeliness is an issue. The survey is introduced to the field in February of each year, covering spending in the prior year. The due date is usually mid-April, but, in the past several years, some agencies did not submit their data until November or December. Since NSF has a policy of not publishing the totals until all agencies have reported, data were not released until February of the following year, one year after the surveys were sent out and more than a year after the end of the reporting time.

Due to these delays in publication, the data for the most recent 2 years are preliminary and projected, which tends to create a false sense of timeliness. The data that were released in February 2008, for example, had preliminary data for fiscal 2006 and projected data for fiscal 2007. The delays also introduce a type of error, since there are sometimes significant differences between the preliminary, projected, and final estimates. In 8 of the last 9 years, budget year projections were higher than the final obligations. In some years, the differences between the first published (projected) and final estimates have varied by 5 percent or more.

Jankowski's presentation listed several shortcomings in the current data. The key data gaps are for federal R&D laboratories, for which there is undercounting of internal versus external R&D, and international science and technology activities, which only identify R&D by foreign location, not performer.

IMPROVEMENT OPPORTUNITIES

A three-person panel explored several recent initiatives that could yield long-term opportunities for more extensive use of federal government administrative data and ancillary data sources for measuring federal R&D spending. The panel was comprised of Andrew Reamer, a fellow with the Metropolitan Policy Program at the Brookings Institution, representing a nongovernment public policy research; Mark Bussow of the Office of Management and Budget; and Jeffrey Alexander, representing a private-sector firm that uses federal data on R&D spending in supporting economic development strategies.

Reamer provided a summary of the development of congressional mandates to provide information on federal spending. He summarized and critiqued eight interrelated mandates for collecting, organizing, and providing information on federal spending. Several of the mandates place OMB in a central role.

There are four primary information and data repositories: the Catalogue of Federal Domestic Assistance (CFDA), the Federal Assistance Award Data System (FAADS), the Federal Audit Clearinghouse (FAC), and the Federal Procurement Data System (FPDS). Two mandated reports require gathering accurate and timely program information—the Information of Federal Assistance to State and Local Governments System (known as 31 USC 1112 (f) system after its legislative mandate) and the Consolidated Federal Funds Report (CFFR). Under the designation of secondary data repositories, Reamer discussed the now defunct RaDiUS repository and the FFATA mandate. He suggested that the RaDiUS system could serve as a prototype for the FFATA effort.

Mark Bussow reported on OMB activities with regard to improving the quality of and access to administrative data that would be useful for measuring federal R&D spending. Under the Federal Funding Accountability and Transparency Act of 2006 (Public Law 109-282), OMB was given responsibility for establishing a publicly available online database containing information about entities that are awarded federal grants, loans, and contracts. The act was to be implemented in two phases. The first phase called for a new database to provide information on entities (corporations, associations, partnerships, sole proprietorships, limited liability companies, limited liability partnerships, states, and localities) that are awarded funds directly from the federal government by January 1, 2008. The second phase called for information on subgrantees and subcontractors that receive funds from a primary recipient by January 1, 2009. The database would provide the following information:

- name of entity receiving award
- amount of award

- type of award (e.g., grant, loan, contract)
- agency funding award
- a North American Industry Classification System (NAICS) code of the recipient or a Catalog of Federal Domestic Assistance (CFDA) number (if applicable)
- program source
- award title that describes the purpose of the funding
- location of recipient
- city, state, congressional district, and country in which award performance primarily takes place
- unique identifier for entity receiving award and of the parent entity of recipient, if one exists
- any other information specified by OMB

OMB has elected to leverage existing systems, functionality, and available data to the fullest extent and has selected three major financial assistance databases as sources of information for the new website: the Federal Procurement Data System (FPDS)—Next Generation (NG), the Federal Assistance Award Data System (FAADS), and Grants.gov. FPDS and FAADS are known to have serious data problems—being incomplete and untimely and having inaccurate entries—so a high priority initially has been to clean up these databases at the source (U.S. Office of Management and Budget, 2007). For access, OMB selected a private "watchdog" organization, OMB Watch, to participate in launching the website, "Fedspending.org," which provides public access to information on federal grants and contracts as mandated under the act.

Bussow said that although OMB is making good progress toward meeting the goals and objectives of the FFATA, this is seen as a long-range process. A full, timely, and accurate database, meeting the needs of multiple users with the ability to substitute for data calls and other data collections, is still perhaps a decade away, but the process has a sense of direction now.

Jeffrey Alexander made the case for timely, accessible, and model-based data for understanding regional innovation. The data are needed for internal analysis, benchmarking, and decisions on investing public funds. The federal R&D funding data are used for correlation analysis (to understand the relationship between R&D funding and patents to determine if there is a clustering effect) and to facilitate innovation and commercialization by identifying, recruiting, and retaining researchers and connecting collaborators. This analysis is hindered by several factors: a mismatch of the need for aggregated budget reports for program administration and data for local purposes; data quality issues, such as data entry errors, incorrect categorization, missing or incomplete records and lack of timeliness; lack of data integration; and outdated, incoherent, irrelevant, and inconsistent taxonomies.

Alexander reported on the kind of analysis that was and could be done with RaDiUS-type of data. Keyword searches and other techniques were able to identify a large number of R&D contracts, grants, and other activities that were associated with very specific technologies. However, the keywords tended to be very brief project descriptions that were not at all standard, nor did they cover all potential uses of the information. The same is true of FAADS and FPDS files.

He suggested some policy changes that could improve the quality and usability of the R&D spending data: the use of triangulation to identify and correct errors; better enforcement of consistent reporting policies; unified format standards and data architectures; and increased use of machine analysis. However, he did not support a unified taxonomy of fields of science and engineering because advances in information technology have created an environment in which multiple taxonomies are supported by such new technologies as text analysis, concept inference, and evolving semantic web programs. The taxonomies most useful for economic development analysis would be self-organizing and self-correcting, which would require computing power, intensive design effort, and commitment of resources.

TAXONOMIES

The second day of the workshop was devoted to discussing issues of taxonomy of fields of science and engineering. Beginning this discussion was Gretchen Gano, librarian for public administration and government information at New York University, who elaborated on the state of classification science and suggested a way of considering the science and engineering classification of the future.

The assumption underlying classification is that there is a hierarchy based on origins, which can be natural or human invented. The classification structure for science and engineering is now standardized, comparable over time, and descriptive, though it is difficult to translate at the boundaries (such as with multidisciplinary fields). In reference to a recent study (Cheney and Park, 2005), she pointed out that interdisciplinary, multidisciplinary, and transdisciplinary fields have emerged that are not well represented in the taxonomy. This has led, for example, to "not elsewhere classified" fields being larger than their peer disaggregated categories.

Gano's definition of interdisciplinarity was taken from a report of the National Research Council (2005a, p. 26): "a mode of research by teams of individuals that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice." There is a process that has emerged as fields go from overlapping to becoming interdisciplinary toward a new field, supported by an infrastructure of academic departments, defined grants, journals, and subject headings.

Gano concluded that NSF should move away from hierarchical fields and subfields toward discipline-spanning classifications of the key elements of scientific practice. As an example, she cited the New York University PRIsm MUlti-object Survey (PRIMUS), a wide-field survey to advance the study of the structure of the universe. It integrates phenomena, data, theory, method, and practice in an integrated system. She suggested that science and engineering fields, on a pilot basis, could be viewed as clusters of attributes that could be mapped to standard disciplinary taxonomies. The information retrieval would be aided by semantic web technology, using the resource description framework (RDF) structure that describes and interchanges metadata on the web. An example of the application of the RDF structure is DBpedia, which extracts structured information from Wikipedia and links to other datasets.

Reporting on the results of an internal review of taxonomy issues, Geri Mulrow of NSF discussed general issues of the collection of taxonomy data and discussed the interface of the NSF taxonomy with other systems. She has been responsible for an internal staff study of taxonomy issues that has reviewed prior Division of Science Resources Statistics (SRS) reports, interviewed NSF division directors and program officers, and interviewed outside researchers.

The SRS Division has published several studies on the taxonomy since 2000. It has also conducted a workshop on OMB directive No. 16 (Cheney and Park, 2005), as well as sponsoring the present workshop. The 2004 workshop concluded that classifications that describe the dynamic science and engineering fields need to be revised periodically and that criteria and procedures are needed for the classification scheme. The updates should be based on input from the disciplines and respondents and data users, understanding that different disciplines view the same topic from different perspectives. However, users also want consistent data and categories over time, so there is a tension between updating the classification structure and continuity.

The main alternative schemes were the NSF classification embodied in OMB Directive No. 16, the Classification of Instructional Programs (CIP), the Standard Occupational Classification (SOC) system, the Frascati Manual, and the National Research Council taxonomy. The studies found general agreement at the major field level but inconsistencies at the subfield level. All reported issues with inter- and multidisciplinary fields. Mulrow's work found that the set of principles for classification that underscored each of the schemes were based on the uses of the classification, all had guidelines for determining how to code units, and all were generally hierarchical in structure.

Her discussions with NSF program divisions led to her conclusion that research is becoming increasingly inter- and multidisciplinary in nature and will continue so in the future. The new and growing areas of research are generally identifiable, at the boundaries of disciplines. Professional associations are a good source of information on emerging trends in the fields; many of them periodically reorganize to accommodate new fields. However, the educational system has generally lagged behind the research community in coming to grips with increasingly multidisciplinary activities.

Another study Mulrow summarized was a 2008 report on the S&E taxonomy, based on interviews with responding federal agencies (Macro International, 2008). The purpose of this study was to gain an understanding of the strengths and weaknesses of the current fields of science and engineering taxonomy used in the federal funds survey; identify the use of the taxonomy across the agencies; and detail the current process used by agencies for allocating and managing their research funds and how they report them to NSF. The findings were instructive and gave a cause for concern. The largest R&D funding agencies do not use the OMB/NSF fields for program management and budgeting, and there is little consistency across the agencies in the fields that they use to track their R&D. The agencies do not use the OMB/NSF fields because they do not relate to their programs, they fail to capture inter- and multidisciplinary research, and they are generally not useful management tools. A good deal of staff judgment is used in coding the fields, and sometimes coding is by formula (percent distribution) or by computer techniques. Agencies tend to manage more by program categories (energy, environment, disease) than by field and can report the categories more readily than the fields.

Mulrow's internal study led her to conclude that it would be useful to build on the current report and gain a deeper understanding of the actual R&D programs in the agencies and to start with a few of the largest agencies to maximize return on the investment. The review of agency program management should include considering the existing agency administrative record systems and linkage mechanisms. As for the issue of modernizing the taxonomy, Mulrow suggested starting with the development of principles and guidelines for a classification system, reviewing the need for multiple classifications of the data, considering network representations of the information, and, in preparation for updating the taxonomy, developing ways to bridge the past with the current with the future classification system.

WORKSHOP AGENDA

Workshop on Modernizing the Infrastructure of the National Science Foundation Federal Funds Survey

September 5-6, 2008
Room 101, Keck Center
500 Fifth Street, NW
Washington, DC 20001

Objectives of Workshop:

1. To explore issues involved with the NSF federal funds and federal support surveys.
2. To learn about user needs for federal R&D expenditure information.
3. To understand federal agency data sources for federal R&D expenditures.
4. To consider short- and long-term changes in the federal funds and federal support surveys.
5. To consider the use of administrative data under the E-Government and Transparency Acts to provide information on federal R&D spending.
6. To consider issues with the taxonomy of fields of science and engineering.
7. To initiate preparation of the final report with recommendations.

Friday, September 5, 2008

Open Session

8:30-9:00 a.m. Welcome (Continental Breakfast served)
Christopher Hill, *Chair,* George Mason University

9:00-10:30 a.m. Overview of User Requirements
Diane DiEuliis, Senior Policy Analyst, Office of Science and Technology Policy, Executive Office of the President

James Wilson, Majority Staff Director, Research and Science Education Subcommittee, Committee on Science and Technology, U.S. House of Representatives

John Sargent, Congressional Research Service

10:30 a.m.-12:00 p.m.	Challenges Facing Data Providers Israel Lederhendler, Director, DIS, Office of Extramural Research, National Institutes of Health Tom Russell, Director of Aerospace, Chemistry, and Materials, Air Force Office of Scientific Research
12:00-1:00 p.m.	Working Lunch
1:00-2:30 p.m.	Strengths and Limitations of Federal Funds/ Support Surveys John Jankowski, National Science Foundation Using Administrative Data to Estimate Federal R&D Expenditures Julia Lane, National Science Foundation
2:30-2:45 p.m.	Break
2:45-4:00 p.m.	Focus on Long-Term Opportunities to Use Federal Government Administrative Data and Other Data Sources for Measuring Federal R&D Spending Mark Bussow, Office of Management and Budget Andrew Reamer, Brookings Institution Jeff Alexander, New Economy Strategies
4:00-5:00 p.m.	Open Discussion

Saturday, September 6, 2008

Open Session

8:00-9:30 a.m.	Issues with the Taxonomy of Fields of Science and Engineering (Continental breakfast served) Gretchen Gano, Librarian for Public Administration and Government Information, New York University

9:30-9:45 a.m.	Break
9:45-10:30 a.m.	Issues with the Collection of Taxonomy Data; Interface of S&E Taxonomy with Other Systems (CIP, SOC, etc). Jeri Mulrow, National Science Foundation

Appendix D

Biographical Sketches of Panel Members and Staff

CHRISTOPHER T. HILL (*Chair*) is professor of public policy and technology at George Mason University. He has served on the professional staff at the National Academy of Engineering, the National Research Council, and the Congressional Research Service. He is a fellow of the American Association for the Advancement of Science. His publications have been in the field of technological innovation and its impact on the economy, the impact of federal regulation on innovation, and the university perspective on issues of federal research and development (R&D) procurement. As vice provost for research, he oversaw completion of the National Science Foundation survey of academic R&D at George Mason University. He has a Ph.D. in chemical engineering from the University of Wisconsin.

WILLIAM B. BONVILLIAN is director of the Massachusetts Institute of Technology's (MIT) Washington, DC, office. In that capacity he works to support MIT's relations with federal research and development (R&D) agencies and its role on national science policy. Previously he served as legislative director and chief counsel to U.S. Senator Joseph Lieberman, working on science and technology policies and innovation issues. He worked extensively on legislation creating the Department of Homeland Security, on intelligence reform, on defense and life science R&D, and on national competitiveness and innovation legislation. He has also previously served as deputy assistant secretary and director of congressional affairs at the U.S. Department of Transportation. He has lectured and taught at Georgetown University and George Washington University and was the recipient of the IEEE Distinguished Public Service Award in 2007. For the National

Research Council, he serves on the Board on Science Education. He has a B.A. from Columbia University, an M.A.R. from Yale Divinity School in religion; and a J.D. from Columbia Law School.

KATY BÖRNER is the Victor H. Yngve professor of information science at the School of Library and Information Science at Indiana University. She is also adjunct professor in the School of Informatics, core faculty of cognitive science, research affiliate of the Biocomplexity Institute, fellow of the Center for Research on Learning and Technology, member of the Advanced Visualization Laboratory, and founding director of the Cyberinfrastructure for Network Science Center. Her research focuses on the development of data analysis and visualization techniques for information access, understanding, and management. She is particularly interested in the study of the structure and evolution of scientific disciplines, the analysis and visualization of online activity, and the development of cyberinfrastructures for large-scale scientific collaboration and computation. She is coeditor of *Visual Interfaces to Digital Libraries* and of a special issue of *Proceedings of the National Academy of Sciences* on mapping knowledge domains. She has a Ph.D. from the University of Kaiserslautern in Germany (1997).

MARY K. FEENEY is an assistant professor at the University of Illinois at Chicago. Her research specializes in public management, mentoring, outsourcing and contracting, and science and technology policy. Feeney's work has been published in *Administration & Society*, the *Journal of Public Administration Research and Theory*, *Public Performance and Management Review*, and *Research Evaluation*. She teaches courses in public management, nonprofit management, and survey research. She has a B.A. in political science from the University of Wyoming, an M.A. in public policy from Rutgers University, and a Ph.D. in public administration and policy from the University of Georgia.

DAVID GOLDSTON is a visiting lecturer at the Harvard University Center for the Environment. Previously, he held a one-year appointment as a lecturer in the Science, Technology and Environment Program at Princeton University's Woodrow Wilson School of Public and International Affairs, and he writes the monthly column "Party of One" on Congress and science policy for the journal *Nature*. From 2001 through 2006, he was chief of staff of the U.S. House of Representatives Committee on Science, which has jurisdiction over much of the federal research and development budget. He was also a key player in most environmental debates in the House from 1995, when he became legislative director to Representative Sherwood Boehlert of New York, until the end of 2006, when he retired from government service. For the National Research Council, he is a member of the Aeronautics and Space Engineering Board. He graduated from Cornell University and completed the course work for a Ph.D. in U.S. history at the University of Pennsylvania.

NANCY J. KIRKENDALL served as director of the Statistics and Methods Group in the Energy Information Administration and was a member of the senior staff from 2002 to 2008, when she retired from federal service. From 1996 to 1999, she served as senior mathematical statistician in the Statistical Policy Branch of the Office of Information and Regulatory Affairs in the U.S. Office of Management and Budget. There she served as the desk officer for the U.S. Census Bureau, chaired the Federal Committee on Statistical Methodology, and led a variety of interagency activities. She taught part time at George Washington University in the Statistics Department from 1978 to 1996 and in the Engineering Management and Systems Engineering Department from 1996 to 2002. She is a past vice president of the American Statistical Association and a past president of the Washington Statistical Society. She has B.S. and M.S. degrees in mathematics from Ohio State University and a Ph.D. in statistics from George Washington University.

JULIE THOMPSON KLEIN is professor of humanities in interdisciplinary studies/English and faculty fellow in the Office for Teaching and Learning at Wayne State University. She has also held visiting posts in Japan and New Zealand and was a Fulbright professor in Nepal and a senior fellow at the Association of American Colleges and Universities. Her area of expertise is interdisciplinary research and education. She received the Kenneth Boulding Award for outstanding scholarship on interdisciplinarity and has lectured and consulted throughout North America, Europe, Latin America, the South Pacific, and Asia. She is past president of the Association for Integrative Studies and former editor of its journal, *Issues in Integrative Studies*, and has served on national task forces and advised public and private agencies. She has a Ph.D. and undergraduate degrees in English from the University of Oregon.

KEI KOIZUMI is assistant director for federal research and development (R&D) in the Office of Science and Technology Policy of the Executive Office of the President. Prior to assuming this position, he served as director of R&D budget and policy programs at the American Association for the Advancement of Science (AAAS). His expertise focuses on the federal budget, federal support for research and development, science policy issues, and R&D funding data. At the AAAS, he was the principal budget analyst, editor, and writer for the annual AAAS reports on federal R&D and for the continually updated analyses of federal R&D on the organization's R&D

website. He is widely quoted in the general and trade press on federal science funding issues and speaks on R&D funding trends and federal budget policy toward R&D to numerous public groups and seminars. He has an M.A. from the Center for International Science at George Washington University and a B.A. from Boston University in political science and economics.

THOMAS J. PLEWES (*Study Director*) is a senior program officer for the Committee on National Statistics and was study director for an earlier National Research Council study 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 and served as chief of the U.S. Army Reserve. 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 George Washington University.

J. DAVID ROESSNER is associate director of the Science and Technology Policy Program at SRI International and professor of public policy emeritus at Georgia Institute of Technology. Prior to joining the Georgia Tech faculty in 1980, he was principal scientist and group manager for industrial policy and planning at the Solar Energy Research Institute in Golden, Colorado. He served as policy analyst with the National Science Foundation's (NSF) research and development assessment program and, subsequently, as acting leader of the working group on innovation processes and their management in the Division of Policy Research and Analysis at NSF. Since 2003, he has been senior evaluation consultant to the National Academies Keck Futures Initiative, a 15-year, $40 million program to foster interdisciplinary research in the United States. His research interests include national technology policy, the evaluation of research programs, management of innovation in industry, technology transfer, and indicators of scientific and technological development. He has B.S. and M.S. degrees in electrical engineering from Brown University and Stanford University, respectively, and M.A. and Ph.D. degrees in science, technology, and public policy from Case Western Reserve University.

MARTHA M. TAYLOR is assistant vice president for research and the director of the Office of Sponsored Programs, the preaward and nonfinancial postaward branch of Auburn University. She came to Auburn in 1989 as a database consultant and then served as assistant director after 3.5 years with the Texas A&M Research Foundation, working in the areas of postaward management and subcontracting. In her position, she is responsible for developing the data on research activities at Auburn to meet the National Science Foundation reporting requirements. Prior to her time in Texas, she worked as a contracts technician for Environmental Sciences and Engineering, Inc. (now QST Environmental), an environmental engineering firm in Gainesville, Florida. She is a member of the National Council of University Research Administrators and the primary representative for Auburn University with the Council on Governmental Relations. She graduated with honors from the University of Florida with a B.S. in business administration.

COMMITTEE ON NATIONAL STATISTICS

The Committee on National Statistics (CNSTAT) was established in 1972 at the National Academies to improve the statistical methods and information on which public policy decisions are based. The committee carries out studies, workshops, and other activities to foster better measures and fuller understanding of the economy, the environment, public health, crime, education, immigration, poverty, welfare, and other public policy issues. It also evaluates ongoing statistical programs and tracks the statistical policy and coordinating activities of the federal government, serving a unique role at the intersection of statistics and public policy. The committee's work is supported by a consortium of federal agencies through a National Science Foundation grant.