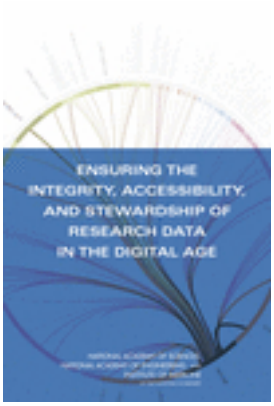
Free Executive Summary



Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age

Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age; National Academy of Sciences

ISBN: 978-0-309-13684-6, 188 pages, 6 x 9, paperback (2009)

As digital technologies are expanding the power and reach of research, they are also raising complex issues. These include complications in ensuring the validity of research data; standards that do not keep pace with the high rate of innovation; restrictions on data sharing that reduce the ability of researchers to verify results and build on previous research; and huge increases in the amount of data being generated, creating severe challenges in preserving that data for long-term use. Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age examines the consequences of the changes affecting research data with respect to three issues - integrity, accessibility, and stewardship-and finds a need for a new approach to the design and the management of research projects. The report recommends that all researchers receive appropriate training in the management of research data, and calls on researchers to make all research data, methods, and other information underlying results publicly accessible in a timely manner. The book also sees the stewardship of research data as a critical long-term task for the research enterprise and its stakeholders. Individual researchers, research institutions, research sponsors, professional societies, and journals involved in scientific, engineering, and medical research will find this book an essential guide to the principles affecting research data in the digital age.

SUMMARY

Advances in digital computing, communications, sensors, and storage technologies are revolutionizing nearly every area of scientific, engineering, and medical research. Today, researchers are employing sophisticated technologies to generate, analyze, and share data to address questions that were unapproachable just a few years ago. They are carrying out detailed simulations to guide theoretical approaches and to validate new experimental approaches. They are working in interdisciplinary and often international teams on complex integrative problems that require inputs from a multitude of perspectives. They are using data generated by others to augment their own data and sometimes to address problems that the original researchers could not have envisioned. Digital technologies have fostered a new world of research characterized by immense datasets, unprecedented levels of openness among researchers, and new connections among researchers, policy makers, and the public.

Even as these new capabilities are expanding the power and reach of research, they are raising complex issues for researchers, research institutions, research sponsors, professional societies, and journals. Digital technologies can complicate the process of verifying the accuracy and validity of research data, in part because of the enormous rate at which data can be generated and the intricate processing those data undergo. The high rate of innovation in digital technologies, a lack of standards, and issues such as privacy, national security, and possible commercial interests can inhibit the sharing of data, which can reduce the ability of researchers to verify results and build on previous research. Huge increases in the quantity of data being generated, combined with the need to move digital data between successive storage media and software environments as technologies evolve, are creating severe challenges in preserving data for long-term use. And these issues are not restricted to large-scale research projects; they can be especially acute for the small-scale projects that continue to constitute the bulk of the research enterprise.

This report examines the consequences of the changes affecting research data with respect to three issues: integrity, accessibility, and stewardship. Because of the enormous range in the detailed procedures and styles of research from field to field, it is impossible to formulate specific recommendations for every field. Instead, for each of the three issues examined in this report, the authoring committee has developed a fundamental principle that applies in all fields of research regardless of the pace or nature of technological change. The report then explores the implications of these three central principles for the various components of the research enterprise.[1]

Developing the policies, standards, and infrastructure needed to ensure the integrity, accessibility, and stewardship of research data is a critically important task. It will require sustained effort on the part of all stakeholders in the research enterprise. The

[1] In this Summary, the principles appear in boldface type and the recommendations drawn from the principles are presented in italic type.

committee believes that the broad principles stated in this report provide the appropriate framework for this undertaking.

ENSURING THE INTEGRITY OF RESEARCH DATA

The fields of science, engineering, and medicine span the totality of physical, biological, and social phenomena. Research in all these fields is based on certain fundamental procedures and convictions. However, each research field has its own characteristic methods and scientific style. Consequently, research is too broad an enterprise to permit many generalizations about its conduct.

One theme, however, threads through its many fields: the primacy of scrupulously recorded data. Because the techniques that researchers employ to ensure the integrity—the truth and accuracy—of their data are as varied as the fields themselves, there are no universal procedures for achieving technical accuracy. The term "integrity of data" also has a structural meaning, related to the data's preservation and presentation. This is the subject of Chapter 4. There are, however, broadly accepted practices for generating and analyzing research. In most fields, for instance, experimental observations must be shown to be reproducible in order to be credible. Even this fundamental principle can have exceptions. For instance, observations with an historical element, such as the explosion of a supernova or the growth of an epidemic, cannot be reproduced. Other general practices include checking and rechecking data to confirm their accuracy and validity and submitting data and research results to peer review to ensure that the interpretation is valid. In addition, some practices may be employed only within specific fields, such as the use of double-blind clinical trials.

Many of the traditional methods for ensuring the integrity of data—whether universal or discipline specific—are being modified as digital technologies alter capabilities and procedures. Because of the huge quantities of data generated by digital technologies, an increasing fraction of the processing and communication of data is done by computers, sometimes with relatively little human oversight. If this processing is flawed or misunderstood, the conclusions can be erroneous. Documenting work flows, instruments, procedures, and measurements so that others can fully understand the context of data is a vital task, but this can be difficult and time-consuming. Furthermore, digital technologies can tempt those who are unaware of or dismissive of accepted practices in a particular research field to manipulate data inappropriately.

Several recent incidents and trends provided an impetus for this study, such as the challenge journals face in preventing inappropriate manipulation of digital images in submitted papers and well-publicized, albeit rare, cases of research misconduct involving fabricated or manipulated data. Assessing the broad set of institutions, policies, and practices that have been put into place to prevent and detect research misconduct, including the fabrication or inappropriate manipulation of data, was beyond the scope of this study. Nevertheless, the committee recognizes that the advance of digital technologies presents special challenges to the individuals and institutions charged with ensuring responsible conduct in research. Since these individuals and institutions will continue to play a critical role in ensuring the integrity of research data, it is important that they adapt their procedures in order to function effectively in the digital age.

The most effective method for ensuring the integrity of research data is to ensure high standards for openness and transparency. To the extent that data and other information integral to research results are provided to other experts, errors in data collection, analysis, and interpretation (intentional or unintentional) can be discovered and corrected. This requires that the methods and tools used to generate and manipulate the data be available to peers who have the background to understand that information.

The traditional way for submitting data and results to the scrutiny of other researchers is through peer review, which allows the validity of data and results to be judged for quality by a research community before dissemination. Although traditional peer review practices remain essential for evaluating the importance and validity of research, it has become clear that these have limitations when it comes to ensuring that digital data have been appropriately collected, analyzed, and interpreted. Fortunately, it has also become clear that the advance of digital technologies is providing new opportunities to ensure data integrity through greater openness and transparency. The emergence and growth of accessible databases such as GenBank and the Sloan Digital Sky Survey illustrate these opportunities in widely disparate disciplines.[2] Yet in many fields, a lack of technological infrastructure, cultural norms and expectations, and other factors act as barriers to openness and transparency.

The integrity of data in a time of revolutionary changes in research practice is too important to be taken for granted. Consequently, this report affirms the following general principle for ensuring the integrity of research data:

Data Integrity Principle: Ensuring the integrity of research data is essential for advancing scientific, engineering, and medical knowledge and for maintaining public trust in the research enterprise. Although other stakeholders in the research enterprise have important roles to play, researchers themselves are ultimately responsible for ensuring the integrity of research data.

This straightforward principle leads to several specific recommendations.

Recommendation 1: Researchers should design and manage their projects so as to ensure the integrity of research data, adhering to the professional standards that distinguish scientific, engineering, and medical research both as a whole and as their particular fields of specialization.

Some professional standards apply throughout research, such as the injunction never to falsify or fabricate data or plagiarize research results. These are fundamental to research, and have been confirmed by leading organizations and codified in regulations.[3] Other standards are relevant only within specific fields—such as requirements to conduct double-blind clinical trials. Researchers must adhere to both sets of standards if they are

[2] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. 2006. "GenBank." *Nucleic Acids Research* 34(Database):D16–D20. Available at http://nar.oxfordjournals.org/cgi/content/abstract/34/suppl_1/D16. See also Robert C. Kennicutt, Jr. 2007. "Sloan at five." *Nature* 450:488–489.

[3] National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. 1992. *Responsible Science: Ensuring the Integrity of the Research Process.* Washington, DC: National Academy Press.

to maintain the integrity of research data, and they can adhere to professional standards only if they fully understand the standards.

Recommendation 2: Research institutions should ensure that every researcher receives appropriate training in the responsible conduct of research, including the proper management of research data in general and within the researcher's field of specialization. Some research sponsors provide support for this training and for the development of training programs.

Researchers, research institutions, research sponsors, professional societies, and journals all are responsible for creating and sustaining an environment that supports the efforts of researchers to ensure the integrity of research data. In some cases, digital technologies are having such a dramatic effect on research practices that some professional standards affecting the integrity of research data either have not yet been established or are in flux. The recent recognition of the inappropriate manipulation of digital images submitted in journal articles illustrates the need for the research enterprise to continue to set clear expectations for appropriate behavior and effectively communicate those expectations.

Recommendation 3: The research enterprise and its stakeholders—research institutions, research sponsors, professional societies, journals, and individual researchers—should develop and disseminate professional standards for ensuring the integrity of research data and for ensuring adherence to these standards. In areas where standards differ between fields, it is important that differences be clearly defined and explained. Specific guidelines for data management may require reexamination and updating as technologies and research practices evolve.

Although all researchers should understand digital technologies well enough to be confident in the integrity of the data they generate, they cannot always be expected to be able to take full advantage of new capabilities. In an increasing number of fields, professionals with expertise specifically in the generation, analysis, storage, or dissemination of data are playing an essential role in taking advantage of digital technologies and ensuring the integrity of research data.

Recommendation 4: Research institutions, professional societies, and journals should ensure that the contributions of data professionals to research are appropriately recognized. In addition, research sponsors should acknowledge that financial support for data professionals is an appropriate component of research support in an increasing number of fields.

ENSURING ACCESS TO RESEARCH DATA

Advances in knowledge depend on the open flow of information. Only if data and research results are shared can other researchers check the accuracy of the data, verify analyses and conclusions, and build on previous work. Furthermore, openness enables the results of research to be incorporated into socially beneficial goods and services and into public policies, improving the quality of life and the welfare of society.

Despite the many benefits arising from the open availability of research data and results, many data are not publicly accessible, or their release is delayed, for a variety of reasons. Data may be withheld because they are being used to generate a commercial product or service, because of confidentiality considerations, or because of national security concerns. Furthermore, in some fields it is acceptable for researchers to have a limited period of exclusivity in which the data are used only by the principal investigators and their immediate associates. In areas of potential commercial applications, patenting considerations, contractual restrictions, and technological constraints also can limit or delay the accessibility of data.

Legitimate reasons may exist for keeping some data private or delaying their release, but the default assumption should be that research data, methods (including the techniques, procedures, and tools that have been used to collect, generate, or analyze data, such as models, computer code, and input data), and other information integral to a publicly reported result will be publicly accessible when results are reported, at no more than the cost of fulfilling a user request. This assumption underlies the following principle of accessibility:

Data Access and Sharing Principle: Research data, methods, and other information integral to publicly reported results should be publicly accessible.

Although this principle applies throughout research, in some cases the open dissemination of research data may not be possible or advisable. Granting access to research data prior to reporting results based on those data can undermine the incentives for generating the data. There might also be technical barriers, such as the sheer size of datasets, that make sharing problematic, or legal restrictions on sharing as discussed in Chapter 3. Nevertheless, the main objective of the research enterprise must be to implement policies and promote practices that allow this principle to be realized as fully as possible.

This principle has important implications for researchers.

Recommendation 5: All researchers should make research data, methods, and other information integral to their publicly reported results publicly accessible in a timely manner to allow verification of published findings and to enable other researchers to build on published results, except in unusual cases in which there are compelling reasons for not releasing data. In these cases, researchers should explain in a publicly accessible manner why the data are being withheld from release.

This principle may seem to apply only to publicly funded research, but a strong case can be made that much data from privately funded research should be made publicly available as well. Making such data available can produce societal benefits while also preserving the commercial opportunities that motivated the research.

As discussed earlier, differences in technological infrastructure, publication practices, data-sharing expectations, and other cultural practices have long existed between research fields. In some fields, aspects of this "data culture" act as barriers to access and sharing of data. With the growing importance of research results to certain areas of public policy, the rapid increase of interdisciplinary research that involves integration of data from different disciplines, and other trends, it is important for fields of research to examine their standards and practices regarding data and to make these explicit.

Data accessibility standards generally depend on the norms of scholarly communication within a field. In many fields these norms are now in a state of flux. In some fields, researchers may be expected to disseminate data and conclusions more rapidly than is possible through peer-reviewed publications. Digital technologies are providing new ways to disseminate research results—for example, by making it possible to post draft papers on archival sites or by employing software packages, databases, blogs, or other communications on personal or institutional Web sites.

Data sharing is greatly facilitated when a field of research has standards and institutions in place that are designed to promote the accessibility of data.

Recommendation 6: In research fields that currently lack standards for sharing research data, such standards should be developed through a process that involves researchers, research institutions, research sponsors, professional societies, journals, representatives of other research fields, and representatives of public interest organizations, as appropriate for each particular field.

If researchers are to make data accessible, they need to work in an environment that promotes data sharing and openness.

Recommendation 7: Research institutions, research sponsors, professional societies, and journals should promote the sharing of research data through such means as publication policies, public recognition of outstanding data-sharing efforts, and funding.

Recommendation 8: Research institutions should establish clear policies regarding the management of and access to research data and ensure that these policies are communicated to researchers. Institutional policies should cover the mutual responsibilities of researchers and the institution in cases in which access to data is requested or demanded by outside organizations or individuals.

PROMOTING THE STEWARDSHIP OF RESEARCH DATA

Research data can be valuable for many years after they are generated. Data that led to initial insights can sometimes be used to generate new findings in the same or entirely different research fields. Existing data can be reanalyzed or combined with new data to verify published results or arrive at new conclusions. In some research areas, accessible databases have become essential parts of the research infrastructure, comparable to laboratories, research facilities, and computing devices and networks.

Maintaining high-quality and reliable databases can be costly, especially over long time periods. Obviously not all data should be preserved, but deciding what to save and what to discard becomes more difficult as increasing quantities of data are generated. Because the future uses of data are difficult to predict, returns on investments in stewardship can be uncertain. Furthermore, in many fields of research, there is no consensus as to who should maintain large databases or who should bear the costs. These problems can be especially difficult for investigators involved in small projects, who can face great challenges in deciding which data will be useful, in documenting those data thoroughly for future uses, and in finding funds from limited budgets for data preservation.

The value of data for long-term use suggests the following general principle for the stewardship of data:

Data Stewardship Principle: Research data should be retained to serve future uses. Data that may have long-term value should be documented, referenced, and indexed so that others can find and use them accurately and appropriately.

Curating data requires documenting, referencing, and indexing the data so that they can be used accurately and appropriately in the future. Data stewardship must start at the beginning of the project, not partway through or at the end of the project.

Recommendation 9: Researchers should establish data management plans at the beginning of each research project that include appropriate provisions for the stewardship of research data.

Because data without accompanying information about how they were derived can be useless, arranging for preserved data to be annotated so that they retain their long-term value is among the most important tasks for researchers establishing a data management plan.

This recommendation is not meant to imply that individual researchers are responsible for ensuring indefinite preservation of their own data, but that they ensure that data that are judged to have potential long-term value are prepared and transferred to the appropriate archives or repositories. Researchers should work in partnership with their institutions, sponsors, and fields to formulate and implement their plans.

Researchers need to participate in the development of policies and standards for data annotation, preservation, and long-term access. Data need not be annotated in such detail that nonspecialists can immediately use them, but guidelines should exist for the

degree of expertise required to use a data collection. Researchers also need to develop procedures for error reporting, tracking, and correction. These policies and standards will vary greatly from field to field because they depend on the nature and potential uses of data. Nevertheless, establishing such policies is the collective responsibility of the researchers in each field.

Recommendation 10: As part of the development of standards for the management of digital data, research fields should develop guidelines for assessing the data being produced in that field and establish criteria for researchers about which data should be retained.

Researchers need a supportive institutional environment to fulfill their responsibilities toward the stewardship of data.

Recommendation 11: Research institutions and research sponsors should study the needs for data stewardship by the researchers they employ and support. Working with researchers and data professionals, they should develop, support, and implement plans for meeting those needs.

The problem of paying for long-term stewardship of research data and other digital scholarly work is difficult, and solutions need to be developed over time. It is important that requirements for improved data management practices not be imposed as unfunded mandates. In the digital age, data management needs to be integrated into research program funding as an essential component of the conduct of research. Where appropriate, grant applications should include costs for data stewardship.

Many issues regarding the integrity, accessibility, and stewardship of research data are common across the research enterprise. Bodies that oversee multiple fields of research should disseminate lessons learned and help to foster interdisciplinary cooperation. Within the U.S. federal government, a recent report by the Interagency Working Group on Digital Data explores the needs for preservation and dissemination of publicly funded research data.[4] At the nongovernmental level, the National Research Council recently established a new Board on Research Data and Information that will address emerging issues in the management, policy, and use of research data at the national and international levels.

[4] Interagency Working Group on Digital Data. 2009. *Harnessing the Power of Digital Data for Science and Society.* Washington, DC: National Science and Technology Council, Executive Office of the President.

Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age

Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age

Committee on Science, Engineering, and Public Policy

Advance Copy
Not For Public Release Before
July 22, 2009 at 11:00 AM EDT

NATIONAL ACADEMY OF SCIENCES,
NATIONAL ACADEMY OF ENGINEERING, *AND*
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

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

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

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

This study was supported by the National Research Council, United States Department of Agriculture, National Aeronautics and Astronautics Administration, United States Geological Survey, United States Department of Health and Human Services, United States Department of Energy, Eli Lilly and Company, Burroughs Wellcome Fund, Nature Publishing Group, The Rockefeller University Press, New England Journal of Medicine, American Chemical Society, Federation of American Societies for Experimental Biology, American Association for the Advancement of Science, American Geophysical Union and IEEE.

The material is based upon work supported by NASA under award #NNX07AP21G. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration.

This material is also based upon work supported by the Department of Energy [Office of Science] under Award Number DE-FG02-08ER15926. Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Cover graphic provided by Well-Formed.Eigenfactor (http://well-formed.eigenfactor.org/), a cooperation between Moritz Stefaner (visualization design) and the Eigenfactor Project (data analysis).

Additional copies of this report are available from the 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

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

COMMITTEE ON ENSURING THE UTILITY AND INTEGRITY OF RESEARCH DATA IN A DIGITAL AGE

DANIEL KLEPPNER (*Co-Chair*), Lester Wolfe Professor of Physics, Emeritus, Massachusetts Institute of Technology, Cambridge

PHILLIP A. SHARP (*Co-Chair*), Institute Professor, The David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge

MARGARET A. BERGER, Professor of Law, Brooklyn Law School, Brooklyn, New York

NORMAN M. BRADBURN, Tiffany & Margaret Blake Distinguished Service Professor Emeritus, University of Chicago, Washington, DC

JOHN BRAUMAN, J. G. Jackson–C. J. Wood Professor Emeritus, Department of Chemistry, Stanford University, Stanford, California

JENNIFER T. CHAYES, Managing Director, Microsoft Research New England, Cambridge, Massachusetts

ANITA JONES, Lawrence R. Quarles Professor of Engineering and Applied Sciences, University of Virginia, Charlottesville

LINDA P. B. KATEHI, Provost and Vice Chancellor for Academic Affairs, University of Illinois, Urbana-Champaign

NEAL F. LANE, Malcolm Gillis University Professor and Senior Fellow of the James A. Baker III Institute for Public Policy, Rice University, Houston, Texas

W. CARL LINBERGER, E.U. Condon Distinguished Professor of Chemistry and Fellow, Joint Institute for Laboratory Astrophysics, University of Colorado, Boulder

RICHARD LUCE, Vice Provost and Director of University Libraries, Robert W. Woodruff Library, Emory University, Atlanta, Georgia

THOMAS O. MCGARITY, Joe R. and Teresa Lozano Long Endowed Chair in Administrative Law, School of Law, University of Texas, Austin

STEVEN M. PAUL, Executive Vice President, S&T and President, Lilly Research Laboratories, Eli Lilly & Company, Indianapolis, Indiana

TERESA A. SULLIVAN, Provost and Executive Vice President for Academic Affairs and Professor of Sociology, University of Michigan, Ann Arbor

MICHAEL S. TURNER, Bruce V. Diana M. Rauner Distinguished Service Professor and Chair, Department of Astronomy and Astrophysics, University of Chicago, Chicago, Illinois

J. ANTHONY TYSON, Distinguished Professor of Physics, Department of Physics, University of California, Davis

STEVEN C. WOFSY, Abbott Lawrence Rotch Professor of Atmospheric and Environmental Sciences, Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts

Principal Project Staff

THOMAS ARRISON, Study Director (after July 2007)
DEBORAH D. STINE, Study Director (up to July 2007)
STEVE OLSON, Consultant Writer
NEERAJ P. GORKHALY, Senior Program Assistant
ALBERT SWISTON, Christine Mirzayan Science & Technology Policy Graduate Fellow
SAGE ARBOR, Christine Mirzayan Science & Technology Policy Graduate Fellow

COMMITTEE ON SCIENCE, ENGINEERING AND PUBLIC POLICY

GEORGE M. WHITESIDES (*Chair*), Woodford L. and Ann A. Flowers Professor of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts
CLAUDE R. CANIZARES, Vice President for Research, Associate Provost, Bruno Rossi Professor of Physics, Massachusetts Institute of Technology, Cambridge
RALPH J. CICERONE (Ex-officio), President, National Academy of Sciences, Washington, DC
EDWARD F. CRAWLEY, Professor of Aeronautics and Astronautics and of Engineering Systems, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge
RUTH A. DAVID, President and CEO of ANSER Institute for Homeland Security (Analytic Services, Inc.), Arlington, Virginia
HAILE T. DEBAS, Chancellor Emeritus, University of California, San Francisco
HARVEY FINEBERG (Ex-officio), President, Institute of Medicine, Washington, DC
JACQUES S. GANSLER, Roger C. Lipitz Chair in Public Policy and Private Enterprise, School of Public Policy, University of Maryland, College Park
ELSA M. GARMIRE, Sydney E. Junkins Professor of Engineering, Dartmouth College, Hanover, New Hampshire
M. R. C. GREENWOOD (Ex-officio), Chair, PGA, and Professor of Nutrition and Internal Medicine, University of California, Davis
W. CARL LINEBERGER, Professor of Chemistry, University of Colorado, Boulder
C. DAN MOTE, JR. (Ex-officio), President, University of Maryland, College Park
ROBERT M. NEREM, Professor and Director, Parker H. Petit Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta
LAWRENCE T. PAPAY, CEO and Principal, PQR, LLC, Maineville, Ohio
ANNE C. PETERSEN, Deputy Director, Center for Advanced Study in the Behavioral Sciences, Stanford University, Palo Alto, California
SUSAN C. SCRIMSHAW, Interim President, Sage Colleges, Troy, New York
WILLIAM J. SPENCER, Chairman Emeritus, SEMATECH, Austin, Texas
LYDIA THOMAS (Ex-officio), Co-Chair, GUIRR, and Chairman and CEO, Mitretek Systems, Falls Church, Virginia
CHARLES M. VEST (Ex-Officio), President, National Academy of Engineering, Washington, DC
NANCY S. WEXLER, Higgins Professor of Neuropsychology, Columbia University, New York, New York
MARY LOU ZOBACK, Vice President for Earthquake Risk Applications, Risk Management Solutions, Inc., Newark, California

Staff
RICHARD BISSELL, Executive Director
MARION RAMSEY, Administrative Associate
PETER HUNSBERGER, Financial Associate

Preface

Data are the foundation on which scientific, engineering, and medical knowledge is built. The generation, analysis, communication, and preservation of data are in a period of profound change, and research is being similarly transformed.

The development and rapid advance of digital technologies have enabled immense quantities of data to be created, processed, and disseminated around the world. These data can capture the characteristics of phenomena in far greater detail and with a dynamic verisimilitude never before possible. Data from different fields are being combined, yielding deep insights into formerly intractable problems. The open sharing of data, tools, and services over the internet is creating new ways of carrying out research and new relationships among researchers. New research topics and fields are emerging between the boundaries of traditional disciplines, and the questions that investigators can address are rapidly expanding.

These changes in the nature and conduct of research are greatly enhancing the capabilities of researchers. However, these changes also are posing challenges, and in some cases they have had negative consequences. A major impetus for this study was a letter sent from the editors of several leading journals to National Academy of Sciences President Ralph Cicerone (see Appendix C) pointing out that the improper manipulation of digital images submitted to scholarly journals has become a significant issue for editors and publishers. More broadly, changes in the use of research data have raised the stakes for the methods traditionally used to ensure the integrity and utility of data. Research data and results are increasingly critical inputs to a widening variety of policy debates and decisions. Transparency on the part of investigators with regard to the collection of data, methods of analysis, and presentation of results is essential for the research enterprise to serve the public as an objective source of unbiased information. In that regard, another major impetus for this report was the recent controversy over the interpretation and use of data to reconstruct historical changes in global temperatures. In this case, the combination of an important policy topic, differences in data-sharing expectations between fields, and unclear expectations among researchers and members of the public opened researchers to heightened scrutiny, skepticism, and even harassment.

As plans for this study took shape, it became clear that the issues involving research data extend well beyond the most immediate connotations of the term “data integrity.” Thus, the charge issued to our committee asked us to look at several critical issues:

An ad hoc committee will conduct a study of issues that have arisen from the evolution of practices in the collection, processing, oversight, publishing, ownership, accessing, and archiving of research data. The key questions to be addressed are:

1. **What are the growing varieties of research data? In addition to issues concerned with the direct products of research, what issues are involved in the treatment of raw data, prepublication data, materials, algorithms, and computer codes?**
2. **Who owns research data, particularly that which results from federally funded research? Is it the public? The research institution? The lab? The researcher?**
3. **To what extent is a scientist responsible for supplying research data to other scientists (including those who seek to reproduce the research) and to other**

parties who request them? Is a scientist responsible for supplying data, algorithms, and computer codes to other scientists who request them?

4. **What challenges do the science and technology community face arising from actions that would compromise the integrity of research data? What steps should be taken by the science and technology community, research institutions, journal publishers, and funders of research in response to these challenges?**
5. **What are the current standards for accessing and maintaining research data, and how should these evolve in the future? How might such standards differ for federally funded and privately funded research, and for research conducted in academia, government, nongovernmental organizations, and industry?**

The study will not address privacy issues and other issues related to human subjects.

At our committee's first meeting, it quickly became apparent that even this wide-ranging charge did not encompass the full range of pressing issues involving research data. Digital technologies have been changing research at a pace that would have been hard to predict even a decade ago. Practices and expectations for data sharing vary considerably from field to field and are rapidly evolving. National and homeland security concerns affect the policy environment governing access to various types of data. In some areas the costs of maintaining collections and transferring them to new digital media raise questions about who is responsible for undertaking and financing long-term stewardship. A growing variety of investigators and research fields face difficult choices involving trade-offs between sustaining existing data collections and performing new research.

The purpose of this report is to explore the evolving roles and responsibilities of researchers, research institutions, research sponsors, journals, publishers, and others in generating, analyzing, disseminating, and preserving research data. Many of the methods used to validate the quality of data, make data available to other researchers, and preserve data for future uses are unique to specific disciplines. Focusing on these discipline-specific methods would yield a report that is both too narrow and too transitory given the transformative influence of rapidly changing technologies.

Instead, we decided to base our report on the broad principles that have characterized science and engineering research for hundreds of years and will continue to do so in the future. In particular, we decided to focus on three broad and intertwined issues that we have characterized as integrity, access, and stewardship. For each of these issues, we state a general principle that applies throughout the research enterprise. We then use these three broad principles to formulate recommendations that apply in more specific circumstances. We have also highlighted, within the text and in sidebars in each chapter, useful efforts by researchers, institutions, research fields, research sponsors, professional societies, and journals to facilitate the realization of our broad objectives. And we have identified issues——some new and some old——that will need continued attention as technology continues to reshape the research enterprise.

Although this report addresses all of the components of the research enterprise, its primary focus is on the roles and responsibilities of the investigator. This is appropriate, given the composition of the committee and the nature of the task. The actions of researchers inevitably influence all the other parts of the research enterprise, and each of

these parts also has responsibilities in maintaining the integrity, accessibility, and stewardship of research data. However, researchers must take the lead in addressing new and pressing issues involving research data. In general, the report attempts to reflect the perspectives of individual researchers in different fields with respect to the generation, preservation, and sharing of research data in science as a whole and in specific fields.

Following the Executive Summary, Chapter 1 introduces the main issues covered in the report by examining the terms used in the report and the varieties of research data. Chapter 2, on the integrity of research data, looks at the challenges to data integrity created by rapidly changing technologies and at responses to those challenges. Chapter 3 discusses the responsibility for researchers to make publicly available the data on which research results are based, and the variety of challenges this poses in different fields and settings. And Chapter 4 describes the long-term value of research data and methods to preserve data for future uses.

The changes in the daily practices and activities of researchers due to the rapidly changing technologies provide a unique opportunity to reinforce and extend the traditional openness and collaborative nature of science.

In preparing this report, our committee has taken advantage of a number of studies by the National Academy of Sciences, the National Academy of Engineering, the Institute of Medicine and the National Research Council. Appendix B provides a list of recent reports on relevant subjects. For example, the committee spent some time reviewing and discussing a recent controversy over the interpretation and use of data to reconstruct historical changes in global temperatures, as described in the 2006 NRC report *Surface Temperature Reconstruction for the Last 2,000 Years.*

The importance of data in research and in societal decisions will continue to increase as science and engineering exert an ever greater influence on society and as digital technologies continue to remake our world. The committee and the members of the Committee on Science, Engineering, and Public Policy hope and trust that this report will stimulate further dialogue to strengthen science and engineering in a data-rich world.

Phillip A. Sharp
Massachusetts Institute of Technology

Daniel Kleppner
Massachusetts Institute of Technology

Acknowledgment of Reviewers

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 National Academies' Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published 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 process.

We wish to thank the following individuals for their review of this report: Frederick Anderson, McKenna, Long & Aldridge LLP; Michael Carroll, American University; Ian Foster, Argonne National Laboratory; John Graham, Indiana University; Myron Gutmann, Inter-University Consortium for Political and Social Research; Henry Horbaczewski, Reed Elsevier Inc.; Jerome Kassirer, Tufts University; Michael Keller, Stanford University; Joan Lippincott, Coalition for Networked Information; David Moorman, Social Sciences and Humanities Research Council of Canada; James Ostell, National Library of Medicine; Robert Pike, Google; David Robinson, Rutgers University; Sandford Shattil, University of California, San Diego; and John White, University of Arkansas.

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 William Press, University of Texas, Austin and Warren Washington, National Center for Atmospheric Research. Appointed by the National Academies, they were responsible for making certain that an 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 this report rests entirely with the authoring committee and the institution.

Contents

SUMMARY 1

1 RESEARCH DATA IN THE DIGITAL AGE 9
Challenges Posed by Research Data in a Digital Age, 13
Descriptions of Terms Used in the Report, 14
The Varieties of Research Data, 18
Structure of the Report, 20

2 ENSURING THE INTEGRITY OF RESEARCH DATA 29
The Roles of Data Producers, Providers, and Users, 34
The Collective Scrutiny of Research Data and Results, 35
Peer Review and Other Means for Ensuring the Integrity of Data, 37
Data Integrity in the Digital Age and the Role of Data Professionals, 44
General Principle for Ensuring the Integrity of Research Data, 44
The Obligations of Researchers to Ensure the Integrity of Research Data, 45
The Importance of Training, 46
Producing Clear, Up-to-Date Standards for Data Integrity:
A Shared Responsibility of the Research Enterprise, 47
The Roles of Data Professionals, 48

3 ENSURING ACCESS TO RESEARCH DATA 55
Barriers to Sharing Data 55
The Costs of Limiting Access to Data, 61
Data Access Issues in Research Affecting Public Policy or Private Interests, 61
Ownership of Research Data and Related Products, 63
Legal and Policy Requirements for Access to Data, 69
The International Dimensions of Access to Research Data,72
General Principle for Enhancing Access to Research Data, 73
Responsibilities of Researchers, 75
Responsibilities of Research Fields, 76
Responsibilities of Research Institutions, Research Sponsors, Professional Societies, and Journals, 77

4 PROMOTING THE STEWARDSHIP OF RESEARCH DATA 85
The Loss and Underutilization of Research Data, 86
Infrastructure and Incentives for the Stewardship of Data, 88
Annotating Data for Long-Term Use, 93
Fostering Data Stewardship for the Broad Research Enterprise, 94

General Principle for Enhancing the Stewardship of Research Data, 95
Responsibilities of Researchers, 96
Responsibilities of Research Institutions, Research Sponsors, and Journals, 98

5 DEFINING ROLES AND RESPONSIBILITIES 103
Assigning Roles and Responsibilities, 103
Researchers, 104
Research Institutions, 105
Research Sponsors, 105
Professional Societies and Journals, 105
Conclusion, 105

APPENDIXES
A Biographical information on the Committee Members 107
B Relevant National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and National Research Council Reports 117
C Letters from Journals 125