



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: 0-309-13685-7, 180 pages, 6 x 9, (2009)

This free PDF was downloaded from:
<http://www.nap.edu/catalog/12615.html>

Visit the [National Academies Press](#) online, the authoritative source for all books from the [National Academy of Sciences](#), the [National Academy of Engineering](#), the [Institute of Medicine](#), and the [National Research Council](#):

- Download hundreds of free books in PDF
- Read thousands of books online, free
- Sign up to be notified when new books are published
- Purchase printed books
- Purchase PDFs
- Explore with our innovative research tools

Thank you for downloading this free PDF. If you have comments, questions or just want more information about the books published by the National Academies Press, you may contact our customer service department toll-free at 888-624-8373, [visit us online](#), or send an email to comments@nap.edu.

This free book plus thousands more books are available at <http://www.nap.edu>.

Copyright © National Academy of Sciences. Permission is granted for this material to be shared for noncommercial, educational purposes, provided that this notice appears on the reproduced materials, the Web address of the online, full authoritative version is retained, and copies are not altered. To disseminate otherwise or to republish requires written permission from the National Academies Press.

**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

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.

Library of Congress Cataloging-in-Publication Data

Committee on Science, Engineering, and Public Policy (U.S.). Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age.

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.

p. cm.

Includes bibliographical references and index.

ISBN-13: 978-0-309-13684-6 (pbk.); ISBN-10: 0-309-13684-9 (pbk.);

ISBN-13: 978-0-309-13685-3 (pdf); ISBN-10: 0-309-13685-7 (pdf)

1. Research—Technological innovations. 2. Information technology—Scientific applications.
3. Electronic information resources—Management—United States. 4. Electronic information resources—Access control. I. Title.

Q180.55.I45C66 2009

001.40285'58—dc22

2009036322

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>.

Copyright 2009 by the National Academy of Sciences. All rights reserved.

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, School 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

WILLIAM OSTENDORFF, 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; Sanford 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	11
Challenges Posed by Research Data in a Digital Age, 19	
Descriptions of Terms Used in the Report, 22	
The Varieties of Research Data, 27	
Structure of the Report, 29	
2 ENSURING THE INTEGRITY OF RESEARCH DATA	33
The Roles of Data Producers, Providers, and Users, 40	
The Collective Scrutiny of Research Data and Results, 41	
Peer Review and Other Means for Ensuring the Integrity of Data, 43	
Data Integrity in the Digital Age and the Role of Data Professionals, 50	
General Principle for Ensuring the Integrity of Research Data, 51	
The Obligations of Researchers to Ensure the Integrity of Research Data, 51	
The Importance of Training, 54	
Producing Clear, Up-to-Date Standards for Data Integrity: A Shared Responsibility of the Research Enterprise, 56	
The Roles of Data Professionals, 57	
3 ENSURING ACCESS TO RESEARCH DATA	59
Barriers to Sharing Data, 63	
The Costs of Limiting Access to Data, 70	
Data Access Issues in Research Affecting Public Policy or Private Interests, 71	
Ownership of Research Data and Related Products, 73	

Legal and Policy Requirements for Access to Data, 80	
The International Dimensions of Access to Research Data, 83	
General Principle for Enhancing Access to Research Data, 84	
Responsibilities of Researchers, 86	
Responsibilities of Research Fields, 88	
Responsibilities of Research Institutions, Research Sponsors, Professional Societies, and Journals, 90	
4 PROMOTING THE STEWARDSHIP OF RESEARCH DATA	95
The Loss and Underutilization of Research Data, 96	
Infrastructure and Incentives for the Stewardship of Data, 99	
Annotating Data for Long-Term Use, 106	
Fostering Data Stewardship for the Broad Research Enterprise, 107	
General Principle for Enhancing the Stewardship of Research Data, 109	
Responsibilities of Researchers, 109	
Responsibilities of Research Institutions, Research Sponsors, and Journals, 112	
5 DEFINING ROLES AND RESPONSIBILITIES	115
Assigning Roles and Responsibilities, 115	
Researchers, 115	
Research Institutions, 118	
Research Sponsors, 119	
Professional Societies and Journals, 119	
Conclusion, 120	
APPENDIXES	
A Biographical Information on the Committee Members	121
B Relevant National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and National Research Council Reports	133
C Letters from Journals	143
INDEX	155

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 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

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

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

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.

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 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

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.

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.

1

Research Data in the Digital Age

In a 1965 article in *Electronics Magazine*, Gordon Moore, the cofounder of Intel, observed that the number of components on an integrated circuit per unit of cost was doubling on a regular basis—a period he later set at 2 years.¹ What came to be known as Moore’s law has become a defining property of the digital age.² For more than half a century, the power of computing available at a given cost has risen exponentially, which has increased computer power by many orders of magnitude. Today, the most powerful computers can perform more than a million billion operations per second. Storage devices can handle petabytes of information.³ Data can be transferred at rates of 10 gigabits (or 10 billion bits) per second (see Box 1-1 for a description of units of size for data). Sensors such as the charged-coupled devices used in modern cameras and telescopes can acquire data from billions of pixels simultaneously. Furthermore, in key areas of computing, Moore’s law continues to hold.⁴ Many measures of computing power continue to double every 1 to 2 years. As a result, the quan-

¹ Gordon E. Moore. 1965. “Cramming more components onto integrated circuits.” *Electronics* 38(19):114–117.

² Michael S. Turner. 2007. “Scientific discovery in the Information Age.” Presentation at the De Lange Conference on Emerging Libraries: How Knowledge Will Be Accessed, Discovered, and Disseminated in the Age of Digital Information, March 6, Houston, TX. Available online at <http://delange.rice.edu/VI/EL/Turner-DeLange-2007.pdf?action=details&event=921>.

³ A petabyte represents a million billion characters, the equivalent of the text in one billion books.

⁴ Not all measures of computing power are increasing exponentially. For example, the transfer rate of data within computers from memory devices to the central processing unit is growing slowly and at a linear rate. Physical limitations on the power of single processors have constrained the continued general application of Moore’s law. However, new algorithms for processors and storage units linked in parallel may lead to resumed exponential increases in computing power in the future.

BOX 1-1 Units of Size for Data

Bit: The fundamental unit of digital information, equivalent to a 1 or a 0, or to an electronic switch being on or off. Bit is short for *binary digit*.

Byte: The information stored in eight bits. A byte can be used to store one character of English text.

Kilobyte: The information stored in approximately 1,000 bytes, which is the equivalent of about 15 lines of text.

Megabyte: The information stored in approximately 1,000 kilobytes. A large novel contains about a megabyte of information, and a standard compact disc holds about 680 megabytes of digital information.

Gigabyte: The information stored in approximately 1,000 megabytes. A typical hard drive (as of 2008) holds about 500 gigabytes of information.

Terabyte: The information stored in approximately 1,000 gigabytes. The printed information stored in the Library of Congress equals approximately 10 terabytes.

Petabyte: The information stored in approximately 1,000 terabytes. All U.S. academic research libraries combined contain about 2 petabytes of information.

Exabyte: The information stored in approximately 1,000 petabytes. According to one estimate,^a human beings have spoken about 5 exabytes of words over the course of our species' history.

^a See <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>.

tity of data being created and stored by businesses, individuals, government, scientific institutions, and individuals is growing rapidly. Figure 1-1 shows one consulting firm's projection of how information and available storage will grow in the coming years.

This exponential increase in computing power has had profound consequences for many aspects of modern society, including scientific, engineering, and medical research.⁵ Using digital technologies, researchers can measure, describe, and model phenomena much more comprehensively and in far greater detail than was possible in the past. They can detect and analyze the products

⁵ Alexander Szalay and Jim Gray. 2006. "Science in an exponential world." *Nature* 440:413–414.

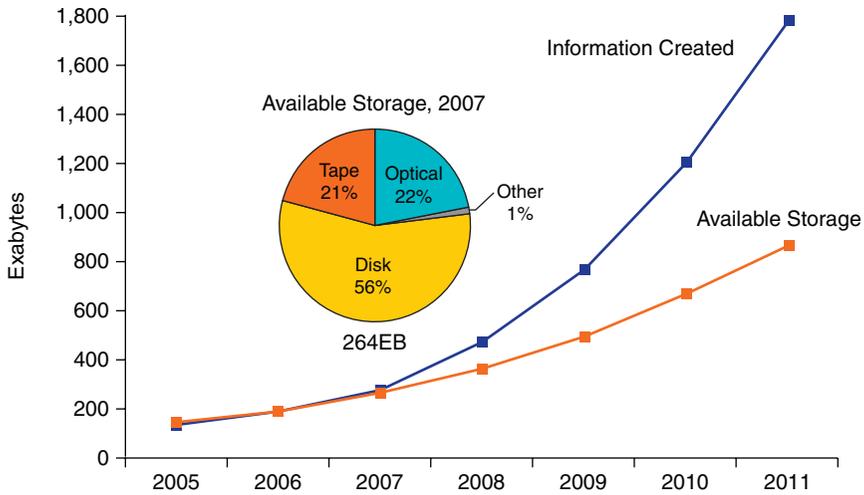


FIGURE 1-1 Projected global information creation and available storage.

NOTE: One exabyte equals one billion gigabytes.

SOURCE: IDC White Paper sponsored by EMC, *The Diverse and Exploding Digital Universe*, March 2008. Available at: <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>.

of high-energy particle collisions to probe the underlying structure of matter. They can extract information about the functioning of nerve cells and construct models of neural processing. They can combine simultaneous measurements of atmospheric and oceanic conditions to predict the effects of pollutants on climates. They can extract patterns of health from extensive databases of genetic and medical records. Examples of the impact of digital technologies on research fields appear as sidebars throughout this report, and the number of such examples could be multiplied many times.

The advances in digital technologies have caused a massive increase in the quantity of data generated by research projects. The proposed Large Synoptic Survey Telescope is expected to gather 30 terabytes of data per night and more than 60 petabytes over its lifetime (see Box 1-2). Particle physics experiments conducted with the Large Hadron Collider at CERN (Figure 1-2) will generate 15 petabytes of data annually. Even relatively small-scale projects can generate immense quantities of data that can be valuable in multiple research fields. These quantities of data are much too large to examine by hand. Instead, computers must conduct the initial analysis of data before the processed and condensed results are reviewed by researchers.

BOX 1-2 Digital Data in Astronomy

As astronomical observatories have become more powerful, they also have become more data-intensive.^a Table 1-1 shows the trend in recent decades. The Sloan Digital Sky Survey (SDSS), for example, has delivered an unprecedented flood of data since it began operation in 2000. The SDSS uses a dedicated 2.5-meter telescope on Apache Point, New Mexico, equipped with two special-purpose instruments. The telescope's camera can image 1.5 square degrees of sky at a time—about eight times the area of the full moon. A pair of spectrographs can measure spectra of—and hence

TABLE 1-1 Data Trends in Astronomy Research

Cosmic Microwave Background (CMB) Surveys: Collect information used to understand the origin and evolution of the universe		
Year	Survey	Data items (pixels)
1990	Cosmic Background Explorer (COBE)	1,000
2000	Boomerang (balloon-borne millimeter-wave telescope)	10,000
2002	Cosmic Background Imager (CBI)	50,000
2003	Wilkinson Microwave Anisotropy Probe (WMAP)	1,000,000
2009	Planck	10,000,000
Galaxy Surveys: Collect two dimensional optical images of galaxies and quasars		
Year	Survey	Objects
1970	Lick Observatory	1,000,000
1990	Automatic Plate Measuring Facility (APM)	2,000,000
2005	Sloan Digital Sky Survey (SDSS)	200,000,000
2009	Visible and Infrared Telescope for Astronomy (VISTA)	1,000,000,000
2015	Large Synoptic Survey Telescope (LSST) ¹	20,000,000,000
Galaxy Redshift Surveys: Collect three dimensional optical catalogs of galaxies and quasars		
Year	Survey	Objects
1986	Center for Astrophysics (CfA)	3,500
1996	Las Campanas Redshift Survey (LCRS)	23,000
2003	2dF Galaxy Redshift Survey	250,000
2005	Sloan Digital Sky Survey (SDSS)	750,000
2007	SDSS color-redshift survey	20,000,000
2015	LSST color-redshift survey	4,000,000,000

NOTE: There are 100 billion galaxies in the observable universe, meaning that LSST will record about 20 percent.

Source: Presentation to the committee by Alex Szalay, Johns Hopkins University, December 2007, updated in 2008 with comments by Tony Tyson and Michael Turner.

distances to—more than 600 galaxies and quasars in a single observation. A custom-designed set of software pipelines keeps pace with the enormous data flow from the telescope.

In its first 5 years of operation, the Sloan telescope searched more than 8,000 square degrees of the northern sky—about a fifth of the entire sky—in five wavelength bands. It recorded some 217 million objects, mostly galaxies, stars, and asteroids, and measured spectra for around 675,000 of these.^b

With funding from multiple sources and countries, the SDSS has followed a policy of freely releasing data annually, with separate Web sites for research users and the general public. A recent release, Data Release 7 (DR7), in November 2008, included some 16 terabytes of images and spectra. Its current phase, SDSS-II, is among the largest astronomical collaborations ever undertaken, involving more than 300 astronomers, astrophysicists, and engineers at 25 institutions around the world.

The SDSS has helped to revolutionize the interactions between a telescope, its data, and its user communities. Because the SDSS data archive is available to any astronomer, roughly half of the 2,100 refereed papers based on SDSS data have come from authors outside the project itself, and that proportion is rising. In fact, for the past 2 years, the SDSS has produced the most high-impact papers of any astronomical observatory.^c At the same time, the project has extended the “reach” of those wishing to participate in frontier astronomy research or to simply enjoy the ability to “be there” as amateur aficionados. The public is offered both the raw data of SDSS and, at a “SkyServer” Web site, a range of search tools to help them use the data. Teachers are encouraged to adapt the projects for use in the classroom. SDSS data also are available through the National Virtual Observatory (<http://www.us-vo.org>), a collaborative effort involving universities, supercomputer centers, observatories, and data repositories.^d

Even bigger projects are under development. For example, the Large Synoptic Survey Telescope (<http://www.lsst.org>) that is currently being developed will generate as much data *each night* as a complete SDSS. As the “Living LSST Document, Version 1.0, of May 15, 2008” put it:

LSST has been conceived as a public facility: The database it will produce, and the associated object catalogs that are generated from that database, will be made available to the world’s astronomical research community and to the public at large with no proprietary period. The software which created the LSST database will be open source. LSST will be a significant milestone in the globalization of the information revolution. LSST will put terabytes of data each night into the hands of anyone who wants to explore it, and in some sense will become an Internet telescope: the ultimate network peripheral device to explore the universe, and a shared resource for all humanity.

^a Alexander Szalay and Jim Gray. 2001. “The world-wide telescope.” *Science* 293:2037–2040.

^b Robert C. Kennicutt, Jr. 2007. “Sloan at five.” *Nature* 450:488–489.

^c J. P. Madrid and F. D. Macchetto. 2006. “High-impact astronomical observatories.” *Bulletin of the American Astronomical Society* 38:1286–1287.

^d Alexander Szalay, Johns Hopkins University, presentation to the committee, December 10, 2008.



FIGURE 1-2 LHC at CERN.

SOURCE: © CERN. See <http://cdsweb.cern.ch/record/42370>.

However, the most consequential changes being fostered by digital technologies involve issues that range beyond the quantities of data generated.⁶ Today, researchers can access a rapidly expanding range of digital information from around the world almost instantaneously. They can use this information to analyze their results, as when biologists compare DNA sequences they have generated to sequences stored in worldwide databases. They can incorporate information from others with their own data to make discoveries that would otherwise have been impossible, as when epidemiologists combine census and economic data to analyze the prevalence of disease. They can analyze data produced by others to answer questions that could not have been anticipated by the data's creators, as when astronomers use digital sky surveys to investigate newly recognized phenomena in distant galaxies. For some areas of science, engineering, and medical research in the digital age, carrying out laboratory experiments to corroborate or disprove hypotheses has given way to a process of hypothesis testing based on computational analysis and modeling.

The creation of inexpensive, complex sensors is contributing to the data explosion by enabling new research approaches in a variety of fields, particularly in the earth sciences. Projects such as the National Science Foundation's Network for Earthquake Engineering Simulation and National Ecological Observa-

⁶ National Research Council. 2001. *Issues for Science and Engineering Researchers in the Digital Age*. Washington, DC: The National Academies Press.

tory Network, as well as the National Aeronautics and Space Administration's Earth Observing System, depend heavily on sensor networks.

Digital technologies also are making possible a new kind of science that depends on simulations combined with experimentation and observation.⁷ Cosmologists can combine simulations of galactic dynamics with astronomical observations of distant galaxies to analyze the early evolution of the universe. Records of calls made with cell phones can be compared to mathematical models of social networks. Researchers can model the functions of cells, simulate the effects of modifying those functions, and then re-create these modifications in real cells to alter biological function and refine the original models. Large-scale simulations of natural phenomena can be as valuable as data drawn from observations of the natural world.

The advances in research enabled by high-performance computing and high-performance communications are contributing to a steady growth of collaborations and interdisciplinary projects. Digital communication technologies enable researchers to communicate and exchange data with colleagues around the world, creating electronic collaborations that can catalyze progress. By making it possible to address more complex and integrative questions, these technologies also catalyze interdisciplinary collaboration. As one indicator of this trend, consider the growth in the number of authors on research papers over time. Over the course of 40 years, according to a computerized analysis of millions of published science and engineering papers, the number of authors for papers in the sciences nearly doubled, from 1.9 to 3.5.⁸ In the environmental sciences, the fraction of papers with multiple authors rose from 25 percent to 82 percent; in economics, it rose from 9 percent to 52 percent.

Collaborations have also become more international. In 2003, 20 percent of all research publications had authors from more than one country, compared with 8 percent in 1988.⁹ Citations to literature produced outside the author's home country rose from 42 percent of all citations in 1992 to 48 percent in 2003.

However, the most far-reaching effects of digital technologies are not evident in traditional measures of research collaboration. Researchers—and especially young researchers—are developing new ways to interact with each other and with the subjects they study.¹⁰ They exchange information in virtual

⁷ The 2020 Science Group. 2006. *Towards 2020 Science*. Redmond, WA: Microsoft Corporation. Available at http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/downloads/T2020S_ReportA4.pdf.

⁸ Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. 2007. "The increasing dominance of teams in production of knowledge." *Science* 316:1036–1039.

⁹ National Science Board. 2006. *Science and Engineering Indicators 2006*. Arlington, VA: National Science Foundation.

¹⁰ Carolyn Y. Johnson. 2008. "Out in the open: Some scientists sharing results." *The Boston Globe*, August 21, p. A1.

communities, write and read blogs on research developments, and are pioneering new methods to conduct research and share their results. In the long run, these developments are likely to have a more profound effect on research than increases in the pace or scale of traditional practices. These developments can be difficult to foresee. For example, research in many fields is moving toward much more open and collaborative models that are both served and driven by technology, and this trend is likely to result in research environments very different from those that have prevailed in the past. Although our committee has not tried to predict the long-term outcomes of this process, ongoing changes can be expected to continue to transform how research is done and how researchers interact with each other.

The rapid spread of digital technologies also is transforming the relationship between researchers and the broader public that supports and expects to benefit from research. When research results that underlie important public policies are available electronically, they can be examined and questioned by any member of the public. Individuals interested in specific issues—whether the regulation of an environmental toxin or the development of therapies for a human disease—can monitor, comment on, and even shape ongoing research.

Similarly, digital technologies have profound implications for scientific, engineering, and medical education.¹¹ Students can have access to research information from instruments in distant locations.¹² Computer owners around the world can contribute to the solution of particular research problems by allowing their computers to become parts of distributed computational networks.¹³ Data from cutting-edge research are being made available on the Internet for use not only by the research community but by educators or anyone else interested in the subject.¹⁴ Members of the public are participating in research projects as varied as analyses of genetic variation and galactic structure.¹⁵ Although fascinating, the full consequences of changing technologies for scientific, engineering, and medical education or for direct public participation in research lie outside the scope of this report.

¹¹ National Research Council. 2002. *Preparing for the Revolution: Information Technology and the Future of the Research University*. Washington, DC: The National Academies Press.

¹² An example is the Education and Outreach Project of the National Virtual Observatory (<http://www.virtualobservatory.org>).

¹³ An example is the SETI@home project (<http://setiathome.berkeley.edu>), which uses computer time provided by volunteers to analyze astronomical data for signs of intelligence.

¹⁴ Ryan Scranton, Andrew Connolly, Simon Krughoff, Jeremy Brewer, Alberto Conti, Carol Christian, Craig Sosin, Greg Coombe, and Paul Heckbert. 2007. "Sky in Google Earth: The next frontier in astronomical data discovery and visualization." Available at http://arxiv.org/PS_cache/arxiv/pdf/0709/0709.0752v2.pdf.

¹⁵ For the analysis of genetic variation, see <https://www3.nationalgeographic.com/genographic>. For the analysis of galactic structure, see <http://www.galaxyzoo.org>.

CHALLENGES POSED BY RESEARCH DATA IN A DIGITAL AGE

Rapid advances in computing and communication technologies have changed the professional responsibilities, interpersonal interactions, and daily practices of researchers. Many of these changes have strengthened the research enterprise, both by enabling researchers to ask new questions of nature and by providing new means of achieving research objectives. At the same time, some changes have raised important issues involving researchers, research institutions, sponsors, and journals.¹⁶ These issues are the focus of this report on the integrity, accessibility, and stewardship of research data.

As discussed in Chapter 2, although advances in digital technologies allow phenomena and objects to be described more comprehensively and accurately, they also can complicate the process of verifying the accuracy and validity of the data (see Box 1-3 for an example). Digital technologies require the translation of phenomena and objects into digital representations, which can introduce inaccuracies into the data. Digital data often undergo several layers of complex processing as they move from an instrument or sensor to the point of being reviewed by a researcher. If this processing is not properly done or is misunderstood, the results can be misleading. In some cases, researchers may intentionally or unintentionally distort data in a misguided attempt to emphasize particular features and downplay others. In the worst cases, researchers can falsify or fabricate data, thereby violating both the ethical and methodological standards of research integrity. Many of these considerations apply as well to data that are not generated or stored digitally, but digital technologies both expand and intensify the challenge of maintaining the integrity of data.

Chapter 3 describes the challenges that researchers face in maintaining the traditional openness of research in a digital age. Electronic technologies provide researchers with many new ways of communicating data to others, but providing other researchers with access to large databases can be difficult and expensive. With smaller, heterogeneous databases, where quality control and documentation tend to be less formal, sociological and technological factors can restrict data sharing. Also, an increasing range of restrictions are being placed on research data as this information becomes more valuable for commercial uses, which can limit the distribution and utilization of data within and beyond the research community.

Even as more research data are being created, their value for future uses is increasing. Chapter 4 describes the need to preserve many research data for long-term use, even in situations where those uses cannot be currently envisioned. Digital storage technologies, application environments, and operating systems change every few years, which means that digital bits must continually

¹⁶ National Research Council. 2001. *Issues for Science and Engineering Researchers in the Digital Age*, Washington, DC: National Academy Press.

BOX 1-3 Digital Data in the Neurosciences

The neurosciences illustrate both the potential value of well-organized and accessible data and the variety of issues raised by the increased importance of data handling and data sharing.

It is not surprising that the neurosciences are rich in the use of and need for data, given the complexity of the nervous system. The brain has roughly a hundred billion neurons and more than 1,000 subdivisions, each with different structures and circuitry. In the past, neurological research has depended heavily on autopsy for clues about function and structure. Now it relies heavily on *in vivo* imaging methods and computational models, both of which depend on computing power and mathematical techniques.

This new universe of neuroscience data is too vast and complex for manual analysis. Large-scale detailed maps of the brain can require some 25 gigabytes of memory per image. Also, neuroscientists must work across multiple scales of resolution because they do not yet know which levels are critical for many neurological processes. They must integrate such diverse datasets as cellular neuroimaging, gene expression data, genotype data, neuronal morphology, and clinical data.

Making neuroscience data widely available holds tremendous potential for helping science and society. This includes:

- Facilitating replication and validation of experimental results,
- Promoting collective analyses of large numbers of experiments by different groups,
- Improving communication within and between groups, and
- Promoting collaboration.

Several very effective databases have been developed in the neurosciences. They include:

- The Cell Centered Database, started in 2002, makes two-dimensional and three-dimensional static and dynamic microscopic data available to the research community. It also links data obtained at cellular and subcellular scales to molecular and higher order structure. It is built on the Biomedical Informatics Research Network and Telescience grid infrastructure for distributed collaboration.
- SumsDB is a repository of brain-mapping data, including surfaces and volumes, with both structural and functional data. It includes more than 500 studies on monkeys, rodents, apes, humans, and others, totaling about 10 percent of the published literature. It also includes a data mining tool called WebCaret so that SumsDB

can be searched online without downloading. Its designers have made attempts to provide metadata and show the source of data, including links to online publications.

Many questions have arisen in developing these and other databases. Which digital data and data stored on film need to be stored? Do calibrations (i.e., the characterization of an instrument's response to known stimulus) need to be stored, and if so which ones? Should proprietary tools be stored so that users can see how the primary data were processed? For now, there is reason to err on the side of depositing too much data, because no one knows what subsequent researchers will need. However, it is likely that just a small percentage of databases will find widespread use, which complicates, rather than simplifies, the task of storage.

Complex databases always include errors. Obvious errors, such as coordinates that lie outside the brain, can be found more easily when data are shared. However, policing data before they are added to a database can be so time-intensive that it can discourage database building. Fortunately, new technologies for assuring the quality of data based on advances in such areas as pattern recognition and learning theory, combined with rapid advances in data processing and storage, are providing new and automated methods for testing the quality of data.

Another problem is that most data assigned to databases in the neurosciences are not adequately annotated, and even those with annotation tend to use nonstandard terminology, making them "islands" of diverse resources. Such databases may not be useful for comparative studies or other purposes.

Issues of who has rights to use data also are far from resolved. A researcher may work for 5 years to assemble data on a transgenic mouse and be reluctant to give the data away. To make data open and accessible, incentives may need to be developed to encourage scientists to share their data.

Another issue is whether journals may be responsible for receiving and storing all primary or supplementary data. Most publications lack a suitable place to enter and store supplementary data, and who should pay for this service remains unresolved.

These issues, most of which we discuss later in this report, are being extensively explored in the research and policy-making communities. Many questions do not yet have clear answers that extend across all research disciplines.

SOURCE: This box draws on presentations to the committee by David Van Essen, Washington University in St. Louis, and Maryann Martone, University of California, San Diego on December 10, 2007.

be transferred from one storage platform and software environment to another if they are not to be lost. Digital data also need to be annotated in sufficient detail that future researchers, sometimes in fields well removed from those of the data's original creators, can both use the data and understand their limitations. Maintaining data collections for long-term use thus requires continued investment and planning, which can compete with expenditures for ongoing research.

DESCRIPTIONS OF TERMS USED IN THE REPORT

In describing issues as broad as those covered in this report, it is essential to have clear understanding of the basic terms.

Research Data

Despite the importance of research data, there exists no standard or widely accepted definition of exactly what research data are. For the purposes of this report, we have treated data as *information used in scientific, engineering, and medical research as inputs to generate research conclusions* (see Box 1-4 for definitions from other reports). This usage encompasses a wide variety of information. It includes textual information, numeric information, instrumental readouts, equations, statistics, images (whether fixed or moving), diagrams, and audio recordings. It includes raw data, processed data, published data, and archived data. It includes the data generated by experiments, by models and simulations, and by observations of natural phenomena at specific times and locations. It includes data gathered specifically for research as well as information gathered for other purposes that is then used in research. It includes data stored on a wide variety of media, including magnetic and optical media.¹⁷

Though our concerns in this report lie largely with the application of digital technologies in research, our examination of the issues is not limited to digital data. Nor does this report address just those areas traditionally considered “science.” It applies to all efforts to derive new knowledge about the physical, biological, or social worlds and thus encompasses research in engineering and in all of the physical, biological, behavioral, and social sciences. The conclusions in the report generally apply to quantitative data. However, many of our conclusions also apply to qualitative data, though we have not focused on the issues unique to qualitative data. Also, this report does not address research in the humanities, which lies outside the committee's charge and expertise.

¹⁷ As a point of comparison, the Office of Management and Budget defines research data as “the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This “recorded” material excludes physical objects (e.g., laboratory samples).” See OMB Circular A-110 at <http://www.whitehouse.gov/omb/circulars/a110/a110.html>.

BOX 1-4 Definitions of “Research Data” from Other Reports

“Data are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors.”^a

“A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.”^b

“Any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment.”^c

^a National Research Council. 1999. *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. Washington, DC: National Academy Press, p. 15.

^b Consultative Committee for Space Data Systems. 2002. *Reference Model for an Open Archival Information System (OAIS)*. Washington, DC: National Aeronautics and Space Administration, p. 1-9. Available at <http://public.ccsds.org/publications/archive/650x0b1.pdf>

^c National Science Board. 2005. *Long-Lived Data Collections: Enabling Research and Education in the 21st Century*. Arlington, VA: National Science Foundation, p. 13.

The term “data” in this report excludes physical objects (including living organisms) and other materials used in research, such as biological reagents or the devices, instruments, or computers that generate experimental or observational data. In many cases, these physical objects can be described in written, numeric, or visual forms, and these descriptions constitute data. However, because materials are tangible whereas data are generally intangible, different issues surround their use, storage, and dissemination. Some of the observations and conclusions in this report apply to materials as well as to data, and on occasion we make this extension of our conclusions explicit. However, the treatment of materials in research introduces issues that are beyond the subject matter of this report.¹⁸

Finally, our definition excludes information that can be important in research but is not used to generate research conclusions, including interpre-

¹⁸ Issues related to sharing research materials in the life sciences have been addressed by a previous National Research Council report. See National Research Council. 2003. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Washington, DC: The National Academies Press.

tive statements, or matters of personal judgment, such as peer reviews, plans for future research, communications with colleagues, or personnel assessments. Of course, the line between research data and subjective judgments is sometimes difficult to draw, since subjective judgments can influence the structure ascribed to data. Nevertheless, a distinction exists, and we do not mean to imply that all of the information associated with research necessarily constitutes research data.

Metadata

As used in this report, the term “metadata” refers to descriptions of *the content, context, and structure of information objects, including research data, at any level of aggregation (for example, a single data item, many items, or an entire database)*. According to the National Science Foundation report *Cyberinfrastructure Vision for the 21st Century*, metadata “summarize data content, context, structure, interrelationships, and provenance (information on history and origins). They add relevance and purpose to data, and enable the identification of similar data in different data collections.”¹⁹ Metadata make it easier for data users to find and utilize data, particularly if they are machine-readable.

Metadata are extremely diverse, ranging from written descriptions of instruments and software to the largely tacit knowledge on which the success of an investigation often depends. They are a critical part of the context needed to assess the integrity of data and use data accurately. Metadata are themselves data, since they consist of descriptive, factual information about data. Thus, conclusions about data in this report generally apply to metadata as well, although special considerations sometimes apply to metadata.

Until fairly recently, the term “metadata” was used primarily by the library community and by individual research communities.²⁰ As digital data has become more important in a variety of disciplines and fields, the scope and value of metadata have grown, leading to the development of metadata standards. Metadata standards represent an agreed set of terminologies, definitions, and values to be provided for data in a given field or community.²¹

¹⁹ NSF Cyberinfrastructure Council (2005), *NSF's Cyberinfrastructure Vision for 21st Century Discovery*, Arlington, VA, National Science Foundation.

²⁰ Tony Gill, Anne J. Gilliland, Maureen Whalen, and Mary S. Woodley. 2008. *Introduction to Metadata*, Version 3.0. Los Angeles, CA: J. Paul Getty Trust. Available at www.getty.edu/research/conducting_research/standards/intrometadata/index.html.

²¹ U.S. Geological Survey, Coastal and Marine Biology InfoBank. USGS CMG “Formal Metadata” Definition. See walrus.wr.usgs.gov/infobank/programs/html/definition/fmeta.html. Accessed December 8, 2008.

Raw and Processed Data

Raw data directly from an instrument or data that have not been documented or processed usually are of little value to anyone except the individuals who generate or collect them. In many fields, capturing data that are “whole” or “perfect” may be difficult or impossible. Instruments may only partially and imperfectly record phenomena. Researchers may not even see the raw data on which their conclusions are based. In some cases, raw data may exist in a computer buffer for only a fraction of a second before they undergo processing. In other cases, raw data may be so voluminous that they cannot be examined in anything other than a processed or condensed form. However, raw data may need to be retained to validate research findings and, in some research fields, to support patent applications, investigate instances of research misconduct, or justify public policies.

Data used to draw conclusions, derive findings, and build models may undergo many changes as they are processed, distributed, and archived. They are analyzed, aggregated, and reformulated by researchers. Data often are organized into structures for long-term storage and access that require the expertise of professionals trained in the management and handling of large databases.

As soon as raw data are processed, the algorithms, computer programs, and other techniques used in that processing become crucial to their understanding. Many data cannot be properly interpreted or used without understanding the processing they have undergone, and it is generally impossible to judge the integrity of processed data without access to the metadata documenting how they were processed. In some cases, this processing may be so machine-dependent that the metadata must include either a thorough representation or a copy of the devices used to do the processing. Consequently, to judge the accuracy and validity of data, researchers, policy makers, and other users of data may need a thorough understanding of the tools and procedures used to analyze those data. In many cases, a high level of expertise is needed to use metadata in order to place data in context.

Given the relatively broad definitions of data and metadata that we have adopted in this report, a great many issues are obviously associated with the generation, use, dissemination, and preservation of research data in the digital age. In this report, however, we focus on three specific issues, which we describe using the terms integrity, accessibility, and stewardship.

Integrity

Integrity describes an uncompromising adherence to ethical values, strict honesty, and absolute avoidance of deception. Integrity also describes the state of being whole and complete, of being totally unimpaired. Thus, the word “integrity” has both an ethical meaning and a structural or methodological meaning. In this report we use the word “integrity” in both senses.

According to one definition, “being assured of data’s integrity means having confidence that the data are complete, verified, and remain unaltered.”²² This is possible only if researchers adhere to professional and ethical standards of their fields. In some research fields, these standards are written, but in many areas they exist as tacit knowledge that is passed from senior researchers to beginning researchers over the course of a research apprenticeship. These professional standards, in turn, describe the methods, procedures, and tools that researchers are expected to employ to minimize error and bias in their work. Consequently, integrity in research has both an individual and a communal meaning. Researchers maintain the integrity of research data by adhering to the professional standards of their fields.

Researchers are expected to describe their methods and tools to others in sufficient detail that the data can be checked and the results verified. Completely and accurately describing the conditions under which data are collected, characterizing the equipment used and its response, and recording anything that was done to the data thereafter are critical to ensuring data integrity. Thus, for experimental data, integrity implies that the data can be reproduced in a test or experiment that repeats the conditions of the original test or experiment. For observational data, data of high “quality” (a term that we sometimes will use as a synonym for data integrity) have been validated through comparison with data whose quality is known or by being generated with an instrument that has been adequately calibrated or tested.

Accessibility

In this report, accessibility refers to the availability of research data to researchers other than those who generated the data. Accessibility is a critical element of integrity, because data must be available to others in order for the validity of those data to be verified. However, in some cases an investigator may not be able to make data available to the public. For example, in private companies, data may need to be restricted for commercial reasons. In such cases, data are frequently made available within the company to evaluate their integrity.

In this report, the term “accessibility” generally implies public access as well as availability to other researchers upon request. Accessibility does not necessarily imply free access, because providing access to data entails financial costs that must be met. Also, access does not necessarily imply that researchers must provide inquirers with the training and expertise they would need to understand or use data. However, data should be accompanied by sufficient metadata for colleagues to assess the integrity of those data.

²² University of Minnesota Research Data Management Online Workshop (www.research.umn.edu/datamgtq1/MDI_020.html).

Stewardship

In the broadest possible sense, the term “utility” in the name of our committee refers to all of the various applications of research data. Both integrity and accessibility are critical elements of utility, because research data must have integrity and be broadly accessible to be effectively utilized.

However, our focus in this report is on a specific aspect of utility that we refer to as data *stewardship*—the long-term preservation of data so as to ensure their continued value, sometimes for unanticipated uses. Stewardship goes beyond simply making data accessible. It implies preserving data and metadata so that they can be used by researchers in the same field and in fields other than that of the data’s creators. It implies the active curation and preservation of data over extended periods, which generally requires moving data from one storage platform to another. The term “stewardship” embodies a conception of research in which data are both an end product of research and a vital component of the research infrastructure.

THE VARIETIES OF RESEARCH DATA

As the examples presented throughout this report illustrate, research data are so varied that they can be described in their entirety only in the most general terms. Different research fields have very different approaches to the treatment of research data. Even at the level of individual research groups, expectations and demands can vary greatly from one investigator to another. This tremendous variety within the research community complicates the task of arriving at conclusions that apply across all fields of research. Research fields are also characterized by diversity in the origins of data and by the size and other characteristics of data collections.

Diversity Across Disciplines

There is great diversity in the ways data are gathered and analyzed both among and within disciplines. The sidebars in this and other chapters describe some of the diversity among disciplines, but individual disciplines also harbor great diversity in the ways data are gathered and analyzed. Data in physics, for example, range from small datasets generated by a “tabletop” experiment to the terabytes of data generated by an accelerator-based experiment. Databases in the social sciences may be freely available to all researchers in some fields and tightly restricted in other fields. Some fields within a discipline may have traditions of storing data for extended periods while others discard data relatively quickly. (In this report, “field” refers to an area of research smaller than a discipline. In many cases, a field can be roughly associated with the community of researchers who follow and publish articles in a relatively small collection

of related journals—what analysts of science have referred to as “invisible colleges.”²³)

Furthermore, some of the most interesting and productive areas of research today involve researchers from multiple disciplines working together on complex, integrative problems.²⁴ In some cases, these areas of multidisciplinary research become so well defined that they evolve into research fields of their own, as in astrobiology. In other cases, researchers may come together to work on a multidisciplinary project and then disband once the project is over. In interdisciplinary research, different traditions of data treatment meet and sometimes clash, and new ways to gather, analyze, and store data may need to be developed to address novel challenges.

Diversity in Origins of Data

The practices for analyzing, disseminating, and storing research data vary greatly from field to field.²⁵ For example, in some fields, observational data can be re-created by other researchers, but in other fields observations are impossible or impractical to make a second time. In these cases, observational data may need to be carefully archived for future use, including uses that cannot currently be foreseen.

Data generated through computer simulations are increasingly important in a variety of fields.²⁶ Data generated entirely by computation can in principle be regenerated, assuming that enough is known about the hardware, software, and inputs used in the computation. However, each of these three components of a computation may be so complex or indeterminate that the computational data have some of the characteristics of observational data. Furthermore, many simulations involve random inputs, so that successive simulations will not be exactly the same. In some cases, sharing and preserving the models and software tools used to create a simulation will be more important for verifying and building upon research than sharing and preserving the data generated. In other cases, the data themselves have value and can represent such a large investment of resources that they may need to be preserved for subsequent use in the same way that unique observational data are preserved.

²³ Daryl E. Chubin. 1983. *Sociology of Sciences: An Annotated Bibliography on Invisible Colleges, 1972–1981*. New York: Garland.

²⁴ National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. 2005. *Facilitating Interdisciplinary Research*. Washington, DC: The National Academies Press.

²⁵ National Research Council. 1995. *Preserving Scientific Data on the Physical Universe: A New Strategy for Archiving Our Nation's Scientific Information*. Washington, DC: National Academy Press.

²⁶ Ghaleb Abdulla, Terence Critchlow, and William Arrighi. 2004. “Simulation data as data streams.” *SIGMOD Record* 33(1):89–94.

Data from experiments may be reproducible if a robust description of the experiment is available. In practice, however, it may not be possible to re-create the exact conditions of the experiment. An experimental apparatus also may be so costly to build or use that experiments can be conducted only once or over a limited time period. If so, long-term preservation of the data generated by the experiment may be essential for optimizing the experiment's value.

Diversity in Types of Data Collections

In this report, we use the term “database” to refer to a collection of data that is organized to permit search, retrieval, processing, and reorganization of stored information. Databases include datasets, which are collections of similar or related data. We use the term “data collection” interchangeably with “database.”

In its report *Long-Lived Data Collections: Enabling Research and Education in the 21st Century*, the National Science Board divided data collections into three broad categories (Box 1-5).²⁷ “Research collections” are the products of one or more focused research projects and typically serve just the research group that generated the data. “Resource collections” serve a single science or engineering community and are generally intermediate in size and budget. “Reference collections” serve large segments of the research and education communities and are often supported by large budgets.

These categories may seem to correspond to small-scale research, intermediate-sized research projects, and large-scale research, but the National Science Board's report shows that such an association can be misleading. Using digital technologies, relatively small-scale projects can generate immense quantities of data that become the basis for research in many related fields. Large-scale reference data collections may be the product of many small projects linked through digital networks. Or large projects may produce focused data collections that serve a narrow research purpose and never become publicly available. Thus, distinguishing research data by the size of the group that generated those data is problematic—in part because of new capabilities created by digital technologies.

STRUCTURE OF THE REPORT

The remainder of this report is organized into three thematic chapters and a final summary chapter. Chapter 2 considers the integrity of data throughout their life cycle, from their collection to their disposal or preservation. Maintaining the integrity of research data is a fundamental obligation of researchers;

²⁷ National Science Board. 2005. *Long-Lived Data Collections: Enabling Research and Education in the 21st Century*. Arlington, VA: National Science Foundation.

BOX 1-5 Three Types of Data Collections

The National Science Board (NSB) has organized data collections into the three categories described below. In addition, the NSB defined “collection” to refer “not only to stored data but also to the infrastructure, organizations, and individuals necessary to preserve access to the data.”

Research data collections are the products of one or more focused research projects and typically contain data that are subject to limited processing or curation. They may or may not conform to community standards, such as standards for file formats, metadata structure, and content access policies. Quite often, applicable standards may be nonexistent or rudimentary because the data types are novel and the size of the user community [is] small. Research collections may vary greatly in size but are intended to serve a specific group, often limited to immediate participants. There may be no intention to preserve the collection beyond the end of a project. One reason for this is funding. These collections are supported by relatively small budgets, often through research grants funding a specific project. (Example: The Fluxes Over Snow Surfaces Project, <http://www.atd.ucar.edu/rtf/projects/FLOSS>.)

Resource or community data collections serve a single science or engineering community. These digital collections often establish community-level standards either by selecting from among preexisting standards or by bringing the community together to develop new standards where they are absent or inadequate. The budgets for resource or community data collections are intermediate in size and generally are provided through direct funding from agencies. Because of changes in agency priorities, it is often difficult to anticipate how long a resource or community data collection will be maintained. (Example: The Arabidopsis Information Resource, <http://www.arabidopsis.org>.)

Reference data collections are intended to serve large segments of the research and education community. Characteristic features of this category of digital collections are a broad scope and a diverse set of user communities including scientists, students, and educators from a wide variety of disciplinary, institutional, and geographical settings. In these circumstances, conformance to robust, well-established, and comprehensive standards is essential, and the selection of standards by reference collections often has the effect of creating a universal standard. Budgets supporting reference collections are often large, reflecting the scope of the collection and breadth of impact. Typically, the budgets come from multiple sources and are in the form of direct, long-term support, and the expectation is that these collections will be maintained indefinitely. (Example: Protein Data Bank, <http://www.pdb.org>.)

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

achieving this objective in the digital age can be either easier or more difficult than in earlier times.

Chapter 3 considers the issues of accessing and sharing research data. The research enterprise is built on the precept that researchers will make the data on which publicly disseminated conclusions are based available to their colleagues so that others can verify and build on those data. Accessibility is vital for ensuring the integrity of research data and facilitating their future use.

Chapter 4 discusses the stewardship of research data, that is, their long-term preservation in databases for various future research uses and other applications. Preserving data collections can be expensive and difficult—so much so that it can compete with the conduct of research. Yet the loss of many kinds of research data also can incur substantial costs.

The final chapter reorganizes recommendations that have appeared earlier in the volume according to different actors within the research community rather than thematically. It also discusses how action can be motivated when responsibility for research integrity, accessibility, and stewardship is shared across the components of the research community. Each part of the research enterprise has much to gain or lose, depending on how research data are managed, and each has a role to play in ensuring the integrity, accessibility, and stewardship of research data.

2

Ensuring the Integrity of Research Data

The fields of science span the totality of natural phenomena and their styles are enormously varied. Consequently, science 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 truth and accuracy of their data are as varied as the fields themselves, there are no universal procedures for achieving technical accuracy. There are, however, some broadly accepted practices for pursuing science. In most fields of science, for instance, experimental observations must be shown to be reproducible in order to be creditable.¹ Other practices include checking and rechecking data to ensure that the interpretation is valid, and also submitting the results to peer review to further confirm that the findings are sound. Yet other practices may be employed only within specific fields, for instance, the use of double-blind trials, or the independent verification of important results in separate laboratories.

Although the pervasive use of high-speed computing and communications in research has vastly expanded the capabilities of researchers, if used inappropriately or carelessly, digital technologies can lower the quality of data and compromise the integrity of research.² Digitization may introduce spurious information into a representation, and complex digital analyses of data can yield misleading results if researchers are not scrupulously careful in monitoring and understanding the analysis process. Because so much of the processing

¹ Even this fundamental principle can have exceptions. For instance, observations with a historical element, such as the explosion of a supernova or the growth of an epidemic, cannot be reproduced.

² The challenges of maintaining data integrity over the long term, including the decay of physical storage media and improper manipulation of archived data, are discussed in Chapter 4.

and communication of digital data are done by computers with relatively little human oversight, erroneous data can be rapidly multiplied and widely disseminated. Some projects generate so much data that significant patterns or signals can be lost in a deluge of information. As an example of the challenges posed by digital research data, Box 2-1 explores these issues in the context of particle physics research.

Because digital data can be manipulated more easily than can other forms of data, digital data are particularly susceptible to distortion. Researchers—and others—may be tempted to distort data in a misguided effort to clarify results. In the worst cases, they may even falsify or fabricate data.

BOX 2-1

Digital Data in Particle Physics

From the invention of digital counting electronics in the early days of nuclear physics, to the creation of the World Wide Web and the data acquisition technology for the Large Hadron Collider (LHC), particle physics has been a major innovator of digital data technology. The LHC, which recently came into operation at the European Center for Nuclear Research (CERN) in Geneva, has spawned a new generation of data processing. The accelerator collides two beams of protons, resulting in about a billion proton-proton collisions every second. These collisions occur at several points around the 27-km circumference of the circular accelerator. This first step of the process is difficult enough to imagine, but the next steps are even more amazing.

Part of the energy carried by the two colliding protons is converted into matter by fundamental processes of nature. Some of these processes are well understood, but others might represent major discoveries that could deepen our understanding of the universe—for instance, the creation of particles that constitute the so-called dark matter inferred from astrophysical measurements.

The spray of energetic outgoing particles from one such collision is called an event.

The particles in the spray have speeds approaching the speed of light. They fly out of the proton-proton collision point into a surrounding region that is instrumented with an array of sophisticated particle detection devices, collectively called a detector. The detector senses the passage of subatomic particles, creating a detailed electronic image of the event and providing quantitative information about each particle such as its energy and its relation to certain other particles.

Each proton-proton collision generates about 1 megabyte of information, yielding a total rate of 1 petabyte per second. It is not practical to record this staggering amount of information, and so the experimenters have devised techniques for rapidly selecting the most promising subset of the data for exhaustive analysis.

Only a tiny fraction of the deluge—perhaps one in a trillion—will be due to new kinds of physical processes of fundamental importance. Once the detector has recorded an event, a high-speed system performs a rapid analysis (within 3 micro-

As an example of how digital data can be inappropriately manipulated, consider the case of digital images in cell biology. When the journals published by the Rockefeller University Press, including the *Journal of Cell Biology*, adopted a completely electronic work flow in 2002, the editors gained the ability to check images for changes in ways that were not possible previously. The *Journal of Cell Biology*, in consultation with the research community it serves, therefore adopted a policy that specified its expectations and procedures:

No specific feature within an image may be enhanced, obscured, moved, removed, or introduced. The grouping of images from different parts of the same gel, or from dif-

seconds) that retains typically 1 in 30,000 of all events. A second rapid analysis step reduces the rate of permanently recorded data down to about 100 events per second.

Research at the LHC is carried about by international collaborations that construct, operate, and analyze the data from each of the four main detectors. The scale of the research borders on the fantastic: Two of the collaborations each have about 2,000 members from 40 different countries; the volume of the ATLAS detector, for example, is about half that of Notre Dame cathedral, and the mass of iron in its gigantic solenoid magnet is approximately that in the Eiffel Tower.

LHC detectors are complex systems that require meticulous calibration, alignment, and quality control procedures. The data from an LHC detector flow from the arrays of devices that track the particles emitted when the protons collide. The data processing system determines the momentum and energy of each particle radiated from a collision, and identifies how the particles are correlated in space and time. The thousands of detection devices, the magnetic field in which the collisions occur, and the properties of the complex digital data acquisition system must all be known accurately. The complexities of data analysis in LHC experiments are comparable to those of the apparatus itself.

Ensuring the integrity of data from a particle physics experiment presents special challenges because no form of traditional peer review would be sufficient. The experiments are so complicated that a knowledgeable outsider who attempted to evaluate the performance of the detection system would require years for the job. Consequently, the particle physics community has developed a method for reliable internal quality assurance that goes beyond straightforward peer review.

As part of each major collaboration, multiple data-analysis teams work to evaluate the performance of the apparatus and analyze the data independently, withholding their final results until the latest possible moment. In effect, in the particle physics community a major portion of the role that was traditionally played by straightforward peer review has been augmented by a process of critical self-analysis.

ferent gels, fields, or exposures must be made explicit by the arrangement of the figure (i.e., using dividing lines) and in the text of the figure legend. If dividing lines are not included, they will be added by our production department, and this may result in production delays. Adjustments of brightness, contrast, or color balance are acceptable if they are applied to the whole image and as long as they do not obscure, eliminate, or misrepresent any information present in the original, including backgrounds. Without any background information, it is not possible to see exactly how much of the original gel is actually shown. Non-linear adjustments (e.g., changes to gamma settings) must be disclosed in the figure legend. All digital images in manuscripts accepted for publication will be scrutinized by our production department for any indication of improper manipulation. Questions raised by the production department will be referred to the Editors, who will request the original data from the authors for comparison to the prepared figures. If the original data cannot be produced, the acceptance of the manuscript may be revoked. Cases in which the manipulation affects the interpretation of the data will result in revocation of acceptance, and will be reported to the corresponding author's home institution or funding agency.

—*The Journal of Cell Biology*, Instructions to Authors,
<http://www.jcb.org/misc/ifora.shtml>

Having developed this policy, the editors at the *Journal of Cell Biology* began to screen all of the images in accepted articles for evidence of inappropriate manipulation. For example, simple brightness and contrast adjustments could reveal inconsistencies in the background of the image that are clues to manipulation. In this way, the editors could determine whether the images presented in a manuscript were an accurate representation of what was actually observed and whether the quality or context in which the images were obtained was apparent.

Over the course of the next 5 years, the editors screened the images in 1,869 accepted papers.³ Over a quarter of the manuscripts contained one or more images that had been inappropriately manipulated. In the vast majority of those cases, the manipulation violated the journal's guidelines but did not affect the interpretation of the data, and the articles were published after the authors revised the images in accordance with the guidelines.

In 18 of the papers—about 1 percent of the total for which the editors sought and obtained the original data—the editors determined that the image manipulations affected the interpretation of the data. The acceptance of those papers was revoked, and they were not published. In only one case did the authors state that the original data could not be found and withdrew the paper.

According to a federal definition of research misconduct developed by the Office of Science and Technology Policy, misconduct consists of fabrication, fal-

³ These figures are from Mike Rossner, The Rockefeller University Press, presentation to the committee, April 16, 2007. For background, see Mike Rossner and Kenneth M. Yamada. 2004. "What's in a picture: The temptation of image manipulation." *Journal of Cell Biology* 166(1):11–15.

sification, or plagiarism of research results.⁴ However, the editors at the *Journal of Cell Biology* do not consider the element of “intent” in their inquiries into potential violations of their guidelines. They obtain the original data directly from the authors, since whether an image has been inappropriately manipulated can be determined only by comparing the submitted figures with the original data. Initial inquiries from the journal emphasize that questions are being asked only about the presentation of data, not its integrity, and inquiries are kept strictly confidential between a journal and authors.

The section on image manipulation in the *White Paper on Promoting Integrity in Scientific Journal Publications* by the Council of Science Editors, which was written by the editors at the *Journal of Cell Biology*, suggests that “journal editors should attempt to resolve the problem before a case is reported. This is because the vast majority of cases do not turn out to be fraudulent.”⁵

Since the *Journal of Cell Biology* adopted its policy, other journals, including the *Proceedings of the National Academy of Sciences* and *Nature*, have begun screening images for evidence of inappropriate manipulation (see Table 2-1). Generally, these journals have screened a subset of papers and have made the additional level of scrutiny known to authors in the hope that this will act as a disincentive to manipulation.⁶ In addition, software is being developed that may automate at least part of the screening process so that more images can be examined with less expense.

Publishers of scientific, engineering, and medical journals continue to grapple with issues related to technological change and ensuring the integrity of published results. Concurrent with the present study, a number of leading journals have held a series of meetings to discuss these issues. One question is whether the additional efforts on the part of journals to screen digital images entail additional responsibilities. For example, suppose a journal screens digital images in a manuscript, finds something suspicious, and after undertaking an inquiry and finding that an image has been fraudulently manipulated rejects the paper. Does the journal have further responsibilities, and if so what are they? According to the *White Paper on Promoting Integrity in Scientific Journal Publications* by the Council of Science Editors, when a journal “suspects an article contains material that may result in a finding of misconduct, the editor can notify some or all of the following parties: the author who submitted the article, all authors of the article, the institution that employs the author(s), the sponsor of the study, or an agency that would have jurisdiction over an inves-

⁴ Office of Science and Technology Policy, Federal Policy on Research Misconduct. Available at <http://ori.dhhs.gov/education/products/RCRintro/c02/b1c2.html>.

⁵ Editorial Policy Committee. 2006. CSE’s White Paper on Promoting Integrity in Scientific Journal Publications. Reston, VA: Council of Science Editors, p. 50.

⁶ Unfortunately, the experience of the editors of the *Journal of Cell Biology* indicates that this is not the case, because the rates at which they see image manipulation have not declined over the past 5 years.

TABLE 2-1 Analysis of Journal Policies

	Nature	Science	PNAS
Data and methods access			
Does the journal require that all data be made available on request to journal editors and reviewers?	Yes	Yes	Yes
Does the journal require deposition of data in a public repository?	Yes	Yes	Yes
Are authors required to provide algorithms or computer programs used in the collection, report, or analysis of data?	No	No	No
Image manipulation			
Is image manipulation prohibited?	No	No	No
Does the journal require that image manipulation be reported?	Yes	Yes	Yes
Does the journal require that digital techniques be applied to the entire image?	Yes	Yes	Yes
Does the journal use software tests to detect image manipulation?	Yes	Yes	Yes
Ethics and Scientific Misconduct			
Is there a specified ethical statement?	Yes	Yes	Yes
Does the journal have a scientific misconduct investigation or reporting policy in place?	Yes ^g	Yes ^h	Yes ⁱ

KEY: PNAS=Proceedings of the National Academy of Sciences; JCB=Journal of Cell Biology and other Rockefeller University Press; NEJM=New England Journal of Medicine; ACS=American Chemical Society journals; AGU=American Geophysical Union journals; FASEB=Federation of American Societies for Experimental Biology journals; IEEE=Institute of Electrical and Electronics Engineers journals; ESA=Ecological Society of America journals; AER=American Economic Review

^a FASEB is reviewing their policies as this goes to press.

^b The authors have to provide the editors with their data and programs AFTER acceptance for publication (data and programs are then posted to a public repository); authors are not required to provide data and other information to reviewers.

^c For certain studies only.

^d Only if the author wishes to cite the data must it be in a public depository. AGU does strongly encourage all authors to deposit their data but it is not a requirement for publication.

^e Encouraged.

tigation of the matter (e.g., ORI [Office of Research Integrity]).”⁷ In practice, however, an editor may be reluctant to initiate action that could have disciplinary consequences.⁸

Another question is whether the high incidence of inappropriate manipulation of images in the above example reflects a lack of experience with applying

⁷ Editorial Policy Committee. 2006. CSE’s White Paper on Promoting Integrity in Scientific Journal Publications. Reston, VA: Council of Science Editors, p. 50.

⁸ D. Butler. 2008. “Entire-paper plagiarism caught by software.” *Nature News* 455:715.

JCB	NEJM	ACS	AGU	FASEB ^a	IEEE	ESA	AER
Yes	No	Yes	Yes	Yes	Yes	Yes	No ^b
Yes	Yes ^c	Encouraged	No ^d	Yes	No	No ^e	Yes
Yes	Yes ^f	Yes	No	Yes	No	No	Yes
No	No	No	No	No	No	No	No
Yes	Yes	No	No	No	No	No	No
Yes	No	No	No	No	No	No	No
Yes	No	No	No	No	No	No	No
Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Yes ^g	Yes	Yes	Yes	Yes	Yes	Yes	No

^f On request.

^g Specifies steps that will be taken in cases of suspected plagiarism and failure to provide data.

^b Policies are “in place regarding reporting scientific misconduct, but these are internal and not listed externally.”

ⁱ “Cases of deliberate misrepresentation of data will result in rejection of the paper and will be reported to the corresponding author’s home institution or funding agency.”

^j “Cases in which the (image) manipulation affects the interpretation of the data will result in revocation of acceptance, and will be reported to the corresponding author’s home institution or funding agency.”

SOURCES: Compiled from journal Web sites. All journals are peer-reviewed publications. Additional information provided by journals 2009.

the standards of science to digital data or an underlying disregard for the standards of science. The recommendations presented later in this chapter address the need for researchers not only to understand the reasons for maintaining the integrity of research data, but also the methods for doing so.⁹

All research data, whether digital or not, are susceptible both to error and

⁹ National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. 2009. *On Being a Scientist: Responsible Conduct in Research*, 3rd ed. Washington, DC: The National Academies Press.

to misrepresentation. Digital technologies can introduce technical sources of error into data analysis, communication, or storage systems. At the frontiers of human knowledge, the data that bear on a problem can be very difficult to separate from irrelevant information.¹⁰ Research methods may not be firmly established, and even the questions being asked may not be fully defined.

Furthermore, researchers may have incentives to structure research or gather data in ways that favor a particular outcome, as in the case of drug studies funded by companies that stand to profit from particular results.¹¹ In addition, researchers can have philosophical, political, or religious convictions that can influence their work, including the ways they collect and interpret data.¹² Because of the many ways in which data can depart from empirical realities, everyone involved in the collection, analysis, dissemination, and preservation of data has a responsibility to safeguard the integrity of data.

THE ROLES OF DATA PRODUCERS, PROVIDERS, AND USERS

The example from the *Journal of Cell Biology* illustrates the different roles that individuals and groups can play in ensuring the integrity of data. For the purposes of this report, we have divided these individuals and groups into three categories—data producers, data providers, and data users—though it should be kept in mind that many individuals and organizations fall into more than one of these categories.

Data producers are the scientists, engineers, students, and others who generate data, whether through observations, experiments, simulations, or the gathering of information from other sources. Often the creation of data is an explicit objective of research, but data can be generated in many ways. For example, administrative records, archaeological artifacts, cell phone logs, or many other forms of information can be adapted to serve as inputs to research. Data also are produced by government agencies in the course of performing tasks for other purposes (such as remote sensing for weather forecasts or conducting the decadal censuses), and these data can be used extensively for research. This report focuses on data produced through activities that are related primarily to research, but the general principles laid out in this report apply to all data used in research.

¹⁰ E. Brian Davis. 2003. *Science in the Looking Glass: What Do Scientists Really Know?* New York: Oxford University Press.

¹¹ Sheldon Krimsky. 2006. "Publication bias, data ownership, and the funding effect in science: Threats to the integrity of biomedical research." Pp. 61–85 in *Rescuing Science from Politics: Regulation and the Distortion of Scientific Research*, eds. Wendy Wagner and Rena Steinzor. New York: Cambridge University Press.

¹² National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. 2009. *On Being a Scientist: Responsible Conduct in Research*, 3rd ed. Washington, DC: The National Academies Press.

Data providers consist of the individuals and organizations who are responsible, whether formally or informally, for making data accessible to others. Sometimes a data provider may be simply the producer of those data, because data producers generally are expected to make data available to verify research conclusions and allow for the continued progress of research. In other cases, data may be deposited in a repository, center, or archive that has the responsibility of disseminating the data. Journals also can be data providers, either through the articles they publish or through the provision of supplementary material that supports a published article.

Data users are the individuals and groups who access data in order to use those data in their own work, whether in research or in other endeavors. At one extreme, the users of data may belong entirely to the community of originating researchers (as in the case of elementary particle physics, which is described in this chapter). At the other extreme, a given body of data may be of wide interest to people outside a research field (as in the case of climate records, which is discussed in Chapter 3). Data producers are generally data users, but the collective body of data users extends beyond the research community to policy makers, educators, the media, the courts, and others. Data users can work in fields quite different from those of data producers, which means that they have an interest in being able to access data that are well annotated in order to use them accurately and appropriately.

As described below, each of these three groups has particular responsibilities in ensuring the integrity of research data.

THE COLLECTIVE SCRUTINY OF RESEARCH DATA AND RESULTS

In Chapter 1, we noted that measures of data integrity have both individual and collective dimensions. At an individual level, ensuring integrity means ensuring that the data are complete, verified, and undistorted. This is essential for science and engineering to progress, but it is not sufficient because progress in understanding the world requires that knowledge be shared. This process of submitting research data and results derived from those data to the scrutiny of others provides for a collective means of establishing and confirming data integrity. When others can examine the steps used to generate data and the conclusions drawn from those data, they can judge the validity of the data and results and accept (perhaps with reservations) or reject proffered contributions to science. Of course, the collective scrutiny of research results cannot guarantee that those results will be free of error or bias. For instance, it is noteworthy that important phenomena such as plate tectonics, chaotic motion in mechanical systems, or the functions of “junk” DNA were overlooked for decades because of theoretical perspectives that shaped the collection of data in those fields. Nevertheless, by bringing multiple perspectives to bear on a common body of

information, the error and bias inherent in individual perspectives can be minimized. In this way, the frontiers of understanding continually advance through the collective evaluation of new data and hypotheses.

Data producers, providers, and users are all involved in the collective scrutiny of research data and results. Data producers need to make data available to others so that the data's quality can be judged. (Chapter 3 discusses the accessibility of research data.) Data providers need to make data widely available in a form such that the data can be not only used but evaluated, which requires that data be accompanied by sufficient metadata for their content and value to be ascertained. (Chapter 4 discusses the importance of metadata.) Finally, data users need to examine critically the data generated by themselves and others. The critical evaluation of data is a fundamental obligation of all researchers.

Completely and accurately describing the conditions under which data are collected, characterizing the equipment used and its response, and recording anything that was done to the data thereafter are critical to ensuring data integrity. In this report we refer to the techniques, procedures, and tools used to collect or generate data simply as methods, where a "method" is understood to encompass everything from research protocols to the computers and software (including models, code, and input data) used to gather information, process and analyze data, or perform simulations. The validity of the methods used to conduct research is judged collectively by the community involved in that research. For example, a community may decide that double-blind trials, independent verification, or particular instrumental calibrations are necessary for a body of data to be accepted as having high quality. Scientific methods include both a core of widely accepted methods and a periphery of methods that are less widely accepted. Thus, discussions of data integrity inevitably involve scrutiny of the methods used to derive those data.

The procedures used to ensure the integrity of data can vary greatly from field to field. The methods high-energy physicists use to ensure the integrity of data are quite different from those of clinical psychologists. The cultures of the fields of research are enormously varied, and there are no universal procedures for achieving technical accuracy. Some practices may be employed only within specific fields, such as the use of double-blind trials. Some of these field-specific methods may be embodied in technical manuals, institutional policies, journal guidelines, or publications of professional societies. Other methods are part of the collective but tacit knowledge held in common by researchers in that field and passed down to beginning researchers through instruction and mentoring.

In contrast to field-specific methods, some methods used to ensure data integrity extend across most fields of research. Examples include the review of data within research groups, replication of previous observations and experiments, peer review, the sharing of data and research results, and the retention of raw data for possible future use.

The importance of understanding the particular methods used (whether field-specific or general) is signaled in some publications by a “methods section” that describes the procedures used to derive a result. In some print journals, methods sections are being squeezed by pressures to cut costs, though conventionally sized or longer methods sections may be available in supplementary material online. Researchers also may abbreviate methods sections to keep some procedures private in order to obscure the processes used to derive data.

To some extent, researchers must simply trust that other researchers have adhered to the methods accepted in a field of scientific, engineering, or medical research. Sometimes it is impossible to specify in enough detail the procedures used to gather or generate data so that others will get exactly the same results. In such cases, assistance from the original researcher may be necessary for other researchers to replicate or extend earlier results.

The importance of understanding the methods of collecting or generating the data emphasizes the importance of understanding the context of data. Most data cannot be properly interpreted without at least some—and frequently detailed—understanding of the procedures, instruments, and processing used to generate those data. Thus, data integrity depends critically on communicating to other researchers and to the public the context in which data are generated and processed.

PEER REVIEW AND OTHER MEANS FOR ENSURING THE INTEGRITY OF DATA

Of all the social processes used to maintain the integrity of the research enterprise, the most prominent is peer review of articles submitted to a scholarly journal for publication. Review of submitted articles by the authors’ peers screens for quality and relevance and helps to ensure that professional standards have been maintained in the collection and analysis of data. It provides a forum in which the collective standards of a field can be not only negotiated but enforced, because of the researchers’ interests in having their results published. Peer review examines whether research questions have been framed and addressed properly, whether findings are original and significant, and whether a paper is clearly written and acknowledges previous work. Peer review also organizes research results so that the most important research appears in specific journals, which allows for more effective communication.

Because peer review is such an effective tool in quality control, it also is used in evaluating researchers. Researchers are judged for purposes of hiring and promotion largely on the basis of publication in peer-reviewed journals. Furthermore, publication in these journals remains the most important way to disseminate quality-controlled contributions to knowledge. The number of peer-reviewed journals is continuing to grow, and importance of peer review has not diminished during the digital era.

However, changes in the way research is conducted, including many changes caused by digital technologies, have put pressure on the peer review system.¹³ The volume or diversity of research data supporting a conclusion may overwhelm the ability of a reviewer to evaluate the link between the data and that conclusion. As supporting information for a finding in a submitted paper increasingly moves to lengthy supplemental materials, reviewers may be less able to judge the merits of a paper. In addition, journals and funders can have trouble finding peer reviewers who are competent and have the time to judge complex interdisciplinary manuscripts.

Peer review cannot ensure that all research data are technically accurate, though inaccuracies in data can become apparent either in review or as researchers seek to extend or build on data. The research system is based to a large degree on trust. As described later in this chapter, training and the development of standards are crucial factors in building trust. Broader cultural forces such as reward systems, the reputation of researchers and their institutions, and social and cultural penalties for violation of trust also serve to build and maintain trust.

A recent example that illustrates both the limitations of peer review and the strengths of the cumulative nature of science is the case of Seoul National University researcher Woo Suk Hwang. Major advances in stem cell technology that were reported by Hwang and his colleagues and published in the journal *Science* were based on fabricated data.¹⁴ The fraud was uncovered and confirmed after the original publication because of continued scrutiny of the results by the research community. Another case involving fabricated data is described in Box 2-2.

Changes in publication practices are affecting peer review. Largely because of advances in digital communications, the scholarly publishing industry is undergoing dramatic changes, some of which are having a major influence on the economics of the industry.¹⁵ Peer review is expensive because of the time devoted to the process by editors, reviewers, and authors responding to reviewers' comments. Changes in the economics of scholarly publishing may put pressure on editors and publishers to lessen the emphasis on peer review as they strive to cut costs and increase efficiency.

At the same time, digital technologies can strengthen peer review by catalyzing and facilitating new ways of reviewing publications. For example,

¹³ Stevan Harnad. 1998. "Learned inquiry and the net: The role of peer review, peer commentary and copyright," *Learned Publishing* 11:183–192. Available at <http://cogprints.org/1694/0/harnad98.toronto.learnedpub.html>. Accessed February 23, 2007.

¹⁴ Mildred K. Cho, Glen McGee, and David Magnus. 2006. "Lessons of the stem cell scandal." *Science* 311(5761): 614–615.

¹⁵ National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. 2004. *Electronic Scientific, Technical, and Medical Journal Publishing and Its Implications*. Washington, DC: The National Academies Press.

BOX 2-2 Breach of Trust

Beginning in 1998, a series of remarkable papers attracted great attention within the condensed-matter physics community. The papers, based largely on work done at Bell Laboratories, described methods that could create carbon-based materials with long-sought properties, including superconductivity and molecular-level switching. However, when other materials scientists sought to reproduce or extend the results, they were unsuccessful.

In 2001, several physicists inside and outside Bell Laboratories began to notice anomalies among the papers. Several contained figures that were very similar, even though they described different experimental systems. Some graphs seemed too smooth to describe real-life systems. Suspicion quickly fell on a young researcher named Jan Hendrik Schön, who had helped create the materials, had made the physical measurements on them, and was a co-author on all the papers.

Bell Laboratories convened a committee of five outside scientists to examine the results published in 25 papers. Schön, who had conducted part of the work in the laboratory where he did his Ph.D. at the University of Konstanz in Germany, told the committee that the devices he had studied were no longer running or had been thrown away. He also said that he had deleted his primary electronic data files because he did not have room to store them on his old computer and that he kept no data notebooks while he was performing the work.

The committee concluded that Schön had engaged in fabrication in at least 16 of the 25 papers. Schön was fired from Bell Laboratories and later left the United States. In a letter to the committee, he wrote that “I admit I made various mistakes in my scientific work, which I deeply regret.” Yet he maintained that he “observed experimentally the various physical effects reported in these publications.”

The committee concluded that Schön acted alone and that his 20 co-authors on the papers were not guilty of research misconduct. However, the committee also raised the issue of the responsibility that co-authors have to oversee the work of their colleagues. The committee concluded that the extent of this responsibility had not been established within the research community. The senior author on several of the papers, all of which were later retracted, wrote that he should have asked Schön for more detailed data and checked his work more carefully, but that he trusted Schön to do his work honestly. In response to the incident, Bell Laboratories instituted new policies for data retention and internal review of results before publication. It also developed a new research ethics statement for its employees.

SOURCE: National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. 2009. *On Being a Scientist: Responsible Conduct in Research*. Washington, DC: The National Academies Press.

some journals have been experimenting with making reviews open and public.¹⁶ In some cases, reviewers' names are known to authors and readers. In other cases, their reviews and authors' responses become part of the online record of publication. More radical innovations, such as the continuous improvement of published materials through wikis and similar approaches, or peer rankings and commentary on published papers, could further change both journals and the institution of peer review.

Although it is clear that traditional peer review processes remain vital for evaluating the importance and relevance of research, the advance of digital technologies is providing new opportunities to ensure the integrity of data. The emergence and growth of accessible databases such as GenBank and the Sloan Digital Sky Survey illustrate these opportunities in widely disparate disciplines.¹⁷

Many researchers post databases, draft papers, oral presentations, simulations, software packages, or other scholarly products on personal or institutional Web sites. Repositories, such as the Nature Precedings repository established by the Nature publishing group for the life sciences, allow researchers to share, discuss, and cite preliminary findings.¹⁸ The Web allows widespread dissemination of critiques, commentaries, blogs, and other communications. All of these communications can be widely disseminated without undergoing a formal peer review process. In these cases, the quality of research results and the underlying data may be uncertain, and other researchers may have questions in deciding whether to rely on that research in their own work.

The processes for reviewing data that are preserved in a repository or otherwise made widely available to researchers can be quite different from the procedures for reviewing data presented in a publication.¹⁹ Trust in the quality of data may require personal knowledge of how the data were collected and analyzed. Metadata that carefully describe the origins and subsequent processing of the data can increase confidence in the validity of the data.

In some cases, digital technologies can assist in ensuring data quality and building trust in the integrity of the data. Verified technical methods for gather-

¹⁶ A number of open access journals maintain open peer review processes. The traditional journal *Nature* experimented with an open peer review process during 2006, finding that the open process was not popular with authors or reviewers. Sarah Greaves, Joanna Scott, Maxine Clarke, Linda Miller, Timo Hannay, Annette Thomas, and Philip Campbell. 2006. "Overview: Nature's peer review trial." *Nature* doi:10.1038/nature05535. Available <http://www.nature.com/nature/peerreview/debate/nature05535.html>. This report is also discussed in an editorial. 2006. "Peer review and fraud." 444:971.

¹⁷ 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.

¹⁸ See <http://precedings.nature.com/>.

¹⁹ Christine L. Borgman. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

ing, analyzing, and disseminating data can establish tight connections between natural phenomena and representations of those phenomena. Digital technologies also can allow for the widespread dissemination of data and research results to potential reviewers and data users. The emergence and growth of accessible databases such as GenBank and the Sloan Digital Sky Survey illustrate these opportunities in widely disparate disciplines.²⁰ (Box 2-3 on clinical research in this chapter describes another example.) However, it can be difficult to verify the integrity of results based on large datasets that have undergone substantial processing.

In cases where research results or underlying data are distributed electronically without undergoing peer review, researchers may be able to find other ways to submit them to collective evaluation. For example, they may be able to submit data to informal review by colleagues or open review by users of electronic documents. To advance science, in some cases it may be desirable to disseminate data and conclusions in ways other than through peer-reviewed publications. Electronic technologies are greatly enhancing this dissemination.

However, widespread dissemination of research results and underlying data that have not been vetted through the social mechanisms characteristic of research poses the risk that the conclusions drawn from available data can be distorted. Furthermore, it can be difficult for a community to assess the validity of evaluations that are outside traditional peer review processes. And academic disciplines and institutions are just beginning to develop methods for evaluating and rewarding researchers for the production of results that have not undergone peer review or have undergone only informal review.²¹

Fields of research may settle on methods that enhance the quality of research without following all the steps of a formal review process. For example, a research community may structure itself to examine and verify research procedures and data, even though the data are not publicly accessible, as happens in high-energy physics. Another example is research in economics, where authors often work on papers for extended periods, presenting preliminary version of their papers (and data) at conferences and receiving official critiques from their colleagues prior to submitting a paper for publication.

In other cases, the accuracy of data may be continuously reviewed as they are incorporated into ongoing research in such a way that their accuracy is checked; for example, this is one of the quality control mechanisms used with

²⁰ 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.

²¹ ACRL Scholarly Communications Committee. 2007. *Establishing a Research Agenda for Scholarly Communication: A Call for Community Engagement*. Chicago: Association of College and Research Libraries. Available at http://acrl.ala.org/scresearchagenda/index.php?title=Main_Page.

BOX 2-3

Using Digital Technologies to Enhance Data Integrity

Digital technologies can pose risks to data integrity, but they also offer ways to *improve* the reliability of research data. By enabling phenomena and objects to be described and analyzed more comprehensively, they make it possible to remove some of the simplifying assumptions inherent in earlier research. They enable researchers to build checking and verification procedures into research protocols in ways that reduce the potential for error and bias. Automated data collection that is quality controlled can be much more accurate when either substituting for or supplementing human observations.

Although examples from many disciplines could be cited, a good example is the use of digital technologies in clinical research, including the conduct of clinical trials and plans to link clinical trial information with individuals' electronic health records.

Access to the data behind the production of new drugs and other medical treatments is often a contentious issue because of the proprietary traditions of the pharmaceutical industry and concerns about the privacy and security of patients enrolled in clinical trials. Nonetheless, the trend in drug development is toward openness, as databases are made more widely available and prepublication information is published in electronic form to make significant findings quickly available. For example, a GlaxoSmithKline Clinical Trial Register has been created to afford online access to factual summaries of clinical trails of marketed prescription medicines and vaccines.^a Although some specialty journals oppose this practice, the general trend toward openness is being pulled by powerful demands for public assurances about accuracy, completeness, and timeliness.

In the United States the federal government has been the primary force behind making drug development data both electronic and public. The Food and Drug Administration (FDA), for example, is moving away from onsite audits of clinical trials to statistically based sampling and electronic audits. The agency is adapting many tools borrowed from the banking, nuclear, and other sectors where security checks and balances have been in place for a long time.

An important catalyst for electronic data handling has been the FDA's issuance of regulation 21 CFR Part 11 in 1997,^b which provided criteria for acceptance of electronic records and electronic signatures. This regulation not only opened the door to electronic submissions but also encouraged the widest possible use of electronic technology in all FDA program areas, including data storage, archiving, monitoring, auditing, and review. A significant goal was that data should be shareable between sponsors and reviewers.

In 2004, FDA made electronic submission mandatory and called for electronic data handling as well, with the primary goal of faster product reviews and acceptance. FDA is currently planning to adopt single standards for the full life cycle of clinical trials, from the protocol through the capture of source data to analysis, submission, and archiving.

Industry has long been viewed as opposed to making data supporting clinical trials or publications public, partly out of a desire to maintain competitive advantages and partly out of concern that data could be misjudged, mishandled, or otherwise abused in a public forum. This attitude is starting to change as the use of the Internet

becomes widespread (the accessibility of data is discussed in more detail in the next chapter).^c

The next frontier of the evolution of clinical research toward an electronic future is the electronic integration of clinical trials data and patients' health records. This integration is anticipated to open new areas of research that feature enhanced risk assessment, improved natural history and epidemiological assessment, more reliable information, and better drug use.

The primary challenge is to develop standards to bridge the different standards and terminologies used in clinical trials with those used in medical recordkeeping. This process presents daunting difficulties, including:

- Health records include a broader range of terminology than clinical trials. For example, a myocardial infarction might be described in a medical record as coronary insufficiency, chest discomfort, or other terms that may be difficult to capture in an electronic system.
- The codes for most electronic health records were developed for reimbursement and billing purposes, not for clinical use or research.
- Health records data are retrospective, which can make it difficult to check for errors.

Questions have been raised about whether digitizing individuals' electronic health records will compromise their security and privacy. Will inappropriate usage be properly restricted? Will companies be able to acquire and share these data? If companies use the data to develop publications, will they later be liable to requests to make the primary data available to others? Another potentially difficult problem is that the merging of two datasets might make it possible to identify patients who have been "de-identified" in each.

Although these and other potential concerns must be addressed, the experience since implementation of 21 CFR Part 11 a decade ago is encouraging. Existing processes, standards, and computer systems have been largely effective in maintaining the accuracy, integrity, and privacy of data. Furthermore, there are grounds to believe that these experiences can be extended to the effective handling of individuals' electronic health records—as witnessed, for example, by the success of the U.S. Department of Veterans' Affairs in developing secure practices.

^a Frank W. Rockhold and Ronald I. Krall. 2006. "Trial summaries on results databases and journal publication" (letter). *Lancet* 367:1633–1635.

^b Food and Drug Administration. 2003. *Guidance for Industry, Part 11, Electronic Records; Electronic Signatures—Scope and Application*. Available at http://www.21cfrpart11.com/files/fda_docs/part11_final_guidanceSep2003.pdf.

^c Eve Slater, Director on the boards of Vertex Pharmaceuticals and Theravance, Inc., presentation to the committee, April 16, 2007.

biological data that are made publicly available as soon as they are generated. The rapid release of validated, high-quality data requires analysis and planning by the researchers who built the data-gathering and processing system (which requires that those researchers be rewarded for their efforts) and the design of systems that incorporate innovative automated data-quality assessment. In these cases, provisions may need to be made for continually updating data as errors are detected and improved methods are developed, resulting in databases that evolve as fields advance.

Table 2-2 summarizes the policies of federal agencies regarding data integrity and data sharing.

DATA INTEGRITY IN THE DIGITAL AGE AND THE ROLE OF DATA PROFESSIONALS

In the digital age, the methods used to maintain data integrity are increasingly complex. As new methods and tools are brought into practice, researchers are continually challenged to understand them and use them effectively. Furthermore, providing data to users inevitably becomes more involved as the size and complexity of databases increase. Because methods continually change as digital technologies evolve, researchers may be required to make a substantial investment of time in order to keep pace.

In some fields, the researchers themselves may be at the forefront of efforts to meet these data challenges, but in many fields the challenges are met at least in part by what we call in this report “data professionals.” These individuals have a very wide range of responsibilities for data analysis, archiving, preservation, and distribution.²² Often, they are the leaders in developing new methods of data communication, data visualization, educational outreach, and other key advances. They also often participate in the development of standards, formats, metadata, and quality control mechanisms. They can bring new perspectives on existing datasets or new ways of combining data that yield important advances. Through their familiarity with rapidly changing digital technologies, they can enhance the ability of others to conduct research. They also are in a unique position to make digital data available to the broadest possible range of researchers, educators, students, and the general public. Educational opportunities, viable career paths, and professional recognition all help ensure that data professionals are in a position to make needed contributions to research.

²² National Science Board. 2005. *Long-Lived Data Collections: Enabling Research and Education in the 21st Century*. Arlington, VA: National Science Foundation.

GENERAL PRINCIPLE FOR ENSURING THE INTEGRITY OF RESEARCH DATA

The new capabilities and challenges posed by digital technologies point to the need for a renewed emphasis on data integrity. The assumption that traditional practices will suffice is no longer tenable as digital technologies continue to transform the nature of research. Researchers must be aware of how the integration of digital technologies into research affects the quality of data. As the generation and dissemination of data become the primary objectives of some research projects, researchers need to find ways to validate the quality of those data. They need to take steps to ensure that digital technologies enhance rather than detract from data integrity.

These observations lead to the following general principle:

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.

In emphasizing the importance of this principle, the committee is not calling for formal assurances of data integrity. Maintaining the quality of research is an essential part of being a responsible and competent researcher. In assigning researchers the ultimate responsibility for data integrity, the committee is asking no more than that researchers adhere to the standards established and held in common by all researchers.

This principle may seem apparent, but its application in the digital age leads to several important recommendations.

THE OBLIGATIONS OF RESEARCHERS TO ENSURE THE INTEGRITY OF RESEARCH DATA

Researchers have a fundamental obligation to their colleagues, to the public, and to themselves to ensure the integrity of research data. Members of the research community trust that their colleagues will adhere to the standards of their field and will be transparent in describing the methods used to generate data. They also assume that colleagues will make available the data on which publicly disseminated research results are based. (Chapter 3 discusses issues of data access in detail.) Members of the general public may be unfamiliar with the standards of a research field, but they, too, trust that researchers will gather, analyze, and review data accurately, honestly, and without unstated bias. If trust among colleagues or the public is misplaced and research data are shown to be inaccurate (or, even worse, fabricated), the consequences can be severe both within science and in the broader society.

TABLE 2-2 Federal Agency Policies on Research Data

Intramural

Are data subject to outside peer review?^b
 Are data sets required to be made available or deposited into appropriate repositories?
 Does training of new scientists include scientific misconduct training?

^a Includes full-time employees of DOE national laboratories owned by the federal government but operated by Management and Operating (M&O) contractors.

^b Presumes work will be published in a peer-reviewed publication.

^c Scientific misconduct training information available for the Jet Propulsion Lab, but not for other facilities.

Extramural Grants^d

	NIH ^b	NSF	USDA ^c	DOC
Are grantees required to share data with other researchers? ^e	Yes	Yes ^f	No ^g	No ^g
Are grantees required to deposit data sets in appropriate repositories?	Yes	Yes ⁱ	No ^g	No ^g
Are grantees required to submit all information regarding computer programs developed or used during the time frame of the grant?	Not Stated	Encouraged	No ^g	No ^g
Are printed “research misconduct” statements in effect, or a link provided to the federal policy?	Yes	Yes	Yes	Yes

^a As a baseline, federal agencies follow OMB Circular A-110, *Uniform Administrative Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations*, which specifies that the Federal Government has the right to obtain, reproduce, publish or use the data first produced under an award, and to authorize others to receive, reproduce, publish or use data. The provisions of the Data Access Act, described in Chapter 3, also apply.

^b NIH’s policy covers “final research data.” Applications seeking more than \$500,000 in direct costs in any single budget period are expected to include a plan for data sharing or state why data sharing is not possible.

^c Entries for this column apply to USDA’s Cooperative State Research, Education, and Extension Service, and may not apply to other parts of USDA.

^d Includes non-NIH grants.

^e Privacy and national security-related exceptions are assumed.

^f Sharing is “expected.” The policy also provides for some exceptions in addition to privacy.

^g No agency-wide written requirement, but sharing is often informally encouraged, and written requirements may cover some specific programs, grants or categories of data (e.g. requirements that genomic data be submitted to GenBank).

^h HHS “expects and supports” sharing of data and tools, including deposit of data into appropriate repositories.

ⁱ Sharing is expected, however, the NSF policy permits necessary flexibility to account for programmatic differences.

NIH	NASA	EPA	NIST	DOE ^a
Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes
Yes	Not stated ^c	Not stated	Not stated	Yes

SOURCES: The table assumes, as a baseline, that agencies have or will implement John H. Marburger, III. 2008. "Principles for the Release of Scientific Research Results." Memorandum. May 28. Available at: www.arl.org/bm~doc/ostp-scientific-research-28may08.pdf. Also see Web sites for NIH (<http://www1.od.nih.gov/oir/sourcebook/ethic-conduct/ethical-conduct-toc.htm>) and JPL (<http://ethics.jpl.nasa.gov/index.html>).

AFOSR	ONR	DOEd	DOE	HHS ^d	EPA	NASA
Not stated	Not stated	No	No ^g	Yes ^b	Yes	Yes
Not stated	Not stated	Not applicable	No ^g	Yes ^b	No	Yes
Not stated	Not stated	No	No ^g	Yes ^b	Yes	Not stated
Yes	Yes	Yes	Yes	Yes	Yes	Yes

SOURCES: Agency Web sites checked December 2008, and communications from agencies 2009.

- NIH: http://grants.nih.gov/grants/policy/nihgps_2003/NIHGPS_Part7.htm
- NSF: http://www.nsf.gov/pubs/policydocs/pappguide/nsf09_1/aag_index.jsp
- USDA: http://www.nsf.gov/pubs/policydocs/rtc/csrees_708.pdf
- DOC: http://oamweb.osec.doc.gov/GMD_grantsPolicy.html
- AFOSR: <http://www.wpafb.af.mil/library/factsheets/factsheet.asp?id=9447>
- ONR: <http://www.onr.navy.mil/02/terms.asp>
- DOEd: <http://www.ed.gov/fund/landing.jhtml?src=ln>
- DOE: <http://www.sc.doe.gov/grants/grants.html#GrantRules>
- HHS: http://www.hhs.gov/grantsnet/docs/HHSGPS_107.doc
- EPA: <http://www.epa.gov/ogd/grants/regulations.htm>
<http://epa.gov/ncer/guidance/>
- NASA: <http://www.hq.nasa.gov/office/procurement/nraguidebook/>

The twin ideals of trust and transparency lead to our first recommendation:

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.²³ Others 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 to maintain the integrity of research data.

THE IMPORTANCE OF TRAINING

The integrity of research data can suffer if researchers inadvertently or willfully ignore the professional standards of their field. Data integrity also can be negatively affected if researchers are unaware of these standards or are unaware of their importance.

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.

The training that is appropriate for researchers varies by field. While every researcher should be familiar with the standards common to all research, other standards may be unique to a particular field. Much of this knowledge is handed down from senior researchers to junior researchers during the course of a person's education and research apprenticeship. In at least some fields, a more formal statement of accepted practices, combined with more explicit instruction in those practices, could enhance the quality and utility of the data produced by those fields. Given the rapid pace of change in many research fields, research focused specifically on methods to ensure the integrity of research data may be necessary.

Today, the actual implementation of training varies greatly from field to field and institution to institution. The National Institutes of Health (NIH)

²³ 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.

requires that graduate and postdoctoral students who are supported by NIH training grants receive instruction in the responsible conduct of research. The Office of Research Integrity at the Department of Health and Human Services supports programs undertaken by the Council of Graduate Schools, the National Postdoctoral Association, and the Laboratory Management Institute at the University of California at Davis to develop education and training programs in the responsible conduct of research.²⁴ Many research institutions also require such training of students or beginning researchers, often in the form of seminars, workshops, or Web-based modules. (Box 2-4 describes one such program.)

A 2002 Institute of Medicine report examined how institutions can create environments that foster research integrity.²⁵ The report points out that although education and training can be helpful, not much is currently known about which approaches are most effective. Institutional self-assessment and external peer review can be valuable tools in developing and improving education and training. Smaller institutions may need to take advantage of consortia or electronic communications to provide their researchers with adequate education and training.

The leaders of research groups have a particular responsibility to see that professional standards are observed in the conduct of research. They should ensure that the members of their groups have opportunities to learn about the proper management of data. Research leaders also have an obligation to set a standard for responsible behavior and to monitor and guide the actions of the members of their groups. Implementing institutional policies at the group level, holding regular meetings to discuss data issues, and providing careful supervision all help to create a research environment in which the integrity of data is understood, valued, and ensured.²⁶

As described earlier, the need for training in the standards of research has been made more urgent by the advance of the digital age. The application of digital technologies in research has fundamentally altered the daily practices and interpersonal interactions of everyone involved in the research enterprise. Researchers need to become familiar with complex and rapidly changing systems to review, visualize, store, summarize, and search for information. They need to understand the technologies and methods they apply to the collection, analysis, storage, and dissemination of data in sufficient detail to have confidence in the integrity of those data. Unless they understand the procedures used to generate, process, represent, and document data, they risk wasting

²⁴ Office of Research Integrity. 2008. Annual Report 2007. Washington, DC: Department of Health and Human Services.

²⁵ Institute of Medicine. 2002. *Integrity in Scientific Research: Creating an Environment That Promotes Responsible Conduct*. Washington, DC: The National Academies Press.

²⁶ Chris B. Pascal. 2006. "Managing data for integrity: Policies and procedures for ensuring the accuracy and quality of the data in the laboratory." *Science and Engineering Ethics* 2:23–39.

BOX 2-4

Training in Data Management

The program *Fostering Integrity in Research, Scholarship, and Teaching (FIRST)* at the University of Minnesota includes an online workshop in research data management. New faculty members, postdoctoral fellows, and graduate students who are acting as principal investigators or otherwise have responsibility for the management of data are required to take the workshop, which takes about an hour to complete.

The workshop is organized around four online case studies in the following areas: ensuring data reliability, controlling access to data, maintaining data integrity, and following retention guidelines. The case study on data retention, for example, is the following:

A group of scientists gathered new research data and published their findings. This exciting research led to a rethinking of some fundamental aspects of superconductivity, and generated a significant amount of discussion. About 3 years after the original publication date, however, a suggestion for a different interpretation of the data was made. To prove that the initial interpretation was correct, the principal investigator (PI) from the project decided to reevaluate the data taken 5 years earlier. Unfortunately, the raw data had been destroyed after they were entered into the computer, and the computer files were thrown out with the computer 1 year ago.

Each case study is followed by a series of questions to answer and links to additional information. Pages that provide answers to frequently asked questions and an opportunity to send additional questions to experts in the responsible conduct of research provide additional resources.

For more information, see <http://www.research.umn.edu/datamgtq1/index.htm>.

resources or reducing the quality of their data and research conclusions. In a profession so dependent on advanced computing and communications, every researcher needs to understand not only how to use computers but how computing affects research.

PRODUCING CLEAR, UP-TO-DATE STANDARDS FOR DATA INTEGRITY: A SHARED RESPONSIBILITY OF THE RESEARCH ENTERPRISE

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 professional standards either have not yet been established or are in flux.²⁷ The research enterprise needs to redouble efforts 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.

To date, research communities have responded to the new challenges of the digital age in a largely decentralized fashion, adapting traditional ethical standards to new circumstances. This decentralized approach is appropriate in that data management practices are so varied across research fields that a “one size fits all” approach would not address important issues, and the imposition of detailed standards from outside a field is unlikely to be effective. In some cases, fields of research within and across disciplines may be able to cooperate in developing standards for ensuring the integrity of research data.

The application of professional standards can be complicated in the case of interdisciplinary research, where investigators in different fields bring different practices to joint projects. In this case, familiarity with the standards and expectations of all the fields represented by that research is preferable to the blanket imposition of overly broad standards. Better education and training in data management for investigators, combined with expanded access to research data across disciplines (which is the subject of the next chapter), will best serve the advance of knowledge and other public interests.

THE ROLES OF DATA PROFESSIONALS

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. Instead, they may have to rely on collaborations with colleagues who have specialized training in applying digital technologies in research. Through their in-depth knowledge of digital technologies and how those technologies can advance

²⁷ The quality standards applied to microarray data in proteomics provide a good example of ongoing efforts to improve the data generated by a rapidly evolving technology. See S. Rogers and A. Cambrosio. 2007. Making a new technology work: The standardization and regulation of microarrays. *Yale Journal of Biology and Medicine* 80:165–178.

knowledge in a particular field, data professionals can make key intellectual contributions to the progress of research.

Data professionals have a wide range of backgrounds, levels of training, and roles in research. Some serve in a support role for research groups; others make substantial intellectual or other contributions to research that warrant professional rewards such as inclusion in a list of authors. The roles of data professionals vary from field to field, but in an increasing number of fields, data professionals are assuming a shared professional responsibility with researchers for maintaining the integrity of research data. Chapters 3 and 4 return to the roles of data professionals in enabling access to and preserving research data. The following recommendation reflects their importance in ensuring data integrity.

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 research cost in an increasing number of fields.

3

Ensuring Access to Research Data

The advance of knowledge is based on the open flow of information. Only when a researcher shares data and results with other researchers can the accuracy of the data, analyses, and conclusions be verified. Different researchers apply their own perspectives to the same body of information, which reduces the bias inherent in individual perspectives. Unrestricted access to the data used to derive conclusions also builds public confidence in the processes and outcomes of research.

Furthermore, scientific, engineering, and medical research is a cumulative process. New ideas build on earlier knowledge, so that the frontiers of human understanding continually move outward. Researchers use each other's data and conclusions to extend their own ideas, making the total effort much greater than the sum of the individual efforts. Openness speeds and strengthens the advance of human knowledge. As an example, Box 3-1 describes how the sharing of genomic data has advanced life sciences research.

Finally, only by sharing research data and the results of research can new knowledge be transformed into socially beneficial goods and services. When research information is readily accessible, researchers and other innovators can use that information to create products and services that meet human needs and expand human capabilities. The Organisation for Economic Co-operation and Development (OECD) describes a new effort to enhance public access to research data (see Box 3-2). According to this approach, "Openness means access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based."¹ As the National Research Council's

¹ "OECD Principles for Access to Research Data from Public Funding," Available at <http://www.oecd.org/dataoecd/9/61/38500813.pdf>.

BOX 3-1 Access to Genomic Data

In biology, the culture of research and the applications of digital technologies have traditionally been heterogeneous, independent, and dispersed. However, the growth of interdisciplinary research, the advent of projects that have generated large volumes of data, and the invention of data-intensive devices such as DNA microarrays and high-throughput sequencers have highlighted the increasing importance of digitization of the biomedical sciences.^a

In the field of genomics, strong forces have pushed in the direction of unrestricted access to data, including directives from funding agencies, requirements from journals that researchers submit data to public repositories, community expectations, and the development of powerful data-sharing systems such as PubMed. In the case of the human genome, for example, the desire by funding agencies, researchers, and the general public for public access to research data led the genomics research community to develop an ethic of unrestricted access. This ethic was formally adopted as the “Bermuda statement” in February 1996:

All human genomic information produced at large-scale sequencing centres should be freely available and in the public domain, in order to encourage research and development and to maximize its benefit to society.^b

At the same time, other forces have had the effect of restricting access to genomics data, including:

- The need to protect patient or individual privacy;
- The principal investigator’s desire to maintain research advantage;
- The danger of misuse (e.g., of virus sequences);
- A profit motive (for data with potential commercial value);
- The tendency to “publish and forget” used data, especially supplementary data.

Committee on Issues in the Transborder Flow of Scientific Data stated in its report *Bits of Power: Issues in Global Access to Scientific Data*, “The value of data lies in their use.”²

The norms and traditions of research reflect the value of openness. Researchers receive intellectual credit for their work and recognition from their peers—and perhaps from the broader community of researchers and the public—when they publish their results and share the data on which those results are based. Some

² National Research Council. 1997. *Bits of Power: Issues in Global Access to Scientific Data*. Washington, DC: National Academy Press.

The generation of complete genome sequences for a growing number of organisms has intensified the digitization of biomedical research. These data have many applications in both basic and applied research, with the lines between the two often being difficult to discern. For example, computational processing and reference to information and knowledge bases about organisms and disease processes allow researchers to reach faster conclusions about the likely results of a therapy.^c The combination of cellular data, genomic profiling, and biological simulation may reduce the failure rate of drug candidates and the cost of testing. In the near future, it will even be possible, given sufficient computing and storage resources, to record the genotype of each person in a secure database. Variations in genes may indicate specific disease susceptibility or responses to known drug types. This information could enable physicians to prescribe a personal immunization and screening schedule or to recommend specific preventive measures for each patient.

Further integration of the biomedical sciences using digital technologies could allow independent investigators to remain the engine of innovative research by participating in “virtual team science.” Early examples of such “cyberinfrastructure”—including the Biomedical Informatics Research Network, myGrid, and the cancer Biomedical Informatics Grid—indicate that it is technically feasible, if not easy, to integrate the many threads of biomedicine. The challenge is to ensure that new “cybersilos” do not replace existing disciplinary and institutional silos.^d

^a “The race to computerize biology.” 2002. *Economist*, Dec. 12, 2002.

^b David R. Bentley. 1996. “Genomic sequence information should be released immediately and freely in the public domain.” *Science* 274:533–534. This statement was written on behalf of the Sanger Institute at the Wellcome Trust Genome Campus and the Genome Sequencing Center at Washington University in St. Louis.

^c Chris Sander. 2000. “Genomic medicine and the future of health care.” *Science* 287:1977–1978.

^d Kenneth H. Buetow. 2005. “Cyberinfrastructure: Empowering a ‘third way’ in biomedical research.” *Science* 308: 821–824.

journals require the submission and public dissemination of the data supporting an accepted manuscript. Funding agencies and research institutions also have policies that require the open sharing of the data on which research conclusions are based. Codes of conduct in a research community, whether explicit or tacit, can exert a powerful influence on researchers to make data accessible.

Advances in information technology—for instance, the advent of grid computing and cloud computing³—will continue to transform the environment for

³ In grid computing, distributed computing resources link experimental apparatus, processing, analysis, and storage; cloud computing involves large-scale, data-intensive, Internet-hosted applications and related infrastructure.

BOX 3-2

OECD Principles and Guidelines for Access to Research Data from Public Funding

From 2004 to 2006 the 30-nation Organisation for Economic Co-operation and Development (OECD) developed a set of guidelines based on commonly agreed principles to facilitate cost-effective access to digital research data generated through public funding. Endorsed by the OECD Council on December 14, 2006, the “OECD Principles and Guidelines for Access to Research Data from Public Funding” serve as objectives for each member country to achieve given its own legal, cultural, economic, and social context.

The Principles and Guidelines cover 13 broad areas:

- Openness
- Flexibility
- Transparency
- Legal conformity
- Protection of intellectual property
- Formal responsibility
- Professionalism
- Interoperability
- Quality
- Security
- Efficiency
- Accountability
- Sustainability

The Principles and Guidelines call “for a flexible approach to data access” under a default principle of openness and recognize “that one size does not fit all.” They also state that “Whatever differences there may be between practices of, and policies on, data sharing, and whatever legitimate restrictions may be put on data access, practically all research could benefit from more systematic sharing.”

NOTE: For more information, see Organisation for Economic Co-operation and Development. 2007. OECD Principles and Guidelines for Access to Research Data from Public Funding. Available at <http://www.oecd.org/dataoecd/9/61/38500813.pdf>.

research and lower the technical barriers to sharing data. As this transformation occurs, researchers are organizing their work in new ways to take advantage of new possibilities. An innovative example is the conduct of research in what can be called an open-knowledge environment.⁴ Building on the methodology pioneered by the open-source software movement, this approach begins with

⁴ *Economist*. 2004. “An Open-Source Shot in the Arm?” June 10.

the identification of a problem that is to be examined in a public forum on the Internet. Researchers from different disciplines, organizations, and countries then can all contribute to solving the problem, with the open sharing of data and ideas that might bear on that problem. An open-knowledge environment allows people with many different backgrounds and viewpoints to interact in a relatively unstructured way while moving toward a common objective. The free flow of information speeds progress, while the global reach of the Internet greatly expands the number and breadth of researchers who can contribute to a project. Another approach to sharing is open-notebook science.⁵ Similarly, blogs, wikis, and other forms of electronic interaction are tools that enable collaborative work on common problems in a generally open research environment.

In the context of this report, sharing research data enhances the data's integrity by allowing other researchers to scrutinize and verify them (as described in the Chapter 2). Sharing also increases the likelihood that data will be preserved for long-term uses, although the stewardship of data requires more than that the data be accessible (as described in the Chapter 4). Thus, the three themes of this report—integrity, accessibility, and stewardship—are intertwined.

BARRIERS TO SHARING DATA

Despite the many benefits to be gained by the sharing of research data and results, even a cursory survey of research activity reveals many circumstances in which access to data is limited.

Because researchers require time to verify data, analyze their data, and derive research conclusions, individual researchers generally are not expected to make all their data public immediately. Individual researchers need latitude to follow hunches, experiment with methods, explore conjectures, and make mistakes. New tools for automatically assessing the quality of data and sharing them with others can facilitate the rapid sharing of digital data, although verifying the reliability of these tools presents its own set of challenges.

Once a research result is published, the norms of science—and often the terms of the research grant or contract—call for the supporting data to be accessible. Researchers may nevertheless try to keep the data private, perhaps to derive additional results without competition from others, for the exclusive use of a student or postdoctoral fellow whose career would be advanced by generating further papers, or just to avoid the effort to put the data in usable form for others. In the worst cases, they may retain data to hide acts of research misconduct or to conceal defects in the dataset.

The norms of a research community may allow keeping data private for a certain period. These norms can be formalized through the terms of a grant

⁵ Katherine Sanderson. 2008. "Data on display." *Nature*. 455:273.

giving the investigator a defined period of exclusive use of the data, with the exclusivity ending upon the publication of results, after a particular length of time, or when data are deposited in a data center or archive.

There is great variation among research fields in their data-sharing norms, to such an extent that different fields can be said to have different data cultures. (Box 3-3 describes aspects of the data culture in economics.) A recent report commissioned by the Research Information Network of the United Kingdom examined data-sharing practices and expectations across a number of fields (Table 3-1).⁶ The report highlights the global importance and relevance of data accessibility in research, as well as the fact that differences between fields are often more important than national differences in determining data-sharing practices. The international aspects of data access and sharing are discussed in more detail below.

Observational astronomy offers a good example of the data-sharing norms that can characterize a field of research. Astronomical data often can be used for multiple purposes and are usually made public, but proprietary periods in which only the members of a research team have access to data are common. The European Southern Observatory (Europe's large optical observatory) and the National Aeronautics and Space Administration have 12-month proprietary periods. The U.S. National Optical Astronomy Observatory has an 18-month proprietary time. These periods provide researchers with an opportunity to make discoveries as a reward for dedicating significant periods of their careers to creating new facilities and developing new techniques. They also provide an opportunity for critical evaluation of the data before they are released.

In the high-energy physics community, collaborations are so large and the experiments so complex—with hundreds of scientists involved with the operation of a single detector—that it could take years for an independent scientist to learn enough to reanalyze the data. The data of each collaboration are treated as proprietary. Other groups that want to undertake the same measurement must form their own large collaboration and repeat the experiment. As explained in Box 2-1, large collaborations in high-energy physics involve elaborate procedures for internal scrutiny of and validation of data.

Cultural norms and expectations in research fields regarding data can change over time. For example, as data sharing has proven increasingly valuable to the advancement of research in many areas of the life sciences, researchers, sponsors, research institutions, and other stakeholders have built new infrastructure and established guidelines to facilitate data sharing. A 2003 National Research Council study (Box 3-4) recommended guidelines for the sharing of

⁶ Alma Swan and Sheridan Brown. 2008. *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs*. Report Commissioned by the Research Information Network. June. Available at: <http://www.rin.ac.uk/data-publication>.

BOX 3-3 Data Sharing Within Economics

Economists rely on an enormous variety of research data—for instance, administrative data from government records, datasets provided by companies to the federal government, or data provided directly to researchers by companies. Some economists rely on methods similar to those used by anthropologists, in which large quantities of data are collected and analyzed. Often the datasets are subject to confidentiality agreements because individuals could be identified from the data. Use of the data may even be restricted to “enclaves,” where a researcher has to work on a nonnetworked computer in a secure room from which materials cannot be removed.

Analysis of economic data may depend critically on highly complex computer programs. These programs, rather than the actual data, can be the most valuable part of an economist’s research, because many datasets are available publicly, whereas a computer program could embody months or years of individual effort. Thus, to assess the original analysis, other researchers often need access to the computer programs as well as to the original data.

As in other sciences, the social sciences have an expectation of reproducibility—that if the data are available and analyzed with the same assumptions, the same results will emerge. But without considerable assistance from the original researchers, actual replication of published results in economics can be time-consuming, tedious, and subject to many errors. Furthermore, journals are reluctant to publish studies that are confirmatory rather than groundbreaking. Social scientists, like other scientists, are more interested in doing their own studies and getting credit for something new than in repeating work that has already been done.

Even if replication is not common, the data should be available to enable replication, but in economics this often is not the case.^a Several years ago two economists wrote to the authors of every paper in the March 2004 issue of the *American Economic Review*, a leading journal in the field, and requested the data to replicate the research. Although the journal has a statement saying “Authors are required to maintain their data and supply it to other researchers upon request,” 14 of the 15 sets of authors to whom the economists wrote said that they did not have the data or would not share them. The authors summarized their findings in an article and submitted it to the *American Economic Review*, which published their paper.

As a result of this and other cases, the *American Economic Review* adopted a new policy. For published articles, the authors must provide both the data and the programs sufficient for the articles’ findings to be replicated. These data and programs are then posted on the journal’s Web site. If the use of the data is restricted, the authors must provide instructions on how to obtain permission to use the data. If some of the data are proprietary, the editors try to work out ways for other researchers to use the data. In addition, the journal is encouraging studies to reanalyze data and replicate results.

The *American Economic Review* is supported by dues from 20,000 members and has the resources to institute such a policy, whereas journals with fewer resources could have difficulty adopting and enforcing the same or similar policies. Also, the data and programs are not requested at the time of submission of an article—only upon acceptance—so that the 92 percent of the papers submitted to the journal that are rejected do not fall under the new guidelines. Some economists have decided not to submit a paper to the *American Economic Review* because they do not want to release their data or software. Nevertheless, because authors want to publish their papers in the journal, it has considerable influence over their actions.

^a Robert A. Moffitt, *American Economic Review*, Presentation to the committee, April 17, 2007.

TABLE 3-1 Summary of the Data-sharing Environment in Various Fields in the United Kingdom

	Culture of sharing data	Infrastructure-related barriers to publishing data	Effect of policy initiatives to encourage data publishing	Overall propensity to publish datasets (with appropriate metadata and contextual documentation)
Astronomy	Strong culture of sharing	Low level of barriers	Policy has medium positive effect	Strong propensity to publish datasets
Chemical crystallography	Medium culture of sharing	Low level of barriers	Policy has little positive effect	Strong propensity to publish datasets
Genomics	Strong culture of sharing	Low level of barriers	Policy has strong positive effect	Strong propensity to publish datasets
Systems biology	Medium culture of sharing	Moderate level of barriers	Policy has strong positive effect	Medium propensity to publish datasets
Classics (Humanities)	Strong culture of sharing	High level of barriers ^a	Policy has medium positive effect	Medium propensity to publish datasets
Social and Public Health Sciences	Weak culture of sharing	Low level of barriers	Policy has little positive effect	Low propensity to publish datasets ^b
RELU ^c	Medium culture of sharing	Low level of barriers	Policy has medium positive effect	Medium propensity to publish datasets
Climate Science	Weak culture of sharing	Low level of barriers ^d	Policy has medium positive effect	Low to medium propensity to publish datasets

^a The Arts and Humanities Data Service was established in 1995 to provide a national service to collect, preserve, and promote electronic resources in the arts and humanities; its funding was eliminated in 2008.

^b This descriptor covers researchers not directly connected with a national data collection.

^c The Rural Economy and Land Use Program is a collaborative research program among several UK research councils.

^d The Natural Environment Research Council provides data centers.

SOURCE: © Research Information Network. 2008. To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. June. <http://www.rin.ac.uk/data-publication>.

BOX 3-4 Sharing Publication-Related Data and Materials

In 2003 the National Research Council Committee on Responsibilities of Authorship in the Biological Sciences released a report that focused directly on the issues discussed in this chapter. In that report, the committee established what it called “the uniform principle for sharing integral data and materials expeditiously” (UPSIDE). They described this principle as follows:

Community standards for sharing publication-related data and materials should flow from the general principle that the publication of scientific information is intended to move science forward. More specifically, the act of publishing is a *quid pro quo* in which authors receive credit and acknowledgment in exchange for disclosure of their scientific findings. An author’s obligation is not only to release data and materials to enable others to verify or replicate published findings (as journals already implicitly or explicitly require) but also to provide them in a form on which other scientists can build with further research. All members of the scientific community—whether working in academia, government, or a commercial enterprise—have equal responsibility for upholding community standards as participants in the publication system, and all should be equally able to derive benefits from it.

The committee also identified five corollary principles associated with sharing publication-related data, software, and materials. For example, the committee stated that “authors should include in their publications the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims.”

The committee noted that its purview extended only to the biological sciences. It also stated, however, that “in the committee’s view, there should be a single scientific community that operates under a single set of principles regarding the pursuit of knowledge. This includes a common ethic with regard to the integrity of the scientific process and a long-held commitment to the validation of concepts of experimentation and later verification or refutation of published observations.”

SOURCE: National Research Council. 2003. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Washington, DC: The National Academies Press.

data and other information supporting research results that emphasize openness and expanded access, including research performed by companies.⁷

Although the charge to our committee excluded privacy and other issues

⁷ National Research Council. 2003. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Washington, DC: The National Academies Press.

related to human subjects from our study, it is important to note that these issues can act as barriers to data access. Some data are not released because of confidentiality or privacy considerations, such as data related to biomedical research or the social sciences. For example, the 1996 Health Insurance Portability and Accountability Act established rules for disclosure of individually identifiable health information (known as protected health information, or PHI).⁸ If PHI is used in research, the researcher must comply with regulations regarding its use and storage in the project. There are instances where PHI may be disclosed, but the need to support published research is not among them. For PHI to be made publicly available, a subject must agree to the disclosure of the information.

For some medical research data, privacy and confidentiality obstacles can be overcome by removing identifiers prior to the private sharing of data or the public release of data. However, this remains an area of ongoing concern and investigation. Efforts are now under way to make medical research data available while ensuring that the data cannot be used to identify individuals.

Research data also can be kept private because they pertain to intelligence, military, or terrorist activities.⁹ Examples include research related to nuclear, radiological, and biological threats; human and agricultural health systems; chemicals and explosives; and information technology infrastructure. National Security Decision Directive 189 (NSDD 189), which was issued by President Ronald Reagan in 1985, states that the policy of the U.S. government is not to restrict, to the maximum extent possible, the products of unclassified fundamental research.¹⁰ The challenge to policy makers and researchers is where to draw the line between classified and unclassified information and how to balance restrictions on access to sensitive information with the potential costs of such restrictions.

Our committee was not asked to examine national security issues in depth. Other National Research Council committees, including the Committee on Scientific Communication and National Security (CSCANS), are directly focused on issues such as classified information, export controls, and nonimmigrant visa policies. A recent CSCANS report points out that many federal government policies and practices since the September 11 attacks have effectively reversed NSDD 189.¹¹ The report calls for a standing entity to review policies in order to

⁸ Institute of Medicine. 2006. *Effect of the HIPAA Privacy Rule on Health Research: Proceedings of a Workshop Presented to the National Cancer Policy Forum*. Washington, DC: The National Academies Press.

⁹ National Research Council. 2007. *Science and Security in a Post 9/11 World: A Report Based on Regional Discussions Between the Science and Security Communities*. Washington, DC: The National Academies Press.

¹⁰ National Policy on the Transfer of Scientific, Technical and Engineering Information. September 21, 1985.

¹¹ *Ibid.*

ensure that the small risks of basic research being misused are balanced with the enormous benefits that accrue from the free exchange of information. Another National Research Council Committee examined the national security implications of access to genomic databases and found that unrestricted access, combined with the development of education programs by professional societies, is the best approach to balancing the advancement of knowledge with protecting the public from misuse of genomic data for bioterrorism threats.¹² The federal government's creation in 2008 of a new category—"Controlled Unclassified Information"—illustrates that restrictions on the sharing of research based on national security concerns will continue to pose challenges to the research enterprise.¹³

When research is carried out or sponsored by public agencies, the general presumption in the United States is that data generated as part of that research should be publicly available.¹⁴ Different considerations apply for research funded by a private company, whether that research occurs within a company or in the academic sector. Though some companies have been experimenting with the benefits of freely sharing results from proprietary research,¹⁵ many companies carefully guard this information as a trade secret and a potential source of commercial advantage. Similarly, an academic researcher may temporarily withhold data in order to file a patent or develop a commercial product, even when the research is publicly funded. These issues are discussed later in this chapter.

The cost of disseminating data can be a barrier to its use. Circular A-130 from the Office of Management and Budget (OMB) stipulates that government-generated data should be available to users at cost sufficient to recover the expense of dissemination but not higher.¹⁶ However, data from private sources, even when purchased by the federal government for research purposes, frequently have high distribution costs and restrictions on redistribution. These costs can be a significant problem for academic researchers who need access to large databases for modeling or data analysis.

Finally, research data may be kept private because the resources are lacking to make data collections available to the public. A project might generate data that could be valuable to researchers in the same or other fields, but the

¹² National Research Council. 2004. *Seeking Security: Pathogens, Open Access, and Genome Databases*. Washington, DC: The National Academies Press.

¹³ George W. Bush. 2008. "Designation and Sharing of Controlled Unclassified Information (CUI)." Memorandum for the Heads of Executive Departments and Agencies. May 9.

¹⁴ Paul F. Uhler and Peter Schröder. 2007. "Open data for global science." *Data Science Journal* 6:OD36–OD53.

¹⁵ Bernard Munos. 2006. "Can open-source R&D reinvigorate drug research?" *Nature Reviews Drug Discovery* 5:723–729.

¹⁶ Office of Management and Budget. No date. Management of Federal Information Resources. Circular A-130. Memorandum for Heads of Executive Departments and Agencies. Available at <http://www.whitehouse.gov/omb/circulars/a130/a130trans4.html>.

investigators who generated those data may not have the resources or capabilities needed to make them available. This is frequently the case in small-scale research that does not have funding set aside for such functions or does not have a robust data management component in place. Alternatively, the data may be available, but the essential metadata needed to understand and use those data may be missing, making the data useless for anyone outside the immediate research team.

In general, researchers have a strong incentive to release the results of research. Their own recognition and advancement in their field generally depend on public dissemination of those results. In contrast, researchers have traditionally had few incentives to make publicly available the data they generate in the course of research. However, those data may have great value for other researchers, and making data publicly accessible can speed the advance of knowledge.

THE COSTS OF LIMITING ACCESS TO DATA

Barriers that restrict access to data, such as withholding data or delaying their release, can result in substantial costs.¹⁷ Once data have been gathered from an instrument or compiled from other sources, it is obviously more cost-effective to share the data than to reconstruct or recompile them. Furthermore, resources spent accessing data then are not available for other research uses.

Limitations on research data also can be barriers to innovation, which incurs costs in the broader society.¹⁸ In today's economy, the creation of new goods and services often depends on access to research data. When access is withheld, economic innovation slows, reducing the returns to investments in research.

Limiting access to research data also hinders the kinds of interdisciplinary and international cooperation that has proven so productive in recent research. When data are restricted to a particular research team or field, other researchers not only cannot use the data but often cannot even ascertain the value of those data to their own research. Similarly, if students are unable to work with new research data, their education and training may be adversely affected.

Limitations on the accessibility of data invariably retard, and can even block, the process of verifying the accuracy of those data. As a result, the quality of the data could be lower than would be the case if they were freely available, again reducing the return on the investment in producing the data.

Finally, researchers who are deprived of access to data are disadvantaged in conducting research and possibly seeking support to do research. This

¹⁷ Uhler and Schröder, *op. cit.*, pp. OD42–OD43.

¹⁸ National Research Council. 1999. *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. Washington, DC: National Academy Press.

problem can be particularly acute in developing countries, where lack of access to data from developed countries can stymie not only the development of research capacity but advances in economic productivity, public health, and well-being.

DATA ACCESS ISSUES IN RESEARCH AFFECTING PUBLIC POLICY OR PRIVATE INTERESTS

Restricting access to data can be costly and wasteful, but there also are circumstances in which providing access to data can entail substantial costs and waste. There are situations in which responding to requests for data could actually slow the progress of research, and there have been instances in which requests for data have been intended to inhibit research.

It is not uncommon for a small research group to lack the resources to make data readily accessible. Especially as data collections grow in size and complexity, small groups may have difficulty providing data to other researchers in the same field, much less making data readily accessible to researchers in fields less directly connected to the research, or to the public.

Access to data can also become an issue of contention in cases where research has important implications for public policy or has a potential for affecting private interests in such areas as the environment or health. An early example was the case of Paul Fischer, who was subpoenaed in the early 1990s by a tobacco company after publication of his research showing widespread recognition among young children of the “Joe Camel” character used in cigarette advertising.¹⁹ Fischer initially was subpoenaed in a lawsuit to which he was not subject. In addition to requesting details about the research that would be considered reasonable and necessary to replicate the results, the subpoena contained more problematic demands, such as personal details about the subjects. According to his own account, Fischer’s institution, the Medical College of Georgia, refused to provide legal support. After Fischer, using his own attorney, had quashed the subpoena, the Medical College of Georgia’s counsel wrote an article that had the effect of alerting R.J. Reynolds to an alternative legal mechanism, the Georgia Open Records Act. Under this act, Fischer ultimately was compelled to turn over all the information except the children’s names.

Perhaps the most famous recent example involved a research project to reconstruct global temperature trends over the last two millennia. A 1998 paper by Michael Mann of Pennsylvania State University and two co-authors made extensive use of proxy studies in which paleoclimatic conditions were inferred from measurements of tree rings, sediments, coral, glaciers, oxygen isotopes, and other phenomena, concluding that global surface temperatures

¹⁹ Paul M. Fischer. 1996. “Science and subpoenas: When do the courts become instruments of manipulation?” *Law & Contemporary Problems* 29:159–167.

were relatively stable for 900 years and then rose rapidly between 1900 and 2000, providing a fingerprint for human-caused climate change.²⁰

After the release of the Third Assessment Report of the Intergovernmental Panel on Climate Change that cited this finding in 2001, it became a point of contention in debates over the reality and causes of global warming. Mann resisted when researchers skeptical of his work requested access to the underlying data and computer programs used in the reconstruction, and controversy ensued.²¹ Two Members of the U.S. House of Representatives issued a letter requesting a wide variety of information from each of the three co-authors of the paper, giving them 18 days to provide, among other things, a curriculum vitae with a list of all studies they authored on climate change and the specific sources of funding; a list of all financial support received from private, state and federal sources for climate-related work; the location of all underlying data archives related to such research and its specific availability; correspondence regarding requests for such data from other researchers, responses to such requests and the researchers' reasons for their decisions, and in-depth responses to inquiries about their work on bristlecone pines and the Intergovernmental Panel on Climate Change.²² This request was viewed by some as intimidation.²³ The National Research Council released a study in 2006 that examined the rapidly emerging field of multiproxy paleoclimate studies.²⁴ The report ultimately affirmed some, but not all, of the key results of Mann's work, while stating that "all research benefits from full and open access to published datasets and . . . a clear explanation of analytical methods is mandatory." The report also points to the need for researchers, professional societies, journals, and research sponsors involved in paleoclimate research to improve access to data and methods.²⁵

This is not an isolated example of a research field with highly charged policy implications. Research data and findings have a substantial influence on a growing number of issues, ranging from arms control to air quality, endangered species, environmental toxins, and school vouchers.²⁶ In many of these cases, researchers are being asked to contribute information in areas

²⁰ Mann, Michael E., Raymond S. Bradley, and Malcolm K. Hughes. 1998. "Global-scale temperature patterns and climate forcing over the past six centuries." *Nature* 392: 779–787.

²¹ Geoff Brumfiel. 2006. "Academy affirms hockey-stick graph." *Nature* 441:1032. The researchers requesting the data and other information were Stephen McIntyre and Ross McKittrick.

²² Letter from Representatives Joe Barton and Ed Whitfield to Dr. Raymond S. Bradley, June 23, 2005. Available at http://republicans.energycommerce.house.gov/108/Letters/062305_Bradley.pdf.

²³ Letter from Dr. Alan I. Leshner to Representative Joe Barton, July 13, 2005. Available at <http://www.aaas.org/spp/cstc/docs/05-7-13climatebarton.pdf>.

²⁴ National Research Council. 2006. *Surface Temperature Reconstructions for the Last 2,000 Years*. Washington, DC: The National Academies Press.

²⁵ The wording in this paragraph has been changed to correct some factual errors.

²⁶ See the list of "Examples of Political Interference in Science" maintained by the Union of Concerned Scientists at http://www.ucsusa.org/scientific_integrity/interference/a-to-z-alphabetical.html.

where government has responsibility for public health and well-being, such as environmental quality regulations or the legal responsibility of manufacturers for product harms. In these areas, the role of research is increasingly being challenged by those who oppose particular regulations, laws, or legal rulings.²⁷ These cases raise important and difficult questions: When are researchers justified in withholding underlying data and methods? What recourse do colleagues, policy makers, and the public have when data or methods underlying research on important policy issues are withheld? What is the line between harassment that unreasonably slows the pace of research and justified requests for information?

These trends point to the need for clearer standards and understandings between researchers, their employers, and the public about the overarching value of openness, as well as the circumstances under which requests or demands for data are reasonable and when they cross the line into the realm of harassment that can slow the advance of knowledge. There are important and complex questions about how to balance the need for important data to be widely accessible, with fundamental issues of academic freedom, confidentiality, and the need for researchers to carry out their studies free of harassment, intimidation, or outside pressure.

OWNERSHIP OF RESEARCH DATA AND RELATED PRODUCTS

Addressing the question of “who owns research data” is a key element of the authoring committee’s charge. There is a range of possible answers, including the researcher, the institution, the sponsor, or nobody, depending on the particular meaning of “ownership” and the context. This section will review the laws and policies relevant to the ownership of research data and related rights to control its dissemination and use. The next section will cover other laws and policies related to research data, focusing on obligations to keep or share data.

To begin with, general principles of property law apply to the media on which data are stored and may also apply to the bits themselves in the case of digital data. One analogy is to the master recording in the music business when analog technology dominated. The owner of the master tape has a property right in the object but does not necessarily own the copyright that controls the copying and distribution of the data stored in the recording. Similarly, the researcher, his or her institution, or the sponsor (depending on the terms of the research grant or contract) may own the medium on which the data are stored.

More important, for the purposes of this discussion, than ownership of the physical storage media are intellectual property rights in a database (some

²⁷ Wendy Wagner and Rena Steinzor, eds. 2006. *Rescuing Science from Politics: Regulation and the Distortion of Scientific Research*. New York: Cambridge University Press.

specific arrangement or organization of the data), in a publication whose central ideas are based on the data, or in an invention that is based on the data. We will consider each of these related issues in turn.

Copyright, Database Protections, and Licensing

In the United States, copyright protection is extended to “original works of authorship fixed in any tangible medium of expression. . . .”²⁸ Copyright holders enjoy the exclusive right to disseminate their creations and to earn a profit by selling or licensing them. Raw data and other facts, however, are not protected as copyrightable subject matter. Databases are copyrightable if the selections or arrangement are original; the mere compilation of facts or data into a collection does not entitle them to protection. These provisions were reinforced by the 1991 Supreme Court ruling in *Feist Publications, Inc. v. Rural Telephone Service Co.*, which limited copyright protection for databases to those arranged and selected in an original manner.²⁹ In addition, the federal government is prohibited from exerting copyright protection over its own publications, including data generated by government entities. Finally, copyright law includes provisions for “fair use” exceptions in which portions of a copyrighted work may be used without permission in teaching, research, and other specified pursuits.

This basic framework has served to support the open flow of research data. Federal agencies have been central in sustaining a strong public domain in data.³⁰ With regard to research data, private companies and nonprofit entities play an important role in creating databases and information services that are utilized by researchers. The existence of copyright protection for creative and original data collections provides an incentive for investments in valuable products and services in the private sector.

Digital technologies have introduced new considerations into copyright laws and enforcement.³¹ Technological barriers to violating copyrights have fallen, posing challenges to copyright-based industries such as music, newspapers, and motion pictures. Before the digital age, the trigger for a copyright violation of a printed document was the act of copying. A photocopy of a document for personal use falls under the fair-use provisions, but copying now can be done almost effortlessly. If a document is made into a PDF file that can be circulated on the Internet, the distinction between private use and publication vanishes.

²⁸ U.S. Code, Title 17, Chapter 1, Section 102. Available at http://www4.law.cornell.edu/uscode/html/uscode17/uscode_sec_17_00000102----000-.html.

²⁹ 499 U.S. 340 (1991). Available at <http://laws.findlaw.com/us/499/340.html>.

³⁰ National Research Council. 2003. *The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium*. Washington, DC: The National Academies Press.

³¹ National Research Council. 2000. *The Digital Dilemma: Intellectual Property in the Information Age*. Washington, DC: National Academy Press.

Digital technologies have also made possible new approaches to commercializing the provision of data and data services.³² Several legal and policy changes of recent years have strengthened the position of copyright holders. These include lengthening of the term of copyright protection and the passage of the Digital Millennium Copyright Act of 1998 (DMCA). The DMCA implemented the World Intellectual Property Organization treaty on copyrights, and criminalized the circumvention of technical measures to prevent copying of digital materials, even in the absence of actual copying. These technical measures include hardware and software-based access controls, increasingly effective forms of encryption, and other forms of digital rights management that limit access to or copying of data.

In addition, in 1996 the European Community enacted a Directive on the Legal Protection of Databases that established a framework for new proprietary rights specific to databases.³³ Experts have warned that a combination of expanded copyright protections, advances in technological means of restricting access to digital content, and database protections of the type that Europe has adopted could enable the assertion and enforcement of proprietary claims to factual matter that previously entered the public domain as soon as it was disclosed.³⁴ The United States and many other countries have not followed the European Union in establishing a new intellectual property regime for databases.

An area where advancing technology and the increased use of contracts and licensing have changed the environment for access is remote sensing and geographic data and services.³⁵ Federal agencies have traditionally acquired full ownership rights to geographic data (such as maps and books) from private entities and have allowed that information to enter the public domain so as to be accessible without restrictions to other uses. However, as digital media have become more prevalent, private data providers have moved to business models focused on selling multiple licenses and access subscriptions to databases. A 2004 National Research Council report recommended approaches agencies should take to licensing geographic data and services in order to maximize their utility, including a recommendation that the federal government should foster

³² National Research Council. 2004. *Licensing Geographic Data and Services*. Washington, DC: The National Academies Press.

³³ Commission of the European Communities. 2005. First Evaluation of Directive 96/9/EC on the Legal Protection of Databases. DG Internal Market and Services Working Paper. December 12. Available at http://ec.europa.eu/internal_market/copyright/docs/databases/evaluation_report_en.pdf.

³⁴ J. H. Reichman and Paul F. Uhlir. 2003. "A contractually reconstructed research commons for scientific data in a highly protectionist intellectual property environment." *Law and Contemporary Problems* 66:315–462.

³⁵ See National Research Council. 2002. *Toward New Partnerships in Remote Sensing: Government, the Private Sector, and Earth Science Research*. Washington, DC: The National Academies Press.

the creation of a National Commons and Marketplace in Geographic Information.³⁶ Such an approach might be relevant to other fields where commercial entities play a major role in data collection and dissemination.

Certainly, copyright protection, licensing, and an active commercial database market can coexist with a strong public domain in digital data. In recent years, efforts have been undertaken to utilize licensing to actively foster an expanded public domain. Although, as noted above, data are not subject to copyright protection, uncertainties about what data users are legally allowed to do with them can inhibit sharing and reuse. For example, it may not be clear whether a particular data collection is copyrightable or whether the creator intends to assert copyright.

The fact that copyright persists for many years—whether it is asserted or not—means that a database may need to be actively placed into the public domain in order for users to be certain that it is free from copyright restrictions and any type of reuse is permitted. Creative Commons and its offshoot, Science Commons, have developed a number of innovations in the area of licensing aimed at facilitating open dissemination, sharing, and use of a wide variety of information, including data. For example, Creative Commons recently launched its CC0 (“CCZero”) protocol that allows creators of copyrightable work, including database generators, to waive all rights they may have to a given work, to the extent possible in the applicable jurisdiction.³⁷

Patents

Patents give researchers, nonprofit organizations, companies, and other entities the right to profit from an innovation. In return, the property owner must make the innovation public, which enables others to build on it. Once intellectual property is patented, it can be freely disseminated while still maintaining its commercial value to a company or research institution.

The Bayh-Dole Act of 1980 has had a major influence on the development of products from publicly funded research. The act granted the rights to inventions with the university, small-business, or nonprofit institution that accepted the research grant supporting the work. To accept this ownership, the university, small business, or nonprofit institution must:

- Report each disclosed invention to the funding agency;
- Elect to retain title in writing within a statutorily prescribed time frame;
- File for patent protection;
- Grant the federal government a nonexclusive, nontransferable, irrevocable, paid-up license to the invention;

³⁶ National Research Council, *Licensing Geographic Data and Services*.

³⁷ <http://wiki.creativecommons.org/CCZero>.

- Actively promote and attempt to commercialize the invention;
- Not assign the rights to the technology, with a few exceptions;
- Share royalties with the inventor;
- Use any remaining income for education and research;
- Give preference to U.S. industry and small business.

For research that is supported exclusively by nonfederal money, the title to any inventions resulting from those data is owned according to the conditions established by the funder. For instance, corporate employees must assign their intellectual property rights to their employer, even sometimes for work done outside the scope of their employment. When research in an academic institution is supported by corporate money, the conditions of ownership must be clearly specified. The conditions often include proprietary control over the outputs of that research. In the case of academic research that is supported by nonprofit organizations, control is established by the granting organization. One example is the Howard Hughes Medical Institute, which retains title to all inventions arising from its support but frequently assigns its rights to the associated university or nonprofit institution.

Trade secrecy may be used as an alternative to patenting. In some cases, inventions and underlying data have been held as proprietary trade secrets by companies and even universities and thus are treated as protected information as long as reasonable efforts are made to maintain secrecy. Researchers and their employers have this option, particularly if they do not plan to seek credit for the findings by reporting or publishing the results. Also, in cases where research at a university is supported by a private company, a research contract may provide for a short delay in publication or sharing data until the patentability of the research findings can be evaluated and, if appropriate, patent applications are filed.

As noted above, academic researchers may have incentives to transfer their research findings to the private sector. These include the desire to see their discoveries translated into useful products or to profit themselves from commercial opportunities made possible by research. If these incentives cause researchers to withhold data, the net effect can be for research data to become less available.

In 2006, a National Research Council Committee examined whether changes in patenting and licensing practices by companies and research institutions pose a threat to continued progress in the rapidly advancing areas of genomics and proteomics research.³⁸ The committee found that although difficulties in accessing proprietary research *materials* are clearly burdening research

³⁸ National Research Council. 2006. *Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation, and Public Health*. Washington, DC: The National Academies Press.

efforts, limited access to data is currently not a serious problem. The committee recommended that the National Institutes of Health (NIH) continue efforts to encourage the free exchange of data and materials through mechanisms such as requiring grantees to develop and adhere to data-sharing plans. The committee also called for efforts on the part of the U.S. Patent and Trademark Office to improve understanding of rapidly emerging technologies in order to avoid the extension of patent protection to inventions that do not meet the patentability standards of novelty, utility, and nonobviousness.

Journals and Access to Data

The interest of scientific, technical, and medical (STM) journals in the integrity of research data, and their role in ensuring it, was discussed in Chapter 2. Because journal articles are the primary means of communicating the results of research, and rely on data to support their findings, journals also play an important role in facilitating access to data. Although research data are not copyrightable, papers incorporating those data are. The conventional arrangement in traditional STM publishing has been for authors to transfer their copyright in the article they have written to the publisher, generally with some retention of rights to use the article.³⁹

The environment for STM journal publishing has changed considerably in recent years, as it has for nearly all publishing and media businesses.⁴⁰ Traditional subscription-access STM journals are published by both commercial and nonprofit entities. Commercial STM publishing has seen significant consolidation, with fewer companies publishing larger numbers of journals. Subscription prices for traditional STM journals have seen steep increases, putting severe pressure on research library budgets.⁴¹ Concurrently, open access STM journals have emerged as a significant part of the scholarly publishing world.⁴² One prominent example of an open access publisher is Public Library of Science (PLOS), which publishes several high-impact journals in the life sciences.⁴³

³⁹ Some universities assert copyright in selected categories of work by faculty, but often grant rights back to faculty for the purpose of traditional academic scholarship. See National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. 2004. *Electronic Scientific, Technical, and Medical Journal Publishing and Its Implications*. Washington, DC: The National Academies Press.

⁴⁰ Ibid.

⁴¹ Judith M. Panitch and Sarah Michalak. 2005. *The Serials Crisis: A White Paper for the UNC-Chapel Hill Scholarly Communications Convocation*. January. Available at <http://www.unc.edu/scholcomdig/whitepapers/panitch-michalak.html>.

⁴² "Open access" refers to publications, data collections, and other digital resources that are available to anyone without charge, and to the scholarly movement that advocates for policies and practices supporting such digital resources. The advocacy movement is referred to in the report as "Open Access," and the publications, data collections, and other digital resources as "open access."

⁴³ See the PLOS homepage at www.plos.org.

Another relevant trend is the growth in open access mandates for published research that have been initiated by research sponsors and research institutions. The most significant of these was adopted in early 2008 by NIH, having been mandated by Congress in the Consolidated Appropriations Act of 2008 and made permanent in the Omnibus Appropriations Act of 2009.⁴⁴ The NIH policy provides that:

The Director of the National Institutes of Health (“NIH”) shall require in the current fiscal year and thereafter that all investigators funded by the NIH submit or have submitted for them to the National Library of Medicine’s PubMed Central an electronic version of their final, peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication: Provided, that the NIH shall implement the public access policy in a manner consistent with copyright law.”⁴⁵

Research institutions are also adopting open access recommendations for faculty research, encouraging faculty to provide electronic copies of their articles for submission to an institutional or other open access repository, generally with an embargo period of 6 to 12 months. This is an international trend, with research institutions or sponsors (both public and private) adopting open access publication recommendations in Europe, Canada, Australia, and India.⁴⁶

The issues raised by the changing environment for scholarly publishing are the subject of continued, vigorous debate. Although they are not within the task statement of this study, it is necessary to review them in this context because access to scholarly publications is related to access to research data at several levels. For example, institutional and governmental repositories that support access to, and stewardship of, faculty articles may serve the same function for data (the data stewardship function of repositories is discussed in Chapter 4). It is also important to note the distinctions between open access to data and open access to publications. Traditional access STM publishers that might look unfavorably on open access publication mandates might support practices and guidelines encouraging open access to data.⁴⁷ Open access mandates for data, to be discussed in the next section, are distinct from open access mandates for publications.

⁴⁴ National Institutes of Health. 2009. The Omnibus Appropriations Act of 2009 Makes the NIH Public Access Policy Permanent: NOT-OD-09-071. March 19. Available at <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-09-071.html>.

⁴⁵ *Ibid.*

⁴⁶ A continuously updated list of open access publication mandates is available at <http://www.eprints.org/openaccess/policysignup/>.

⁴⁷ See International Association of Scientific, Technical & Medical Publishers. 2009. Briefing Document (for Publishing Executives) on Institutional Repositories and Mandated Deposit Policies. January; International Association of Scientific, Technical & Medical Publishers. 2006. Databases, Datasets, and Data Accessibility—Views and Practices of Scholarly Publishers. June. Available at <http://www.stm-assoc.org/documents-statements-public-co/>.

LEGAL AND POLICY REQUIREMENTS FOR ACCESS TO DATA

The Data Access Act and the Information Quality Act

Various government laws, regulations, and policies influence the accessibility of research data. Among these are the Data Access Act (DAA) of 1999 and the Information Quality Act (IQA) of 2001, also known as the Data Quality Act.⁴⁸

The DAA is also known as the “Shelby Amendment,” after its sponsor, Senator Richard Shelby of Alabama. It was passed as a rider to an appropriations bill in 1999. The DAA requires that data from federally funded research be made available to requesting parties under Freedom of Information Act procedures if the research is: (1) used to support an agency action, and (2) performed by a university or other nonprofit institution.⁴⁹ In response, OMB modified its Circular A-110 to read as follows:

[I]n response to a FOIA [Freedom of Information Act] request for research data relating to published research findings under an award that were used by the federal government in developing an agency action that has the force and effect of law, the federal awarding agency shall request, and the recipient will provide within a reasonable amount of time, the research data so that they can be made available to the public under FOIA.

The provision established which types of research data are subject to disclosure and the procedures, standards, and exemptions that apply in requesting and disclosing those data. Before the provision was published, persons could only obtain raw data that were in possession of a federal agency, whereas the revised provision provided access to data that are in possession of a grantee institution. If even a small amount of public money was used to produce data, those data may be subject to DAA requests. However, studies conducted by industry or by others without the use of public funds are not covered by the data-sharing requirements, even if the studies are employed in the formulation of public policy or regulations. Also, as interpreted by OMB, the provision applies only to data supporting regulations with a “major” impact on the economy and is prospective, covering studies launched after the OMB guidelines were put into effect.

The DAA was controversial at the time the legislation passed and when OMB was developing the specific changes to Circular A-110. Participants in a 2001 National Research Council workshop pointed out future problems that

⁴⁸ National Research Council. 2002. *Access to Research Data in the 21st Century: An Ongoing Dialogue Among Interested Parties: Report of a Workshop*. Washington, DC: The National Academies Press.

⁴⁹ Wendy Wagner and David Michaels. 2004. “Equal treatment for regulatory science: Extending the controls governing the quality of public research to private research.” *American Journal of Law & Medicine* 30:119–154.

might be encountered in implementing the DAA, suggesting that this approach might not be an ideal way to ensure public access to data underlying federal policies and regulations.⁵⁰ At the same time, the DAA does not appear to have led to any contentious cases during the decade since it went into effect. For example, a 2003 General Accounting Office report found that two agencies had received a total of 42 requests under the DAA up to that time, and that none of the requests had actually met the Circular A-110 criteria.⁵¹

The IQA was passed as a two-sentence rider to the 2001 Consolidated Appropriations Act. The IQA called on OMB to issue regulations for “ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by Federal agencies.” In response, OMB issued guidelines that all agencies “must embrace a basic standard of quality as a performance goal, and agencies must incorporate quality into their information dissemination practices.”⁵²

The guidelines state that “if an agency is responsible for disseminating influential scientific, financial, or statistical information, agency guidelines shall include a high degree of transparency about data and methods to facilitate the reproducibility of such information by qualified third parties.”⁵³ For “original and supporting data,” agencies are to consult with “relevant scientific and technical communities” and determine which data are subject to the reproducibility requirement.⁵⁴ “Reproducibility” here means a high level of transparency about research design and methods, which is meant to negate any need to replicate work before dissemination. For “analytic results” there must be “sufficient transparency about data and methods that an independent reanalysis could be undertaken.”⁵⁵ This means that “independent analysis of the original or supporting data using identical methods would generate similar analytic results, subject to an acceptable degree of imprecision or error.”⁵⁶ In cases where the public does not have access to data and methods (privacy, security, trade

⁵⁰ See National Research Council, *Access to Research Data in the 21st Century*. In particular, Chapter 6, which reports on workshop chair Richard Merrill’s summary remarks, is a concise statement of the longer-term shortcomings of DAA.

⁵¹ General Accounting Office. 2003. University Research: Most Federal Agencies Need to Better Protect against Financial Conflicts of Interest. GAO-04-31. November. Washington, DC: General Accounting Office.

⁵² Office of Management and Budget. 2002. “Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies; Notice; Republication.” *Federal Register* 67(36):8451–8460. Available at <http://www.noaa.gov/stories/feb22.pdf>. This *Federal Register* entry includes the final guidelines as well as a discussion of the comments received.

⁵³ *Ibid.*, p. 8455.

⁵⁴ *Ibid.*

⁵⁵ *Ibid.*, p. 8456.

⁵⁶ *Ibid.*

secrets), “agencies shall apply especially rigorous robustness checks to analytic results and document what checks were undertaken.”⁵⁷

A committee organized by the National Research Council’s Committee on Science, Technology, and Law held several workshops in 2002 to discuss OMB’s IQA guidance and the agency responses that were being developed. The summary of those workshops reviews a number of the issues agencies faced in developing their own implementing guidelines.⁵⁸

Federal and Journal Policies Affecting the Availability of Data

Table 2-2 shows federal agency policies toward availability of data generated directly by agencies as well as data generated by external grantees. In 2008 the federal government released its Principles for the Release of Scientific Research Results in response to the America COMPETES Act of 2007.⁵⁹ These principles promote sharing of data from research undertaken by federal civilian agency employees.

For federally sponsored research performed by external organizations, the grants guides of agencies vary in how strongly data sharing is encouraged or required. A 2007 Government Accountability Office (GAO) assessment of agency policies toward grantees in climate science found that although agencies encouraged data sharing, the specific requirements varied from program to program.⁶⁰ For example, the National Science Foundation (NSF) grants guide states the expectation that grantees make their data “widely available and useful” within a “reasonable time.” Specific NSF programs might require that data be deposited in a specific repository within a set time period following data collection. The GAO report also found that agencies generally do not monitor whether data-sharing requirements are being met and have not overcome barriers to sharing, such as the lack of appropriate data archives in some subfields of climate science.

Although specific federally sponsored research programs include a range of data-sharing mandates, no federal agency has yet adopted an agencywide open access data mandate, analogous to NIH’s open access publication mandate. NIH does require that grant proposals above a certain size include a data management plan consistent with NIH’s Data Sharing Policy, which is discussed further below in the section on “Responsibilities of Research Institu-

⁵⁷ Ibid., p. 8457.

⁵⁸ National Research Council. 2003. *Ensuring the Quality of Data Disseminated by the Federal Government: Workshop Report*. Washington, DC: The National Academies Press.

⁵⁹ John H. Marburger, III. 2008. Principles for the Release of Scientific Research Results. Memorandum. May 28. Available at www.arl.org/bm~doc/ostp-scientific-research-28may08.pdf.

⁶⁰ Government Accountability Office. 2007. Climate Change Research: Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research. September. Available at <http://www.gao.gov/new.items/d071172.pdf>.

tions, Research Sponsors, Professional Societies, and Journals.” Federal agencies are creating new open access data resources, such as the Department of Energy’s Data Explorer program, an open access repository of data from DOE-sponsored research, and National Library of Medicine efforts such as GenBank, which is discussed elsewhere in the report.⁶¹

Some private research sponsors such as the Wellcome Trust have adopted open access data mandates for their grantees.⁶² As shown in Table 2-1, an increasing number of STM journals have adopted open access data mandates for authors.⁶³

THE INTERNATIONAL DIMENSIONS OF ACCESS TO RESEARCH DATA

The advent of digital networks has enabled and stimulated global access to all types of digital information, including research data. Access to research data online means that researchers can use the data on a global basis, enhancing the universal progress of science to solve common problems and develop new knowledge. Both the benefits and the costs of unrestricted and restricted access are thus amplified in the international context.

The United States has been a leader in promoting openness to public sector information, as well as to publicly funded research data. Despite the trends in fields with commercial potential toward more proprietary treatment of academic research and the post-September 11 increase in national security restrictions on some sensitive data sources and types, the overall policy trend may be seen as moving toward greater access to both governmental and academic research data sources. The international dimensions of access to research data are being shaped both from the bottom up and the top down.

At the informal working level of the individual investigator, data are now shared across geographic boundaries as easily as they once were with the colleague next door. Countless international data exchanges are made among scientists on a daily basis, or through the posting of datasets on individual researchers’ Web sites.

At a more formal level, international research projects establish data-sharing protocols that reflect the norms of the fields in which they are operating. Some of the larger research or infrastructure programs are establishing data centers or federated networks for sharing of data resources. The first international network of such data centers, the World Data Center system, was formed following the 1957 International Geophysical Year to help bridge the gap in cooperation and data exchanges during the cold war.

⁶¹ See <http://www.osti.gov/dataexplorer/>.

⁶² See <http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm>.

⁶³ See, for example, http://www.nature.com/authors/editorial_policies/availability.html.

With the advent of global digital networks over the past two decades, both international cooperation in research and the formation of networked data resources on regional and global levels have become commonplace. Examples include the Global Biodiversity Information Facility, the International Federation of Digital Seismograph Networks, the International Nucleotide Sequence Database Collaboration, the International Virtual Observatory Alliance, and the Global Earth Observation System of Systems, to name but a few. Almost all fields of inquiry have some data centers or networks designed to provide access to data. In most cases, the U.S. research community has been the organizing force for the collaborative data-sharing networks.

Greater access to research data from public funding also is receiving more attention at the national policy levels of many countries, in part because such data resources are now seen as being major research infrastructure components. For example, the Research Councils of the United Kingdom adopted a more open policy for their data holdings in 2006. The Ministry of Science and Technology in China initiated the Scientific Data Sharing Project in 2002, in recognition of the fact that “[t]he insufficient use of China’s massive data holdings has been an urgent problem.”⁶⁴ Many other countries are similarly reviewing or revising their national policies and myriad institutional ones to make better use of their data resources.

Finally, some international scientific, engineering, and medical organizations at both the intergovernmental and nongovernmental levels, such as the International Council of Scientific Unions, the Committee on Data for Science and Technology, and the OECD, are developing data-sharing policies and guidelines for adoption by members and the international research community. For example, the OECD in 2007 published its Principles and Guidelines for Access to Research Data from Public Funding, which are summarized in Box 3-2. The InterAcademy Panel, an organization of national science academies, supports a program to expand access to digital scientific information to researchers in developing countries.⁶⁵

GENERAL PRINCIPLE FOR ENHANCING ACCESS TO RESEARCH DATA

Because of the huge increase in the quantity of research data being generated, it is possible to say both that more data are being publicly disseminated than have ever been before and that more data are being withheld from public

⁶⁴ Jinpei Cheng, 2006. The development of China’s scientific data sharing policy. In National Research Council. *Strategies for Preservation of and Open Access to Scientific Data in China: Summary of a Workshop*. Washington, DC: The National Academies Press. Available at: http://www.nap.edu/catalog.php?record_id=11710.

⁶⁵ See the program’s Web site: <http://www.interacademies.net/CMS/Programmes/4704.aspx>.

access today than have ever been before. Many fields of research have moved toward more open data-sharing policies as the value of data has increased and as digital technologies have enabled information to be disseminated more broadly. At the same time, heightened interest in the commercial applications of research data has caused some forms of data to be more restricted.

As described earlier in this chapter, there are legitimate reasons why some research data are not made publicly available, ranging from privacy concerns to technical barriers. Yet the basic principle that should guide decisions involving research data supporting publicly reported research results is clear:

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

This principle applies throughout research, but in some cases the open dissemination of research data may not be possible or advisable when viewed from the perspective of enhancing research in science, engineering, or medicine. Access to research data prior to reporting results based on those data might undermine the incentives to pursue the research. 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 the section on “Legal and Policy Requirements for Access to Data.” Also, “accessible” does not necessarily imply that data should be disseminated for free, though free or marginally priced distribution is the ideal. Nor are researchers responsible for providing data users with instruction or training in the use of their data, though they do have a responsibility to provide metadata, analysis software, models (including code and input data) and other information necessary for practitioners to validate and build on the results. Where researchers have proprietary interests in such tools, they have the option of protecting those interests through applying for patents and/or asserting copyright, as appropriate, in advance of publicly reporting results.

This principle is a standard that is not currently being met in some areas of research. Yet it provides a yardstick against which to measure current initiatives and future plans. Researchers know that the information they generate should be available to others to advance the frontiers of knowledge. The objective therefore must be to implement policies and promote practices that allow this principle to be realized as fully as possible.

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. In many cases, making such data available can produce societal benefits while not threatening the commercial opportunities that led to the data’s generation. Note that this principle covers data underlying publicly reported results. When a researcher working at a corporate lab seeks to publish results, patent applications can be filed in advance

of publication, so that making data accessible at the time of publication will not compromise commercialization of the invention in question. If a company decides to protect an invention as a trade secret, it might be assumed that researchers will not publish papers about the invention and the question of providing access to data will not arise.

In the past few years we have also seen private companies announce plans to make significant data resources available on an open access basis. For example, Merck has spun off a nonprofit, open access platform known as Sage.⁶⁶ Sage is aimed at helping researchers to build new databases aimed at more effectively modeling disease. Where possible, public policies should encourage the release of such data, and privately funded researchers and their managers should explore possible means of making data available.

The Access and Sharing Principle is consistent with recommendations from National Academies committees that have previously addressed data access. A 2003 report, *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*, puts forward the “uniform principle for sharing integral data and materials expeditiously (UPSIDE).”⁶⁷ The UPSIDE principle calls on researchers employed in the academic, government, and commercial sectors to provide data and materials needed to support published findings, and to “provide them in a form on which other scientists can build with further research.” The 1997 report *Bits of Power: Issues in Global Access to Scientific Data* states that “full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from publicly funded research.”⁶⁸

RESPONSIBILITIES OF RESEARCHERS

As with the integrity of research data, the primary responsibility for sharing data lies with the researchers who produced them. (In addition, other parts of the research enterprise have responsibilities for sharing data, as described later in this chapter and in the next chapter.) Only researchers know their data well enough to ascertain what information must be publicly available to allow others to verify their results and build on their work. Only researchers are in a position to work with research institutions, research sponsors, and journals to make data available in a way that they can be understood and used effectively by others. Thus, our committee recommends that:

⁶⁶ Bryn Nelson. 2009. “Something wiki this way comes.” *Nature* 458(13, March 4). doi:10.1038/458013a.

⁶⁷ National Research Council. 2003. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Washington, D.C.: The National Academies Press.

⁶⁸ National Research Council. 1997. *Bits of Power: Issues in Global Access to Scientific Data*. Washington, DC: National Academy Press.

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.

Making data available does not necessarily mean providing them at no cost. The next chapter discusses the need for research projects to develop plans for the management and sharing of data from the initial stages of a research program. Chapter 4 also describes the evolving infrastructure for providing data access and stewardship, whose components include institutional and disciplinary repositories.

Fulfilling this recommendation also requires that researchers be familiar with any possible constraints on the release of data. Although this information is usually known to researchers and their managers from the outset of a research project, agreements may be informal, may be understood differently by different parties (such as principal investigators and graduate students), or may change during the course of a research project. Requiring that researchers clarify and agree to these arrangements places the responsibility on researchers to oversee the accessibility of research data and to decide whether to participate in research where data accessibility is limited. Researchers who are considering becoming involved in a project where data accessibility is restricted need to ask themselves whether the benefits of participating in that project outweigh the benefits of transparency in generating and disseminating data.

Research thrives under conditions where data are available to others. If data are not available, there should be a clear and public reason why those data are being withheld from dissemination. Indeed, justifications for not making data available should be understood by the researcher, sponsor, and institution. Dissemination of the reasons why data are being withheld could be published with journal articles, posted on Web sites, stated in the publicly accessible award statements of research sponsors or research institutions, or made available by some other means. The important point is that the reasons should be publicly available so that others can review and comment on the grounds for withholding data.

As discussed in the following section, the committee believes that research fields, research sponsors, research institutions, and journals have considerable ability to set appropriate standards and expectations regarding data access and sharing, and to develop the necessary incentives. Some are taking leadership roles in setting standards and instituting incentives. The committee believes that continued efforts taken by these stakeholders can create an environment in which the Data Access and Sharing Principle is widely followed in the research enterprise, and in which a bureaucratic framework of regulations and enforcement will not need to be imposed.

RESPONSIBILITIES OF RESEARCH FIELDS

As emphasized earlier, there are major differences between research fields in the handling of data, including technological infrastructure, publication practices, and data-sharing expectations. In some fields, aspects of their data culture act as barriers to access and sharing of data. Because of the growing importance of research data and the rate at which practices are changing in research, it is important for various fields and disciplines to examine their standards and practices regarding data and to make these explicit.

The development of plans for data management and sharing is greatly facilitated when a field of research has standards and institutions in place designed to promote the accessibility of data.

Recommendation 6: In research fields that currently lack standards for the sharing of 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.

The development of standards and institutions can occur in different ways depending partly on the field of research in which it occurs. The process can be led by journal editors, professional societies, ad hoc bodies of researchers established to solve particular problems, or permanent institutions charged with overseeing data management issues. National Academies committees and advisory committees to federal agencies can play constructive roles. In large, complex fields, multiple initiatives may be undertaken to address various aspects of standard setting. Input and participation from international stakeholders will often be needed.

The life sciences provide useful examples of the standards-setting process. As described in Box 3-4, a National Academies committee developed broad standards for the sharing of research data in the life sciences. Similarly, as described in Box 3-5, a journal-led effort incorporating community input developed the Paris Guidelines for the management of protein data. Both examples demonstrate how standards can be put in place to deal with existing or new issues affecting the management of research data.

The Principles for the Release of Scientific Research Results, released in 2008 and discussed in the earlier section on “Federal and Journal Policies Affecting the Availability of Data,” establish data-sharing standards for research conducted by employees of federal civilian agencies.⁶⁹ One section of the principles states:

⁶⁹ John H. Marburger, III. 2008. Principles for the Release of Scientific Research Results. Memorandum. May 28. Available at www.arl.org/bm~doc/ostp-scientific-research-28may08.pdf.

BOX 3-5 The Paris Guidelines

In some fields, journals have played a major role in developing standards for data collection, sharing, and preservation. In 2004, for example, the journal *Molecular and Cellular Proteomics* (MCP) developed standards for the management of protein data.^a These standards were revised 1 year later based on community input, resulting in the “Paris Guidelines.”^b These guidelines were made available in a checklist format, in a tutorial, and in MCP-hosted workshops to educate researchers about the details of the requirements for publication and data submission.^c

MCP’s standard requires all relevant quantitative data to be made available at a level in which it is possible to reproduce the reported results. Methods can reference previously published standards but any deviations must be explained. In particular, authors must submit along with the manuscript the data that have the greatest potential for misinterpretation—for instance, mass spectrographic spectra for post-translationally modified proteins—for the journal to publish.

Data considered less important but worthy of access are recommended for submission to the journal as supplementary material to be deposited in a nonjournal repository, which therefore may not be archival.^d In addition, an institutionally based government-funded data depository was recommended (“Tranche”) that has a distributed storage system similar to Bit Torrent, thereby lessening costly bandwidth problems caused by downloading large amounts of data over the Internet.

In this way the Paris guidelines ensure that the most important data are deposited for perpetual and accessible storage while second-tier data also are accessible without placing too large a burden on the journal as the sole repository for data.

^a Steven Carr, Ruedi Aebersold, Michael Baldwin, Al Burlingame, Karl Clauser, and Alexey Nesvizhskii. 2004. “The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data.” *Molecular and Cellular Proteomics* 3:531–533.

^b Ralph A. Bradshaw, Alma L. Burlingame, Steven Carr, and Ruedi Aebersold. 2006. “Reporting protein identification data: The next generation of guidelines.” *Molecular and Cellular Proteomics* 5:787–788.

^c See http://www.mcponline.org/misc/Tutorial_MCP_final.pdf.

^d For an example of supplementary data, see <http://www.mcponline.org/cgi/content/abstract/6/7/1123>.

Research data produced by scientists working within Federal agencies should, to the maximum extent possible and consistent with existing Federal law, regulations, and Presidential directives and orders, be made publicly available consistent with established practices in the relevant fields of research.

This principle is consistent with the Data Sharing and Access Principle stated above. This report advocates that the principle apply not just to federal scientists but to all research where results are publicly reported.

A wide range of issues must be considered in setting data standards, including dissemination, usage restrictions, periods of exclusive use, documentation requirements, financial provisions, ownership, licensing terms, infrastructure needs, technological compatibility, and sustainable preservation. These issues vary greatly from field to field, depending on particular traditions and requirements. Although it is not impossible to prescribe a standard set of practices to which all researchers should adhere—indeed, the general principles stated in this report apply to all researchers—every field collectively and every researcher individually must address issues of data accessibility.

RESPONSIBILITIES OF RESEARCH INSTITUTIONS, RESEARCH SPONSORS, PROFESSIONAL SOCIETIES, AND JOURNALS

For researchers to make their 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

As noted earlier in this chapter, research institutions, research sponsors, professional societies, and journals are undertaking a range of initiatives to promote the sharing of research data. In taking the next steps, research institutions and research sponsors need to create incentives for researchers to share data, just as they have incentives to maintain the integrity of research data and to publish their findings. Researchers need both formal and informal ways of being acknowledged and rewarded for making research data accessible and usable. For example, in some cases tenure and promotion decisions could take into account efforts to promote the accessibility of data, the creation of publication-based metrics, or service to a community or institution.

Data professionals also have an important role to play in ensuring the accessibility of research data. In close cooperation with researchers in a field, data professionals can anticipate the needs of data users and establish data management systems that meet those needs. Their contributions to making data accessible, as well as ensuring the integrity of data, need to be recognized.

One way for research sponsors and journals to promote data accessibility is to establish the terms of access and sharing expected of institutions and investigators. For example, NIH explicitly requires that all grant applications for more than \$500,000 in direct costs in a single year must include a data management plan that embodies the principles of the NIH Data Sharing Policy. This policy says that “data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and

proprietary data.” The data management plan becomes part of the proposal, and “NIH expects that plan to be enacted. . . . In the case of noncompliance (depending on its severity and duration) NIH can take various actions to protect the Federal Government’s interests.”⁷⁰ These actions are not specified but may affect the review of future proposals.

As discussed above, research institutions, research sponsors, and journals have considerable leverage in encouraging data access and sharing on the part of researchers. Several leading research institutions have announced open access publication recommendations, which encourage faculty to deposit their publications in their institutional repository. Such recommendations could be extended to data. Some federal research programs and journals have adopted open access data policies that require or encourage researchers to deposit underlying data in a disciplinary or institutional repository (see Tables 2-1 and 2-2). Depending on the program or discipline, adopting and effectively enforcing such open access data policies may be an appropriate way for research institutions, research sponsors, and journals to implement this recommendation.

The Council on Government Relations points out that “few institutions have formal policies and procedures for access to and retention of research data.”⁷¹ As described above, the terms of research contracts and grants and other regulations often specify that research institutions are responsible for retaining data and providing access. Given the current lack of formal policies and procedures, we make the following recommendation.

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.

The knowledge needed to develop data access policies is not widespread or fully developed. Research institutions and sponsors may need to come together to identify best practices and policy models. Organizations such as the Association of American Universities, the Association of Public and Land-Grant Universities, the Association of Research Libraries, and the Council on Government Relations can contribute to this process.

Disputes between researchers and their institutions regarding control of data are not unusual. For example, faculty members may be denied tenure and seek to take their research data with them, while the institution may seek

⁷⁰ National Institutes of Health Office of Extramural Research. 2003. NIH Data Sharing Policy and Implementation Guidance.

⁷¹ Council on Government Relations. 2006. Access to and Retention of Research Data: Rights and Responsibilities. March. Washington, DC: Council on Government Relations.

to keep it. Or researchers and institutions may have different perspectives on how to respond to outside requests for access to data, including requests made under the auspices of the DAA or in connection with litigation. As described earlier in this chapter, requests for information can go beyond research data to information about a researcher's personal life.

Procedures for handling requests for data that either intentionally or inadvertently hamper the progress of research need special attention. Although the data from publicly funded research should be accessible in general, exploiting the norms of science to slow or stop the progress of research harms society. For example, institutional policies might stipulate that an institution will come to the aid of researchers in disputes with third parties, but researchers also must comply with institutional policies.

Many journals play a critical role in ensuring access to the data that support the publications appearing in those journals (see Box 3-6 for an example). Access to those data may be lost as journals evolve under the pressures of dramatic changes being catalyzed by digital technologies. The following chapter covers the responsibilities of journals to make data accessible in the context of the long-term preservation of research data.

BOX 3-6 Promoting Reproducibility in Medical Research

As of April 1, 2007, the *Annals of Internal Medicine* instituted a new policy designed to help the research community evaluate and build on published results. Authors of original research articles in the *Annals* are required to include a statement indicating whether the study protocol, data, and statistical code are available to readers and under what terms the authors will share this information. Sharing is not mandatory, but authors are required to state whether they are willing to share the protocol, data, and statistical code. Authors are not asked whether they are willing to make this information available until after a manuscript is accepted for publication.

According to an article announcing the new policy, the goal of the new requirement is to promote “reproducible research” in which independent researchers can reproduce results using the same procedures and data as the original investigators. Reproducible research does not require unlimited access to data and methods, but it requires access to as much of the dataset and statistical procedures as is necessary to reproduce the published results. As the article states:

Major cultural shifts in research must occur before a world of completely reproducible research can exist. These shifts include increasing the technical capacity of many research teams, further developing acceptable data-sharing mechanisms, and supporting—both professionally and financially—the publishing of reproducible research. . . . We hope that shining a spotlight on the availability of the study protocol, data, and statistical code for every *Annals* research report will be seen as a small but important step toward biomedical research that the public can really trust. At the same time, it will enhance what is perhaps the main function of a journal: to provide a transparent medium for a conversation about science.^a

^aFor more information, see Christine Laine, Steven N. Goodman, Michael E. Griswold, and Harold C. Sox. 2007. “Reproducible research: Moving toward research the public can really trust.” *Annals of Internal Medicine* 146:450–453.

4

Promoting the Stewardship of Research Data

Realizing the full value of research data requires that the data be accessible to the community of researchers and others who might be able to use them. The data need to be accompanied by sufficient metadata for them to be found easily, understood in context, and used appropriately. Data need to be stored in repositories using up-to-date technologies until a decision is made that the information is no longer needed. Data useful for ongoing research or historical purposes may need to be stored indefinitely. These issues of useful accessibility, annotation, curation, and preservation are the heart of what we term in this report the *stewardship* of research data.

Digital technologies are having a revolutionary effect on every aspect of data stewardship. The Internet provides a mechanism for making data available to anyone anywhere in the world. Powerful new computers and sophisticated software can automate part of the process of annotating data. Data repositories offer a means for preserving digital data for the indefinite future. Though the infrastructure necessary for data stewardship is still taking shape, much of the technological capability needed to realize the full value of research data already exists.

Secondary use of data is of growing importance in an increasing number of fields. In astronomy, for example, the Sloan Digital Sky Survey, a project for which the open provision of both processed and raw data over the Internet is central, is the facility responsible for the most high-impact papers in astronomy in recent years.¹ Repositories of genomic data, such as the Trace Archive of the National Center for Biotechnology Information (NCBI), have become essential components of the national and global infrastructure for life sciences research

¹ Juan P. Madrid and F. Duccio Macchetto. 2006. High-impact astronomical observatories. *Bulletin of the American Astronomical Society Electronic Edition* 38(4). Available at <http://www.aas.org/publications/baas/v38n4/BAASv38n4Madrid.pdf>.

(see Figure 4-1). In other areas, such as clinical data, the potential gains from data reuse are clear, even though technical and other barriers stand in the way of realizing that potential.²

This technological capability has given rise to a powerful new vision of how some areas of research can be conducted.³ Known as e-science or cyberinfrastructure, this approach to research involves decentralized collaborations of researchers who draw on remote sensors and facilities, very large data collections, and powerful computing resources. These distributed resources are interconnected so that they can be shared in a flexible, secure, and coordinated manner. Individuals and groups can build and make available services and tools that extend across research fields.⁴ In an interconnected grid of facilities, instruments, and computers, the collective knowledge of scientific, engineering, and medical research resides not just in published books and articles but in the grid itself.

THE LOSS AND UNDERUTILIZATION OF RESEARCH DATA

E-science has been partially implemented in a number of research fields, but in others information technology is not being used to advantage.

Today, much research data that could be of value in the future are lost because of the lack of provisions for preserving them: Research notebooks are discarded; computer hard disks crash, destroying unique data; an investigator changes fields, retires, or dies and leaves behind data that are poorly organized, haphazardly stored, or otherwise unusable.

Digital data are often stored in formats that rapidly become technologically obsolete. Data stored on paper can survive for decades or centuries before the paper breaks down and becomes unreadable. In the digital age, however, the longevity of storage media sometimes seems to conform to an inverse Moore's law, with accelerating technological advances hastening the demise of superseded media. Many scientists have data on floppy disks, hard drives, or zip drives that new generations of computers cannot read. One expert raises the possibility of a "digital dark age," in which large amounts of digital data stored in a variety of proprietary file formats are permanently lost.⁵

Digital media also decay over time, a phenomenon known as "bit rot." Many old magnetic tapes molder in boxes and are now essentially worthless.

² James J. Cimino. 2007. "Collect once, use many: Enabling the reuse of clinical data through controlled terminologies." *Journal of AHIMA* 78(2):24–29.

³ National Science Foundation Cyberinfrastructure Council. 2007. *Cyberinfrastructure Vision for 21st Century Discovery*. Arlington, VA: National Science Foundation. Available at <http://www.nsf.gov/pubs/2007/nsf0728/index.jsp>.

⁴ Ian Foster. 2005. "Service-oriented science." *Science* 308:814–817.

⁵ Phil Ciciora. 2008. "Digital dark age' may doom some data." University of Illinois at Urbana-Champaign News Bureau. October 27. Available at news.illinois.edu/news/08/1027data.html.

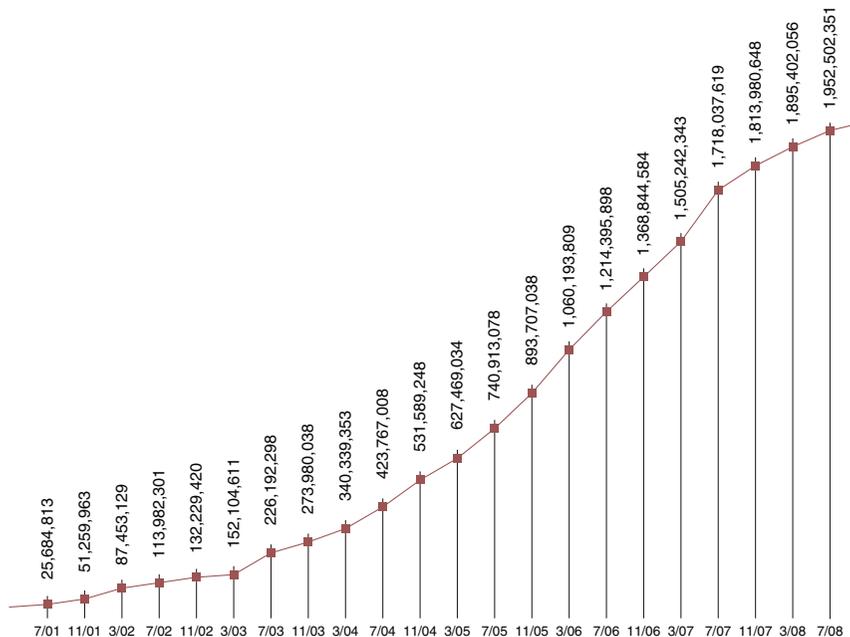


FIGURE 4-1 National Center for Biotechnology Information Trace Archive through September 2008

SOURCE: National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=graph_query&m=stat&s=graph).

Data stored on CD disk drives begin to degrade within a few years. Unless provisions are made to move data from one storage medium to another, the data are lost relatively quickly. Of course, if data are judged to be valuable to a research community, resources can be devoted to replication so as to minimize the risk of digital media decay. As generations of applications, data formats, operating systems, and digital archives interoperate and succeed one another, multiple locations and systems for data access and sharing might be engaged to preserve a given data collection. Ensuring that archived data are not altered due to human error or intentional mischief is an additional challenge for large data repositories, particularly those utilizing automated processes to ingest large datasets.⁶ Table 4-1 shows the various risks to long-term digital data reliability and the time frames in which they might be expected to occur.

⁶ National Research Council. 2005. *Building an Electronic Records Archive at the National Archives and Records Administration*. Washington, DC: The National Academies Press. See Chapter 4 in particular.

TABLE 4-1 Long-term Data Reliability Issues

Entity at Risk	What Can Go Wrong?	Frequency
File	Corrupted media, disk failure	1 year
Disk	Simultaneous failure of two copies	5 years
System	Systematic errors in vendor software; malicious user; operator error that deletes multiple copies	15 years
Archive	Natural disaster, obsolescence of standards	50-100 years

SOURCE: Francine Berman, SDSC, presentation to the committee, September 2007.

The loss of valuable data is especially a problem in small research projects. Large projects often have data management plans and funds set aside for data storage and dissemination. Individual investigators, however, typically face much greater challenges in deciding which data may be useful in the future, in documenting those data thoroughly, and in finding funds from limited budgets for adequate data curation and preservation. Furthermore, although large projects can generate immense quantities of data, small research projects can themselves produce substantial quantities and varied kinds of data.

Some research fields that formerly consisted almost exclusively of small projects, such as molecular biology or ecology, have moved in part toward larger and more data-intensive programs. Some of these fields have groups that oversee the collection and annotation of data for use by others. The social sciences, for example, have long sponsored a specialized institution that has data stewardship as part of its mission (see Box 4-1). Other fields, despite generating much larger quantities of data, continue to be characterized by largely disparate and often inadequate data management efforts.

Not all research data should be preserved, but deciding what to save and what to discard becomes increasingly difficult as ever larger quantities of data are generated. Furthermore, there is a financial trade-off between creating new data and preserving old data. While the cost of storage per bit is declining rapidly, as described in Chapter 1, data stewardship requires a long-term commitment of attention and resources. As the secondary use of data becomes more important for fields and disciplines, they need to develop guidance for researchers, research sponsors, and research institutions on what data should be preserved, and whether new organizations or capabilities are needed to perform stewardship functions. A 2002 National Research Council report on geosciences data and collections is a useful example of how research fields can develop criteria for prioritizing the data and collections that should be preserved, and for making the trade-offs between creating new data and preserving existing data.⁷

⁷ National Research Council, 2002. *Geoscience Data and Collections: National Resources in Peril*. Washington, DC: The National Academies Press.

The discussion of neuroscience data issues in Box 1-3 illustrates the challenges facing data-intensive fields that need to develop policies, standards, and new organizational approaches to data stewardship.

Ownership considerations influence the stewardship of research data, just as they do access to the data. As discussed in Chapter 3, the institutions that receive research grants are generally acknowledged to be the owners of the data and other “intangible property” resulting from that research.⁸ However, for practical reasons, researchers may retain possession of the data on behalf of the institution, and institutions may specify in policies or contracts that investigators are to serve as the custodian of data and as the responsible party for preserving and retaining data.⁹ Indeed, investigators often assume that they are the owners of the research data that they produce, which can create problems when they move to a different institution and their original institution exerts its ownership rights over the data.

INFRASTRUCTURE AND INCENTIVES FOR THE STEWARDSHIP OF DATA

Each group associated with the generation, use, and preservation of research data has different incentives and expertise with respect to the stewardship of those data.

Researchers

Although the researchers who generate the data have the greatest stake in their use, they do not necessarily have a strong interest or incentive in preserving data, especially in small-scale projects. Most researchers prefer to pursue new goals rather than devote effort to making their existing and past data useful for others. Figure 4-2 shows the results of a survey by the Inter-University Consortium for Political and Social Research (ICPSR). Many National Science Foundation (NSF)- and NIH-sponsored projects that promised to create social science data have not followed through. Investigators typically have little expertise in data annotation or long-term database management.

This resistance to sharing on the part of faculty is changing over time, and this can be expected to accelerate as the value of publicly accessible data becomes more apparent in a wider range of disciplines, and as infrastructure for

⁸ Council on Government Relations. 2006. *Access to and Retention of Research Data: Rights and Responsibilities*. Washington, DC: Council on Government Relations.

⁹ For example, the National Institutes of Health (NIH) requires that primary research data be retained for at least 3 years after the closeout of a grant or contract agreement. See http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm.

BOX 4-1

Data Stewardship and Accessibility in the Social Sciences

The Inter-University Consortium for Political and Social Research (ICPSR) is an interdisciplinary institution established in 1962 to provide data stewardship and access for a wide range of datasets from the social sciences. Part of a global network of social science data archives, ICPSR is the world's largest archive of digital social science data and is hosted by the University of Michigan.^a It is supported by dues from more than 600 member institutions, plus support from government agencies and other research sponsors.

ICPSR, which currently houses 7,500 studies and 500,000 data files, has recommended guidelines, but not requirements, for submission of data. As part of its mission, ICPSR proactively seeks out data at risk of being lost. It also emphasizes the importance of preparing good documentation, or metadata, which are critical to data interpretation and to successful data sharing and preservation. These metadata include project summaries, descriptions of data collection instruments, summary statistics, database dictionaries, and bibliographies. As technology progresses, ICPSR migrates data to new storage media and maintains sets of redundant copies in various locations.

Ownership and access to data in the social sciences is determined by funding, with contract-funded data belonging to the sponsor and grant-funded data belonging to the grantee (typically a university). ICPSR does not acquire copyright to databases but instead requests permission to redistribute. Barriers to data access and sharing in the social sciences include generally weak federal requirements to archive and provide access to research data and the heterogeneity of expectations across fields (with economics, demography, sociology, and criminology having a stronger tradition of data sharing than anthropology and epidemiology).

In a recent ICPSR study on data-sharing and archiving practices, researchers surveyed principal investigators from NIH- and NSF-funded projects and asked whether their projects had produced data and, if so, whether the data had been archived (see Figure 4-2). Of the 1,599 responses received as of late 2008, 327 studies had been archived, 876 studies were still in the hands of researchers, and 396 studies had been "lost."

making data available on a long-term basis diffuses more widely and becomes easier to use.

Research Institutions, Research Libraries, and Repositories

Institutional and disciplinary digital data repositories have been growing steadily. The emergence of open access software tools for building repositories (such as DSpace, EPrints, and Fedora), external repository hosting services,

Preserving and sharing social sciences data involves the risk of violating an individual's privacy. Each data collection is reviewed to see if it could reveal individual identities. If such information is found, it is removed, masked, or collapsed in the public-use version. ICPSR staff are trained and certified in disclosure risk limitation procedures. Original restricted data can be requested under terms of a contract, and the most sensitive data can be viewed onsite in a nonnetworked "data enclave" with significant security checks.

ICPSR also has a strong educational component. Workshops and courses on research methods in the quantitative social sciences are offered to graduate students and faculty from around the world, mainly in the summer. ICPSR also provides data-driven instructional modules at the undergraduate level to enable teachers to integrate data into the curriculum.

Over time, ICPSR's archival model has proven to be an effective approach to ensuring data integrity, facilitating data sharing, and providing data stewardship across a range of fields and many institutions. Because many social science data are used for secondary analysis, and because the social sciences reward academic producers of general-purpose data, universities see the value of ICPSR, which makes the membership funding model sustainable.

The emerging world of massively complex and voluminous data raises new challenges. There will be no single repository and no single harmonization scheme. Unrestricted access is needed to realize the full value of data, which may lead to greater risk of disclosure and confidentiality breaches. New tools need to be developed to enable the merging of disparate data and communication across disciplines. Building new, dynamic communities around data and cutting-edge research questions will require the collaborative efforts of technologists and domain scientists. A greater focus by institutions and federal sponsors on data preservation and access also will be needed.

a <http://www.icpsr.umich.edu>.

and advances in the cost performance of storage technologies have enabled a proliferation of repository efforts. Private foundations such as the Andrew W. Mellon Foundation have played an important role in supporting repository software development, and continue to invest in new capabilities for the digital stewardship of scholarly work.¹⁰

¹⁰ See the description of the Mellon Foundation's Research in Information Technology program: http://www.mellon.org/grant_programs/programs/rit.

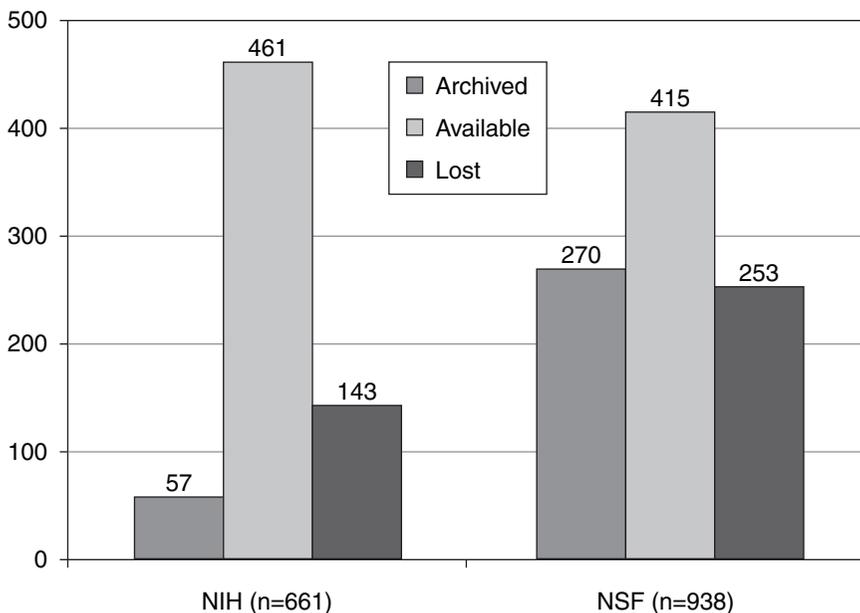


FIGURE 4-2 ICPSR LEADS project findings of NSF- and NIH-sponsored awards that created social science data

NOTE: This figure reflects survey results through November 2008 of principal investigators of 1,599 NIH and NSF awards that indicated social science data creation.

SOURCE: Inter-university Consortium for Political and Social Research (ICPSR). We would like to acknowledge the National Digital Information Infrastructure and Preservation Partnership program at the Library of Congress for supporting this work (NDIIPP Cooperative Agreement 8/04).

Disciplinary repositories accept data and publication submissions regardless of the institutional affiliation of the researcher. One longstanding example is the arXiv publication repository at Cornell University, which focuses on physics and related fields.¹¹

Research institutions typically have more experience with the long-term preservation of data than do individual researchers, especially since many institutions are accustomed to running libraries or archiving offices. In recent years, many research institutions have created their own repositories to house data and publications resulting from research at the institution. One example is the IDEALS repository at the University of Illinois at Urbana-Champaign.¹² UIUC faculty, staff, and students can deposit materials into IDEALS, which

¹¹ <http://arxiv.org/>.

¹² <http://www.ideals.uiuc.edu/>.

can then be accessed by anyone over the Internet. Many repository efforts are led by university libraries, which have begun exploring the new issues posed by research data and other digital information as increasingly central components of the scholarly record.¹³

These efforts are part of a trend in which some research institutions, large research universities in particular, are reassessing their institutional role in the dissemination and stewardship of scholarship, both that of their own faculty and more broadly.¹⁴ During the time when the scholarly record was primarily print-based, a relatively small number of research libraries, most connected with research institutions, saw comprehensive stewardship of scholarship as part of their missions. Likewise, in the digital age, some research institutions and their libraries are likely to play leadership roles in the stewardship of research data.

Institutional repositories naturally face challenges—for instance, building faculty awareness and participation—even at large institutions.¹⁵ A recent report on the role of research libraries in providing repository services identifies several key issues as repositories develop and grow.¹⁶ The issues include building new services (as the focus expands from publications, theses, and dissertations to research data, courseware, images, and other content), engaging with the larger networked environment (as the demand grows for higher-level, cross-repository services), attending to the “demand side” (meeting the needs of heterogeneous user groups), and sustainability (going beyond money to organizational commitment).

Smaller institutions that seek to fulfill a stewardship mission face even greater challenges. The size and complexity of digital datasets can overwhelm small institutional libraries or archives, which traditionally have dealt with analog textual information. Yet new partnerships and approaches hold the promise of overcoming many of these barriers. For example, the National Institute for Technology and Liberal Education now offers institutional repository services to member institutions for an annual fee.¹⁷

¹³ Anna Gold. 2007. “Cyberinfrastructure, data, and libraries.” *D-Lib Magazine* 13(9/10). Available at <http://www.dlib.org/dlib/september07/gold/09gold-pt1.html>.

¹⁴ Clifford A. Lynch. 2008. A matter of mission: Information technology and the future of higher education. Pp. 43–50 in *The Tower and the Cloud*, ed. Richard Katz. Boulder, CO: EDUCAUSE. Available at <http://www.educause.edu/thetowerandthecloud>.

¹⁵ Philip M. Davis and Matthew J. L. Connolly. 2007. Institutional repositories: Evaluating the reasons for non-use of Cornell University’s installation of DSpace. *D-Lib Magazine* 13(3/4). Available at <http://www.dlib.org/dlib/march07/davis/03davis.html>.

¹⁶ ARL Digital Repository Issues Task Force. 2009. *The Research Library’s Role in Digital Repository Services*. Washington, DC: Association of Research Libraries. Available at <http://www.arl.org/bm~doc/repository-services-report.pdf>.

¹⁷ http://www.nitle.org/index.php/nitle/information_services/dspace_services.

Federal Agencies, Data Centers, and Digital Archives

Federal agencies and other funding organizations can play key roles in preserving research data. In some fields, such as the earth and environmental sciences, federal agencies play a central role in the collection and stewardship of research data. For example, the National Oceanographic and Atmospheric Administration (NOAA) collects, manages, and disseminates a wide range of climate, weather, ecosystem and other environmental data used by scientists, engineers, resource managers, policy makers, and others in the United States and around the world. NOAA must deal with the challenges of an increasing volume and diversity of its data holdings—which include everything from satellite images of clouds to the stomach contents of fish—as well as a large number of users.

A recent National Research Council report offered nine general principles for effective environmental data management, along with a number of guidelines on how the principles could be applied at NOAA.¹⁸ The principles and guidelines developed for NOAA are consistent with the principles laid out in this study, and represent an example of how they apply to an agency with significant data management responsibilities in the earth sciences. The description of NOAA's data management challenges also illustrates the challenges of providing access and stewardship for large, heterogeneous datasets.

In some fields, federal agencies have established large digital archives that house important collections of data provided by grantees and other external researchers. NCBI at the National Library of Medicine is perhaps the best example. NCBI houses several data and literature collections, provides education, and develops software for various computational biology applications. GenBank, which has been discussed previously, is a large database of nucleotide sequences that has become an essential national and global resource in the life sciences.¹⁹

Federal agencies have traditionally supported the data management needs of the research fields with which they work most closely. NSF is undertaking a large initiative explicitly focused on developing capabilities to meet longer-term data stewardship needs across science and engineering fields.²⁰ The Sustainable Digital Data Preservation and Access Network (DataNet) program intends to make about five awards totaling \$100 million over 5 years to organizations that will “provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline.” By adapting to and driving technological changes in serving their given domains,

¹⁸ National Research Council. 2007. *Environmental Data Management at NOAA: Archiving, Stewardship, and Access*. Washington, DC: The National Academies Press.

¹⁹ 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.

²⁰ See <http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.pdf>.

the awardees would be helping to demonstrate the feasibility of long-term digital stewardship.

In other fields where federal agencies themselves are not as central to data collection and stewardship efforts, federal capabilities may be more limited. In these cases, nonfederal research sponsors need the support and active participation of research institutions and communities if they are to help ensure the long-term preservation and availability of data. Also, sponsors may be more interested in the initial development of data collections than in maintaining those collections over long periods as an open-ended commitment.

The federal government can also foster data exchange among research institutions and companies in specific, highly applied areas. For example, the Government-Industry Data Exchange Program (GIDEP) is a joint activity of the military services, other federal agencies such as the National Aeronautics and Space Administration and the Department of Energy, defense and space contractors such as Lockheed Martin, Boeing, and Raytheon, and even the Canadian Department of National Defence.²¹ GIDEP has existed since the 1950s, and is a mechanism for sharing research, development, design, testing, acquisition, and logistics information among government and industry participants in order to reduce or eliminate expenditures.

In recent years, other organizations and networks, including data centers, have taken on important roles in the stewardship of research data. The San Diego Supercomputer Center (SDSC), managed by the University of California at San Diego, is a high-performance computing center and a national data hosting facility, providing an integrated set of data services (access, manipulation, management, and storage).²² SDSC is a data services provider for the Protein Data Bank and the National Virtual Observatory (NVO). For NVO, SDSC stores two replicants of the Sloan Digital Sky Survey as well as other sky surveys, over 88 terabytes in all. SDSC DataCentral also hosts over 100 community data collections, including Molecular Dynamics Simulation Data (chemistry), Human Brain Dynamics Resource data (neuroscience), and Employment Responses to Global Markets data (economics).

SDSC's agreements with research communities vary substantially with regard to standards, sharing, formats and ontologies, usage scenarios, and intellectual property. SDSC utilizes multiple levels of data reliability and data integrity mechanisms.

Research communities and data centers such as SDSC need to develop common understanding on key issues such as trust, expectations, incentives/penalties, and privacy/security/confidentiality. Good long-term stewardship requires resources for increased capacity, up-to-date reliability tools, and skilled people. Developing sustainable economic models for long-term stewardship is

²¹ <http://www.gidep.org/>.

²² Francine Berman, Director, SDSC, presentation to the committee, September 17, 2007.

a challenge. In some very high-priority areas, federal support on an “infinite mortgage” basis might be sustainable. In other cases, some combination of relay funding, user fees, endowments, and other mechanisms may need to be employed.

Companies and Journals

Opportunities for new public-private partnerships for data stewardship also exist. For example Google had announced a free service named Palimpsest that would make massive datasets accessible to researchers, but canceled the official launch of the project in late 2008.²³ At the same time, Amazon has launched a service to host large public datasets, allowing researchers to upload their own data.²⁴ Researchers would be charged fees for online data storage and data analysis capability. Many datasets have become so large that they are impossible to download over the Internet in a reasonable time.

Some journals play a role in maintaining the data submitted to support published articles. Journals are also participating in initiatives such as Portico, an archive of electronic scholarly literature.²⁵ However, many journals lack the financial resources for maintaining databases for extended periods. And many journals face financial constraints, especially as they make the transition to electronic publication, which could threaten their ability to preserve and supply data either now or in the future.

ANNOTATING DATA FOR LONG-TERM USE

As noted in Chapter 2, raw data are typically of use only to the research group that generated them. To be useful to others, data must be accompanied by metadata that describe the content, structure, processing, access conditions, and source of the data in a form that permits the data to be used by researchers, educators, policy makers, and others. For computational data, for example, annotation might mean preserving the software used to generate the data along with a simulation of the hardware on which the software ran (or, in some cases, the hardware itself). For observational data, the documentation of the hardware, instrumental calibrations, preprocessing of data, and other circumstances of the observation are generally essential for using the data. In some cases, these metadata can be generated automatically, but annotation can be a labor-intensive process.

²³ Alexis Madrigal. 2008. Google shuts its science data service. *Wired Science*. December 18. Available at <http://blog.wired.com/wiredscience/2008/12/googlescienceda.html>.

²⁴ Aaron Rowe. 2008. Amazon hosting, crunching massive public databases. *Wired Science*. December 5. Available at <http://blog.wired.com/wiredscience/2008/12/massive-amounts.html>.

²⁵ See the Portico Web site: <http://www.portico.org/>.

Different types of users of data generally have different needs for annotation. Researchers in the same field can be expected to need less metadata than a researcher in a quite different field or a nonresearcher. Making data usable in the latter case can be difficult and involved, and researchers do not have a responsibility to make data understandable to a nonexpert. However, guidelines should exist for the degree of expertise required to use a dataset.

E-science that ranges widely across research fields requires standardized interfaces and protocols to enable useful communication across widely separated research fields. However, there is a trade-off between the demands of interoperability between research fields and detailed annotation within a field.²⁶

FOSTERING DATA STEWARDSHIP FOR THE BROAD RESEARCH ENTERPRISE

Most of the discussion in this chapter involves overseeing and promoting data stewardship in individual fields of research. There is also the question of how the broad research enterprise should develop data management standards and long-term strategies across all fields of research, both within and outside government. Many issues are common to multiple fields.

In late 2007, the Blue Ribbon Task Force on Sustainable Digital Preservation and Access was created to “analyze previous and current models for sustainable digital preservation, and identify current best practices among existing collections, repositories and analogous enterprises.”²⁷ The Task Force is developing recommendations and a research agenda aimed at catalyzing and supporting sustainable economic models for stewardship of digital information, including research data. The Task Force is supported by NSF, the Andrew W. Mellon Foundation, and several other organizations. NSF’s DataNet program, described earlier in this chapter, is seeking to develop technologies and organizational capabilities that would be broadly applicable to long-term data stewardship in science and engineering.

Within the U.S. federal government, the Interagency Working Group on Digital Data under the National Science and Technology Council has been examining the needs for preservation and dissemination of publicly funded research data. In January 2009 the working group released its report, *Harnessing the Power of Digital Data for Science and Society*. The report provided goals and implementation plans for the federal government to work, as both leader and partner, with other sectors to enable reliable and effective digital data preservation and access. The working group noted, as we have in this report, that “communities of practice are an essential feature of the digital landscape”

²⁶ Christine L. Borgman. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

²⁷ See blueribbontaskforce.sdsc.edu.

and that “preservation of digital scientific data is both a government and private sector responsibility and benefits society as a whole.” To provide reliable management of digital scientific data, the working group calls for “a comprehensive framework of transparent, evolvable, extensible policies and management and organizational structures that provide reliable, effective access to the full spectrum of public digital scientific data.”

The goals and recommendations of the working group are complementary to those of our committee. The working group recommends that federal agencies “promote a data management planning process for projects that generate preservation data.” These plans should identify the types of data and their expected impact, specify relevant standards, and outline provisions for protection, access, and continuing preservation. The working group’s report points out that not all digital scientific data need to be preserved and not all preserved data need to be preserved indefinitely. Stakeholders that should be involved in decisions about which data to preserve include research communities, data professionals, data users, entities such as professional organizations and governments, and preservation organizations.

In addition, the working group calls for the creation of a subcommittee on digital scientific data preservation, access, and interoperability under the National Science and Technology Council that would track and recommend policies on such issues as national and international coordination; education and workforce development; interoperability; data systems implementation and deployment; and data assurance, quality, discovery, and dissemination.

At the nongovernmental level, in fall 2008 the National Research Council established a new Board on Research Data and Information. The board is engaged in planning, program development, and administrative oversight of projects dealing with the management, policy, and use of digital data and information for science and the broader society. The board’s primary objectives are to:

1. Address emerging issues in the management, policy, and use of research data and information at the national and international levels.
2. Through studies and reports of the National Research Council, provide independent and objective advice, reviews of programs, and assessment of priorities concerning research data and information activities and interests of its sponsors.
3. Encourage and facilitate collaboration across disciplines, sectors, and nations with regard to common interests in research data and information activities.
4. Initiate or respond to requests for consensus studies, workshops, conferences, and other activities within the board’s mission, and provide oversight for the activities performed under the board’s auspices.
5. Broadly disseminate and communicate the results of the board’s activities to its stakeholders and to the general public.

GENERAL PRINCIPLE FOR ENHANCING THE STEWARDSHIP OF RESEARCH DATA

Data are a critical part of the research infrastructure, with an importance comparable to that of laboratories, research facilities, and computing devices and networks. Researchers need to access data quickly and from multiple sources. Data need to be annotated so that they can be used by researchers in a wide variety of fields. Data need to be migrated to successive storage platforms as technologies evolve. These observations lead to the committee's third general principle.

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.

As with the two previous broad principles, this principle is not a recommendation but a general statement of intent that can guide specific actions. Also, as with the Data Access and Sharing Principle, the Data Stewardship Principle's reference to future uses should be seen as limiting rather than broadening the scope of the principle. Decisions must continually be made about which data to save and which data to discard. General heuristics offer some guidance on these decisions.²⁸ Observational data that cannot be re-collected are candidates for being archived indefinitely. Experimental data may or may not be saved depending on whether the experimental conditions can be reproduced precisely at minimal cost. In general, decisions about data retention require focused attention within each research group and field.

Many critical questions involving the retention of data are not directly addressed by the Data Stewardship Principle. For how long should data be retained? In what format and by whom? Who should pay for the preservation of data? These questions can be answered only by the researchers, research institutions, research sponsors, and policy makers who have responsibility for data stewardship.

RESPONSIBILITIES OF RESEARCHERS

As with ensuring the integrity and accessibility of data, researchers have unique responsibilities for data stewardship. As stated in an editorial for its issue on "petabyte science," which appeared in September 2008, the journal *Nature* states that "Researchers need to be obliged to document and manage

²⁸ National Science Board. 2005. *Long-Lived Data Collections: Enabling Research and Education in the 21st Century*. Arlington, VA: National Science Foundation.

their data with as much professionalism as they devote to their experiments. And they should receive greater support in this endeavor than they are afforded at present.”²⁹

Through their planning and actions, researchers facilitate or complicate the retention of data. Researchers need to provide much of the metadata that can allow data to be used in the future by colleagues who may be in quite different fields. Only the researchers and data professionals directly involved in a project know their data well enough to judge what should be preserved and what should be discarded. The heterogeneity of data and the variety of possible needs argue that policies and strategies be set by those within a field, not outside it.

Among the most important tasks for researchers establishing a data management plan is to arrange for preserved data to be annotated in such a way that they retain their long-term value. Annotation might include computer codes, algorithms, or other processing techniques used in the course of research. Furthermore, this information should be sufficient to allow other researchers not only to verify previous results but to extend those results into new areas.

Data stewardship must start at the beginning of a project, not partway through or at the end of a 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.

At a minimum, data management plans for research projects should provide for compliance with the relevant legal and policy requirements covering research data. These would include institutional policies, sponsor requirements, federal law (e.g. the 1996 Health Insurance Portability and Accountability Act), and state law as appropriate. Under certain circumstances (e.g., when the data can be reproduced cheaply, no secondary use is anticipated), provisions for stewardship of the data beyond what is legally required may not be necessary. In other cases, the data management plan would specify whether the data would be deposited in an institutional and/or disciplinary repositories, annotation and metadata specifications, and other elements.

This recommendation does not imply that individual researchers are responsible for ensuring indefinite preservation of their own data, only that they ensure that it is prepared and transferred to the appropriate archives or repositories. Also, researchers should be working in partnership with their institutions, sponsors, and fields in formulating and implementing their plans.

Researchers need to participate in the development of policies and standards for data access, annotation, and preservation, including standards regard-

²⁹ Editorial. 2008. “Community cleverness required.” *Nature* 455(7209). Available at <http://www.nature.com/nature/journal/v455/n7209/pdf/455001a.pdf>.

ing the degree of expertise needed to use the data. Establishing such policies is the collective responsibility of the researchers in each field, given the potential value of data to future researchers in that field and others.

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.

As research data become more voluminous, complex, and valuable, a need may arise to formalize the process of making data management decisions within research fields. As with data access and data integrity, international participation may be needed in the development of data management standards, or international organizations might take the lead role. Often ad hoc groups can provide guidance, such as National Research Council committees, federal agency advisory groups, or collaborative efforts such as the one undertaken by the Ecological Society of America and described in Box 4-2. In some fields it might become desirable to charge a data oversight board with this responsibility. Such a board could serve many functions including the following:

- Make recommendations about whether data should be stored in special repositories or by individuals.
- Determine how long particular kinds of data need to be preserved and who is responsible for the quality of the data as they move from one storage platform to another.
 - Inventory and publicize good practices for data management.
 - Conduct assessments of which datasets offer the most potential future value and which can be sacrificed.
 - Organize interactions with specialized support organizations, either nonprofit or commercial, to store and distribute data.
 - Evaluate access and preservation to identify problems and ensure that data with the greatest potential utility are being preserved.

As was discussed in Chapter 3, science, engineering, and medical research is a global enterprise. A wide range of governmental and private entities around the world have developed expertise in areas related to data stewardship, many working at the level of disciplines and fields.³⁰ Professional societies and indi-

³⁰ Raivo Ruusalepp. 2008. *Infrastructure Planning and Data Curation: A Comparative Study of International Approaches to Enabling the Sharing of Research Data*. Data Curation Centre and Joint Information Systems Committee (UK). November. Available at http://www.dcc.ac.uk/docs/publications/reports/Data_Sharing_Report.pdf

BOX 4-2 The Ecological Society of America's Data-Sharing Initiative

The Ecological Society of America (ESA), which was founded in 1912, consists of more than 10,000 scientists from diverse fields studying ecological restoration, biotechnology, ozone depletion, species extinction, and many other topics.^a All of the ESA journals archive their electronic publications using Portico, which preserves “scholarly literature published in electronic form and ensure[s] that these materials remain accessible to future scholars, researchers, and students.”^b Funded by the Andrew W. Mellon Foundation, Ithaca, the Library of Congress, and JSTOR, Portico was launched in 2005 and has almost 6 million journal articles archived. The ESA requires that data and information on methods and materials needed to verify conclusions be made available to editors of its journals on request, and strongly encourages authors to register their data in ESA’s official registry (data.esa.org).

ESA also has devoted considerable attention to making unpublished foundational data accessible. In 2004 it formed a joint working group to promote data sharing and archiving. Representatives of the working group came from many organizations and a wide range of fields. Over the course of three meetings, the working group discussed the promotion and design of data registries,^c the role of data centers,^d and obstacles to data sharing.^e

In addition, ESA is working to establish a National Ecological Data Center (NEDC), which would be a repository for metadata and datasets. The NEDC would feature a directory of connected data centers, an online manual, training, and free access.^f

^a <http://www.esa.org/aboutesa/>.

^b http://www.portico.org/about/portico_brochure.pdf.

^c http://www.esa.org/science_resources/DocumentFiles/DataRegistry_WorkshopReportFinal.pdf.

^d http://www.esa.org/science_resources/DocumentFiles/ESA_Data_Centers_Wkshp_notes.pdf.

^e http://www.esa.org/science_resources/DocumentFiles/DataObstacles_Wkshp_notesFinal.pdf.

^f http://esa.org/science_resources/DocumentFiles/visionstatement_nedc.pdf.

vidual U.S. researchers should be encouraged to participate in and lead international efforts to improve research data stewardship.

RESPONSIBILITIES OF RESEARCH INSTITUTIONS, RESEARCH SPONSORS, AND JOURNALS

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.

Research institutions and research sponsors have an interest in seeing data used to full advantage. Research data represent a sizable investment of human and financial resources, and preserving those data typically costs less than generating them in the first place. Nevertheless, maintaining high-quality and reliable databases can have significant costs. Because future uses of data are difficult to predict, the return on those costs can be uncertain. In many fields, there still is no consensus as to who should maintain large databases or who should bear the costs.

Depending on the field, data management plans might include incentives for proper data stewardship (including research sponsor policies and conditions for grants and contracts), investments in technological and institutional tools, standardization of interfaces, and the support of data centers. The examples of the Ecological Society of America and ICPSR (Boxes 4-1 and 4-2) show how fields and coalitions of fields can develop policies and capacity for data stewardship over time.

Research institutions, including research libraries, can play leadership roles in the stewardship of research data, both those produced by their own faculties and more broadly. As with the preservation of scholarship in the print era, not every institution will be positioned to develop comprehensive capabilities by itself. Coalitions and partnerships among institutions and between institutions and agencies can accomplish much of this work.

It is important that requirements for improved data management practices not be imposed as unfunded mandates. They need to be integrated into research program funding as an essential component of the conduct of research. Where possible, grant applications should include costs for data stewardship.

The questions of who pays, how much, and for how long are at the heart of the problem of how to ensure long-term stewardship of research data. It has been suggested that only the federal government is positioned to guarantee the preservation of research data, and that a federal data archive or system of archives analogous to the Library of Congress should be established to undertake this mission.

This chapter discusses the variety of federal resources and programs related to research data stewardship that already exist, many of which involve partnerships of various types with research fields and research institutions. Many of them are relatively new. This committee was not in a position to comprehensively evaluate whether the current, largely decentralized, approach is likely to meet the needs of the research enterprise. The relevant communities are actively engaged in addressing these issues, through groups such as the Blue Ribbon Task Force for Sustainable Digital Preservation and Access mentioned earlier.

5

Defining Roles and Responsibilities

ASSIGNING ROLES AND RESPONSIBILITIES

Periods of rapid technological change offer strong incentives as well as unique opportunities for examining current policies and instituting new policies to address changing circumstances. Every part of the research enterprise is being affected by the changes in how research is being planned, conducted, and used, and each has responsibilities for ensuring the integrity, accessibility, and stewardship of research data. However, shared responsibilities can create problems. When responsibility is shared, each group can assume that the other groups should be the ones taking action. As a speaker at one of the committee's meetings memorably described the problem, "If two people are responsible for feeding a dog, that dog's going to starve."

The remainder of this chapter revisits the recommendations made in the three preceding chapters by briefly describing the roles and responsibilities of the major sectors of the research enterprise in ensuring the integrity, accessibility, and stewardship of research data. In that regard, it functions as a summary of the report's recommendations, though the recommendations are resorted according to the groups responsible for each action (see Table 5-1). It also discusses some of the particular responsibilities incumbent on parts of the research enterprise to avoid inaction caused by an overly diffuse allocation of responsibilities.

RESEARCHERS

Researchers have particular obligations in each of the three areas discussed in this report. As data producers, providers, and users, they know best how to generate data of high quality, disseminate data to others so that the data are useful, and preserve the data for future uses. In some fields they may need to work in close association with data professionals. They might also carry out

TABLE 5-1 Responsibilities of Groups Within the Research Enterprise

Recommendation	Research Institutions and Research Libraries				
	Researchers	Research Sponsors	Professional Societies	Journals	Public
Data Integrity					
Manage projects to ensure data integrity	✓				
Receive appropriate training for data management	✓				
Participate in development of professional standards for management of data	✓	✓	✓	✓	
Provide support for training in data management		✓	✓		
Recognize appropriateness of financial support for data professionals		✓			
Help ensure that contributions of data professionals are recognized and rewarded	✓	✓	✓	✓	
Data Access and Sharing					
Make data and other information integral to reported results accessible in a timely manner	✓				
Ascertain whether data are publicly accessible and, if not, whether restrictions are appropriate	✓				
If data are not accessible, explain why publicly	✓				
If necessary, develop standards for data accessibility (with parties appropriate for field)	✓	✓	✓	✓	✓
Except where restrictions are justified, require that data be made available		✓		✓	
Promote sharing of data through public recognition of outstanding data-sharing efforts, funding, and publication policies		✓	✓	✓	
Establish clear policies for management of and access to research data					✓

Ensure availability of data in short and long term	✓	✓	✓	✓	✓
Data Stewardship					
Include provisions for stewardship of data in data management plans	✓				
Document, reference, and index data with long-term value	✓				
Develop process to generate guidance for researchers about data retention	✓	✓			
Develop and implement plans with researchers to meet needs for data stewardship	✓	✓	✓		
Develop incentives and tools for data stewardship		✓	✓		
Publicize and promote proper data stewardship		✓	✓		✓
Consider supporting data centers and archives, and stewardship tools		✓	✓		✓

their responsibilities through informal groups or formal organizations created with the involvement of funding agencies or professional societies.

In a period of rapid technological change, researchers can be challenged to master all of the information they need to fulfill their responsibilities toward data. Training in the responsible conduct of research that includes guidance on the management of data can clarify and emphasize researchers' responsibilities (Chapter 2). Many research data have potential uses and users that may not be obvious from the perspective of a single research field. Courses, seminars, or Web-based modules in data management can list and describe these potential uses and users, providing researchers with a more comprehensive set of factors to consider in making decisions about data accessibility and stewardship.

Researchers also need to be aware of the many considerations surrounding data when they are considering possible restrictions on data and the appropriateness of any such restrictions (Chapter 3). Restrictions may be necessary, yet most restrictions on the accessibility of data have costs for the research community. Because of these costs, researchers have a responsibility to provide compelling reasons for any limitations on the accessibility of data, which requires that they fully understand and are able to justify these limits.

Finally, researchers are the ones best positioned to plan both how data will be made available and how they will be preserved and curated for long-term use (Chapter 4). When standards for data accessibility and stewardship do not exist in a field, researchers need to be involved in—and most likely will lead—the process of developing such standards.

The integrity, accessibility, and stewardship of research data are too important to be secondary considerations or afterthoughts in the development of a research plan. Provisions for maintaining these three qualities of research data should be part of every research plan, whether a sponsor requires such provisions or not.

RESEARCH INSTITUTIONS

Research institutions, including colleges, universities, medical schools, and other nonprofit organizations, have a major influence on the policy environment in which research is conducted. Their support—or lack of support—for data integrity, accessibility, and stewardship can have a major effect on the quality and usability of research data. Research institutions need to have clear written policies regarding data management and communicate these policies to researchers. Organizations such as the National Association of State Universities and Land-Grant Colleges, the American Association of Universities, the Committee on Government Relations, the Committee on Institutional Cooperation, and others can help formulate and disseminate these policies.

Research institutions need to support training in data management (Chapter 2). They should establish an expectation that researchers will undertake

such training and provide the financial support for researchers to be able to do so. Research institutions and sponsors also facilitate the development of data professionals by providing career paths for these individuals, supporting their training, and recognizing and rewarding their contributions.

Researchers have many incentives for maintaining the integrity of the data they generate. They have fewer incentives, in general, for making their data widely available, and fewer still to invest the time and resources needed to ensure the stewardship of data. Policy initiatives are therefore essential if research data are to achieve their maximum value.

Research institutions have a special responsibility to be proactive in making research data accessible (Chapter 3). Research grants and contracts typically give research institutions ownership rights in research data, and so those institutions have a particular interest in seeing that research data are available, that restrictions on the accessibility of research data are justified, and that procedures exist for responding to requests for research data. Both formal policies and informal expectations help to avoid conflicts over data accessibility.

Research institutions also can and should play the leading role in stewardship of its scholarship and knowledge resources (Chapter 4).

RESEARCH SPONSORS

Research sponsors, including government agencies, philanthropies, private companies, and other funders, also have an interest in all three of the qualities discussed in this report. But they have a particular responsibility toward data stewardship (Chapter 4). The infrastructure needed for data stewardship is much less developed than is the infrastructure for publishing research conclusions. Also, the long-term preservation of data in a usable form can be costly, and research data are so varied across fields that different systems are needed for different fields.

Funders can maximize the value of the research they fund by also taking steps to support the stewardship of data. They need to work with researchers in the fields they sponsor to develop incentives for researchers to invest in data stewardship, and they need to consider support for the data centers and tools that facilitate stewardship.

PROFESSIONAL SOCIETIES AND JOURNALS

Finally, professional societies and journals have important roles to play in all three of the areas explored in this report. They can help develop and disseminate guidelines for a research field and then help monitor and enforce compliance with those guidelines. Journals are directly responsible for the long-term preservation of the articles they publish, and an increasing number of journals are assuming responsibility for maintaining the data on which research conclu-

sions are based. And journals and professional societies can help ensure that the contributions of data professionals are recognized and rewarded through such mechanisms as prizes, publication, and recognition at disciplinary meetings.

In general, more dialog is needed among researchers, research institutions, and research sponsors about the need for education and training, how sponsors should support the stewardship of data, the role of data professionals, and how institutions and sponsors should respond to reasonable and unreasonable requests for research data. Professional societies and journals can catalyze these dialogues within research fields, providing a base of knowledge that can then be applied across disciplines.

CONCLUSION

During periods of rapid change, an emphasis on specific policies may be less useful than reiterating and reemphasizing the fundamental principles that should guide action. Thus, we close by restating three general principles that have motivated our recommendations in the areas of data integrity, accessibility, and stewardship.

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.

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

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.

Appendix A

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

COMMITTEE MEMBERS

DANIEL KLEPPNER, Co-Chair, is professor emeritus at the Massachusetts Institute of Technology (MIT) and co-director of the MIT-Harvard Center for Ultracold Atoms. He has made fundamental contributions to atomic physics and quantum optics. His research encompasses spectroscopic tests of extreme precision and novel quantum phenomena. He was director of the MIT-Harvard Center for Ultracold Atoms from 2000 to 2006, and from 1987 to 2000 he was associate director of the MIT Research Laboratory of Electronics. In 1960, along with Norman Ramsey, he developed the Hydrogen maser, later used as an atomic clock of unprecedented stability. Applications of this early work range from coordination of radio signals in long-baseline radio astronomy, to satellite-based global positioning systems.

In the 1970s, Dr. Kleppner was a pioneer in the physics of Rydberg atoms, demonstrating the inhibition of spontaneous emission from them. This was a pioneering step in the development of cavity quantum electrodynamics, the study of the radiative properties of atoms in confined spaces. Kleppner's investigations of Rydberg atom spectra in high electric and magnetic fields provided deep physical insight into the implications of classical chaos for quantum systems.

Professor Kleppner and MIT colleague Professor Thomas Greytak were among the first to search for quantum degeneracy effects in ultra-cold gases. After a 20-year-long quest, in 1998, they achieved Bose-Einstein condensation (BEC) in hydrogen. In the meanwhile, they developed tools instrumental to the 1995 discovery of BEC in alkali atoms by MIT alumni Eric Cornell and Carl Wieman, and MIT's Wolfgang Ketterle. These include the technique of evaporative cooling, developed in collaboration with Harald Hess. Bose-Einstein

condensates and fermionic degenerate samples of cold atoms represent a new form of matter at the lowest temperatures ever achieved. These species are now the subject of intense investigation in laboratories around the world.

In addition to these research achievements, Dr. Kleppner has been a dedicated teacher at the undergraduate and graduate levels, and has served on numerous national committees charged with investigating key scientific or social issues. His honors include election to the National Academy of Sciences, the American Academy of Arts and Sciences, the American Philosophical Society, and the Academies of Science (Paris), and the Davison-Germer Prize, Leo Szilard Lectureship Award and Lilienfeld Prize of the American Physical Society, the Oersted Medal of the American Association of Physics Teachers, the Frederick Ives Medal of the Optical Society of America, the Wolf Prize, and the 2006 National Medal of Science.

PHILLIP A. SHARP, Co-Chair, is Institute Professor at the Massachusetts Institute of Technology. Much of Dr. Sharp's scientific work has been conducted at MIT's Center for Cancer Research (now the Koch Institute), which he joined in 1974 and directed from 1985 to 1991. He subsequently led the Department of Biology from 1991 to 1999 and the McGovern Institute from 2000 to 2004. His research interests have centered on the molecular biology of gene expression relevant to cancer and the mechanisms of RNA splicing; his landmark achievement was the discovery of RNA splicing in 1977. This work provided one of the first indications of the startling phenomenon of "discontinuous genes" in mammalian cells. The discovery that genes contain nonsense segments that are edited out by cells in the course of utilizing genetic information is important in understanding the genetic causes of cancer and other diseases. Dr. Sharp's research opened an entirely new area in molecular biology and forever changed the field. For this work he shared the 1993 Nobel Prize in Physiology or Medicine with Dr. Richard Roberts who did work in parallel at Cold Spring Harbor.

Dr. Sharp has authored more than 350 scientific papers and serves on many scientific committees, including the National Cancer Institute's Advisory Board, which he chaired for two years (2000–2002). His work has been honored with numerous awards including the Gairdner Foundation International Award, General Motors Research Foundation Alfred P. Sloan, Jr. Prize for Cancer Research, Louisa Gross Horwitz Prize, and Albert Lasker Basic Medical Research Award. He is an elected member of the National Academy of Sciences, the Institute of Medicine, the American Academy of Arts and Sciences, and the American Philosophical Society.

A native of Kentucky, Dr. Sharp earned a B.A. degree from Union College, Kentucky, and a Ph.D. in chemistry from the University of Illinois at Champaign-Urbana in 1969. He did his postdoctoral training at the California Institute of Technology, where he studied the molecular biology of plasmids from bacteria

in Professor Norman Davidson's laboratory. Prior to joining MIT, he was senior scientist at Cold Spring Harbor Laboratory.

Dr. Sharp is co-founder of Biogen, Inc., 1978, chairman of the Scientific Board (to 2002) and member of the board of directors. He is also co-founder of Alnylam Pharmaceuticals (2002), where he serves as chairman of the Scientific Board and as a member of the company's board of directors.

MARGARET A. BERGER is widely recognized as one of the nation's leading authorities on scientific evidentiary issues, in particular DNA evidence, and is a frequent lecturer across the country on these topics. She is a recipient of the Francis Rawle Award for outstanding contributions to the field of postadmission legal education by the American Law Institute/American Bar Association for her role in developing new approaches to judicial treatment of scientific evidence and in educating the legal and science communities about ways to implement these approaches. Professor Berger serves as a member of the National Academy of Sciences Committee on Science, Technology, and Law. She recently completed her service as a member of the National Commission on the Future of DNA Evidence in which she served as the reporter for the Working Group on Post-Conviction Issues. She has been called on as a consultant to the Carnegie Commission on Science, Technology, and Government, and has served as the Reporter to the Advisory Committee on the Federal Rules of Evidence. She is the author of numerous amicus briefs, including the brief written for the Carnegie Commission on the admissibility of scientific evidence in the landmark case of *Daubert v. Merrell Dow Pharmaceuticals, Inc.* She has also contributed a chapter on "The Supreme Court's Trilogy on the Admissibility of Expert Testimony" to the *Reference Manual on Scientific Evidence* (2nd ed. 2000). Her textbook, *Evidence: Cases and Materials* (9th ed. 1997) (with Weinstein, Mansfield, and Abrams), is the leading evidence casebook. Professor Berger has been a member of the faculty of Brooklyn Law School in New York since 1973, and holds the Suzanne J. and Norman Miles Chair.

NORMAN M. BRADBURN, the Tiffany and Margaret Blake Distinguished Service Professor Emeritus of the University of Chicago, serves on the faculties of the Irving B. Harris Graduate School of Public Policy Studies, the Department of Psychology, the Graduate School of Business, and the college. He is a former provost of the university (1984–1989), chairman of the Department of Behavioral Sciences (1973–1979), and associate dean of the Division of the Social Sciences (1971–1973). From 2000 to 2004 he was the assistant director for social, behavioral and economic sciences at the National Science Foundation. Bradburn is currently a senior fellow at the National Opinion Research Center (NORC). Associated with NORC since 1961, he has been director of NORC and president of its board of trustees.

A social psychologist, Bradburn has been at the forefront in developing

theory and practice in the field of sample survey research. He has focused on psychological well-being and assessing the quality of life, particularly through the use of large-scale sample surveys; nonsampling errors in sample surveys; and research on cognitive processes in responses to sample surveys. His book, *Thinking About Answers: The Application of Cognitive Process to Survey Methodology* (with Seymour Sudman and Norbert Schwarz; Jossey-Bass, 1996), follows three other publications on the methodology of designing and constructing questionnaires: *Polls and Surveys: Understanding What They Tell Us* (with Seymour Sudman; Jossey-Bass, 1988); *Asking Questions: A Practical Guide to Questionnaire Construction* (with Seymour Sudman; Jossey-Bass, 1982; 2nd edition with Brian Wansink, 2004) and *Improving Interviewing Method and Questionnaire Design* (Jossey-Bass, 1979).

Bradburn serves on the board of directors of the Chapin Hall Center for Children. He was chair of the Committee on National Statistics of the National Research Council/National Academy of Sciences (NRC/NAS) from 1993 to 1998, and is past president of the American Association of Public Opinion Research (1991–1992). Bradburn chaired the NRC/NAS panel to advise the Census Bureau on alternative methods for conducting the census in the year 2000. The report, published as *Counting People in the Information Age*, was presented to the Census Bureau in October 1994. He was a member of the NRC/NAS Panel to Review the National Assessment of Educational Progress and the Panel to Assess the 2000 Census. He is currently one of the domain chairs for the Key National Indicators Initiative at the National Academy of Sciences. Bradburn was elected to the American Academy of Arts and Sciences in 1994. In 1996 he was named the first Wildenmann Guest Professor at the Zentrum für Umfragen, Methoden und Analyse in Mannheim, Germany.

JOHN BRAUMAN was born in Pittsburgh, Pennsylvania, in 1937. He attended Massachusetts Institute of Technology (S.B., 1959) and the University of California at Berkeley (Ph.D., 1963). He was a National Science Foundation postdoctoral fellow at University of California at Los Angeles, and then took a position at Stanford University where he is J. G. Jackson–C. J. Wood Professor of Chemistry Emeritus. He was department chair, associate dean for natural sciences, and has been associate dean of research since 2005. He also currently serves as the Home Secretary of the National Academy of Sciences.

Dr. Brauman has received a number of awards including the American Chemical Society Award in Pure Chemistry, Harrison Howe Award, Guggenheim Fellowship, R. C. Fuson Award, Arthur C. Cope Scholar Award, James Flack Norris Award in Physical Organic Chemistry, National Academy of Sciences Award in Chemical Sciences, Linus Pauling Medal, Willard Gibbs Medal, and National Medal of Science. He is a member of the National Academy of Sciences, the American Academy of Arts and Sciences, the American Philosophical Society, a fellow of the American Association for the Advancement of Science,

and an honorary fellow of the California Academy of Sciences. He received the Dean's Award for Distinguished Teaching from Stanford University in 1976. Dr. Brauman has served on many national committees and advisory boards. He was deputy editor for Physical Sciences for *Science* from 1985 to 2000 and is currently the chair of the senior editorial board.

Dr. Brauman's research has centered on structure and reactivity. He has studied ionic reactions in the gas phase, including acid-base chemistry, the mechanisms of proton transfers, nucleophilic displacement, and addition-elimination reactions. His work includes inferences about the shape of the potential surfaces and the dynamics of reactions on these surfaces. He has made contributions to the field of electron photodetachment spectroscopy of negative ions, measurements of electron affinities, the study of dipole-supported electronic states, and multiple photon infrared activation of ions. He has also studied mechanisms of solution and gas-phase organic reactions as well as organometallic reactions and the behavior of biomimetic organometallic species.

JENNIFER T. CHAYES is managing director of the new Microsoft Research New England lab in Cambridge, Massachusetts which opened in July 2008. Before this, she was research area manager for Mathematics, Theoretical Computer Science and Cryptography at Microsoft Research Redmond. Chayes joined Microsoft Research in 1997, when she co-founded the Theory Group. Her research areas include phase transitions in discrete mathematics and computer science, structural and dynamical properties of self-engineered networks, and algorithmic game theory. She is the co-author of almost 100 scientific papers and the co-inventor of more than 20 patents.

Chayes has many ties to the academic community. She is affiliate professor of mathematics and physics at the University of Washington, and was for many years professor of mathematics at UCLA. She serves on numerous institute boards, advisory committees and editorial boards, including the Turing Award Selection Committee of the Association for Computing Machinery, the board of trustees of the Mathematical Sciences Research Institute, the advisory boards of the Center for Discrete Mathematics and Computer Science and the Miller Institute for Basic Research in Science, the U.S. National Committee for Mathematics and the Committee on Assuring the Integrity of Research Data of the National Academies, the Advisory Committee on Women in Computing of the Association for Computing Machinery, the Leadership Advisory Council of the Anita Borg Institute for Women and Technology, and the Selection Committee for the Anita Borg Award for Technical Leadership. Chayes is a past chair of the mathematics section of the American Association for the Advancement of Science, and a past vice president of the American Mathematical Society.

Chayes received her B.A. in biology and physics at Wesleyan University, where she graduated first in her class, and her Ph.D. in mathematical physics

at Princeton. She did her postdoctoral work in the mathematics and physics departments at Harvard and Cornell. She is the recipient of a National Science Foundation Postdoctoral Fellowship, a Sloan Fellowship, and the UCLA Distinguished Teaching Award. She has twice been a member of the Institute for Advanced Study in Princeton. Chayes is a fellow of the American Association for the Advancement of Science, and a National Associate of the National Academies.

Chayes is best known for her work on phase transitions, in particular for laying the foundation for the study of phase transitions in problems in discrete mathematics and theoretical computer science; this study is now giving rise to some of the fastest known algorithms for fundamental problems in combinatorial optimization. She is also one of the world's experts in the modeling and analysis of random, dynamically growing graphs—which are used to model the Internet, the World Wide Web and a host of other technological and social networks. Among Chayes' contributions to Microsoft technologies are the development of methods to analyze the structure and behavior of various networks, the design of auction algorithms, and the design and analysis of various business models for the online world.

Chayes lives with her husband, Christian Borgs, who also happens to be her principal scientific collaborator. In her spare time, she enjoys overworking.

ANITA K. JONES is a university professor and the Lawrence R. Quarles Professor of Engineering and Applied Science at the University of Virginia. She came to the University in 1988 to serve as chair of the Department of Computer Science. Professor Jones served as the director of defense research and engineering for the U.S. Department of Defense from 1993 to 1997, where she managed the department's science and technology program. She has served on the boards of several government organizations including as the vice chair of the National Science Board. She is a member of the National Academy of Engineering, the Defense Science Board, the Charles Starke Draper Foundation, the board of trustees of InQTel, the governing board of Science Foundation Arizona, and the MIT Corporation Executive Committee. Professor Jones is a fellow of several professional societies and she has been awarded honorary doctorate degrees by Carnegie Mellon University and Duke University. She has been awarded the Department of Defense Award for Distinguished Public Service, the Ada Lovelace Award from the Association of Women in Computing, and the Founder's Award of the Institute of Electrical and Electronics Engineers. The U.S. Navy named a seamount in the North Pacific Ocean (51° 25' N and 159° 10' W) for her.

LINDA P. B. KATEHI is the provost and vice chancellor for academic affairs at the University of Illinois at Urbana-Campaign and professor of electrical and computer engineering. She holds a joint appointment with the Program of

Gender and Women Studies at the University of Illinois. As a faculty member, Professor Katehi has focused her research on the development and characterization of three-dimensional integration and packaging of high-frequency circuits with particular emphasis on MEMS devices, high-Q passives, and embedded filters. She pioneered the development of on-wafer packaging for high-density, high-frequency monolithic Si-based circuit and antenna architectures that led to low-cost, high-performance integrated circuits for radar, satellite, and wireless applications. Her work in this area has led to numerous national and international technical awards and to distinctions as an educator. Professor Katehi holds 13 U.S. patents and has authored more 500 papers published in refereed journals and symposia proceedings.

Professor Katehi is a member of the National Academy of Engineering, a fellow of American Association of the Advancement of Science (AAAS), and a fellow of IEEE. She serves on many scientific committees including the Nominations Committee for the National Medal of Technology, the board of AAAS, the Kauffman National Panel for Entrepreneurship, the National Science Foundation (NSF) Advisory Committee to the Engineering Directorate, the National Research Council (NRC) Telecommunications Board, the NRC Army Research Lab Advisory Committee on Sensors and Electronics Division, the NSF Advisory Committee to CISE, the National Aeronautics and Space Administration Aeronautics Technical Advisory Committee, and the Department of Defense Advisory Group on Electron Devices.

Professor Katehi earned her diploma degree from the National Technical University of Athens, Greece, in 1977 from the School of Mechanical and Electrical Engineering. Following her undergraduate studies, she worked for 2 years as a senior engineer in the Naval Research Lab and joined the University of California at Los Angeles as a graduate student in fall 1979, completing an M.S.E.E. in December 1981 and a Ph.D. in electrical engineering in 1984. From 1984 to 2002 she was a faculty member of the Electrical Engineering and Computer Science Department of the University of Michigan in Ann Arbor, where she served as the associate dean for academic affairs from 1998 to 2002. From 2002 until 2004 she served as the dean of engineering and as faculty member of the Electrical and Computer Engineering Department at Purdue University.

NEAL F. LANE is the Malcolm Gillis University Professor at Rice University. He also holds appointments as a senior fellow of the James A. Baker III Institute for Public Policy, where he is engaged in matters of science and technology policy, and in the Department of Physics and Astronomy. Prior to returning to Rice University, Dr. Lane served in the federal government as assistant to the president for science and technology and director of the White House Office of Science and Technology Policy from August 1998 to January 2001, and as director of the National Science Foundation (NSF) and member (ex officio) of the

National Science Board. Prior to joining NSF, Dr. Lane was provost and professor of physics at Rice University in Houston, Texas, a position he had held since 1986. He first came to Rice as an assistant professor in the Department of Physics and later became professor of physics and space physics and astronomy. He left Rice from mid-1984 to 1986 to serve as chancellor of the University of Colorado at Colorado Springs. In addition, from 1979 to 1980, while on leave from Rice, he worked at the NSF as director of the Division of Physics. Dr. Lane's many writings and presentations include topics in theoretical atomic and molecular physics and science and technology policy. Dr. Lane has received numerous prizes and awards. He is a fellow of the American Academy of Arts and Sciences. He also serves on several boards and advisory committees. Born in Oklahoma City in 1938, Dr. Lane earned his B.S., M.S., and Ph.D. degrees in physics from the University of Oklahoma.

W. CARL LINEBERGER is currently serving as professor of chemistry at the University of Colorado. He was elected to the National Academy of Sciences in 1983. His work is primarily experimental, using a wide variety of laser-based techniques to study structure and reactivity of gas-phase ions. Recent studies have been directed toward elucidating the structure of transient reaction intermediates, to developing understanding of the gradual evolution of physical properties from an isolated molecule to a solvated species, and to real-time investigations of reaction dynamics.

RICHARD LUCE is vice provost and director of libraries at Emory University. He is responsible for managing the main library—including specialist libraries in business, chemistry, music and media, as well as the Manuscript, Archives, and Rare Books Library—and coordinating university-wide library policy with the directors of the health, law, theology, and Oxford College libraries. Prior to joining Emory, Mr. Luce was the research library director at Los Alamos National Laboratory (1991–2006). Known as an information technology pioneer and organizational innovator, he managed a world-class scientific research library and forged regional, national, and international public information and technology collaborations. In 1999 he was a co-founder of the Open Archives Initiative to develop interoperable standards for author self-archiving systems. In October 2003 he co-organized the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, and in 2004, the Brazilian Declaration on Open Access. He holds numerous advisory and consultative positions supporting digital library development, electronic publishing, and scholarly communication. He was the senior advisor to the Max Planck Society's Center for Information Management (2000–2006) and an executive board member of the National Information Standards Organization (1998–2004). He was the recipient of the 2005 Fellows' Prize for Leadership at Los Alamos National Laboratory, the first ever awarded to a nonscientist.

Mr. Luce was the course director of the International Spring School on the Digital Library and E-Publishing for Science and Technology in Geneva and a founding member and chair of the Alliance for Innovation in Science and Technology Information. He received a Distinguished Performance Award from Los Alamos for his contributions supporting science and technology. Prior to Los Alamos, Mr. Luce held positions as the first executive director of the Southeast Florida Library Information Network, director of Colorado's Irving Library Network, and assistant director of the Boulder Public Library in Colorado. He speaks extensively in the areas of digital libraries and scientific communication, quality and change management, and strategic planning. Luce holds a bachelor's degree in political science from the University of San Diego, a master's degree in public administration from San Diego State University, and a master's degree in library and information science from the University of South Florida.

THOMAS O. MCGARITY is Joe R. and Teresa Lozano Long Endowed Chair at the University of Texas at Austin School of Law. He was articles editor of the Texas Law Review. Thomas McGarity has studied both administrative law and environmental law. He also teaches torts. He is currently serving as co-reporter for rulemaking on the American Bar Association's restatement project of the Administrative Procedures Act and related statutes. He received his J.D. from the University of Texas. He has written three influential books: *Workers at Risk* (Praeger, 1993) (co-author), *The Law of Environmental Protection* (West, 2nd ed., 1991) (co-author), and *Reinventing Rationality: The Role of Regulatory Analysis in the Federal Bureaucracy* (Cambridge University Press, 1991). His recent articles include "On the Prospect of Daubertizing Judicial Review of Risk Assessment" (*Law & Contemporary Problems* 2003). He currently serves as president of the Center for Progressive Reform.

STEVEN M. PAUL is the executive vice president for science and technology and president of Lilly Research Laboratories (LRL), a division of Eli Lilly and Company. He also is a member of the corporate policy and strategy and operations committees and the company's senior management council, a group of top Lilly executives who implement corporate strategies, ensure corporate performance, and identify corporate issues and opportunities. In 2005, Dr. Paul was named Chief Scientific Officer of the Year at one of the annual pharmaceutical achievement awards. He joined Lilly in April 1993 as vice president of central nervous system discovery and decision phase medical research in LRL and was named vice president, therapeutic area discovery research and clinical investigation, in 1996. Dr. Paul became group vice president of therapeutic area discovery research and clinical investigation for LRL in 1998. Paul received a B.A. degree, magna cum laude with honors, in biology and psychology from Tulane University in 1972. He received an M.Sc. degree in anatomy and neuroanatomy

and his doctor of medicine degree, both in 1975, from the Tulane University School of Medicine. Prior to joining Lilly, Paul served as scientific director of the Intramural Research Program of the National Institute of Mental Health (NIMH); professor of psychiatry at Tulane University School of Medicine; and chief of the clinical neuroscience branch, as well as chief of the section on pre-clinical studies at NIMH. Dr. Paul is a member of various professional societies, and he was listed as one of the most highly cited neuroscientists in the world (1980–2000) by the Institute for Scientific Information. Dr. Paul serves on the editorial boards of numerous scientific journals and on several NIH extramural and intramural committees. Paul serves on the board of directors of the Lilly Foundation, the Foundation of the NIH, Butler University and the Indianapolis Zoological Society. He is a member of the Institute of Medicine.

TERESA A. SULLIVAN became provost and executive vice president for academic affairs at the University of Michigan in 2006. She is also professor of sociology in the College of Literature, Science, and the Arts. Prior to coming to the University of Michigan, Dr. Sullivan was executive vice chancellor for academic affairs for the University of Texas System, a position she held from 2002 until May 2006. In that role, she was the chief academic officer for the nine academic campuses within the University of Texas System. Her responsibilities included developing tuition-setting procedures, initiating and supporting educational and research collaborations among the various campuses, and developing external collaborations. Dr. Sullivan first joined the University of Texas at Austin in 1975 as an instructor and then assistant professor in the Department of Sociology. From 1977 to 1981, she was a faculty member at the University of Chicago. Dr. Sullivan returned to Texas in 1981 as a faculty member in sociology. In 1986, she was named to the Law School faculty as well. Dr. Sullivan also held several administrative positions at Texas including vice president and graduate dean (1995–2002), vice provost (1994–1995), chair of the Department of Sociology (1990–1992), and director of Women’s Studies (1985–1987). Dr. Sullivan’s research focuses on labor force demography, with particular emphasis on economic marginality and consumer debt. The author or co-author of six books and more than 50 scholarly articles; her most recent work explores the question of who files for bankruptcy and why. Dr. Sullivan has served as chair of the U.S. Census Advisory Committee. She is past secretary of the American Sociological Association and a fellow of the American Association for the Advancement of Science. A graduate of James Madison College at Michigan State University, Dr. Sullivan received her doctoral degree in sociology from the University of Chicago.

MICHAEL S. TURNER is the Bruce V. and Diana M. Rauner Distinguished Service Professor at the University of Chicago. He was born in Los Angeles, California, attended University High School, received his B.S. in physics from

the California Institute of Technology (1971) and his Ph.D. in physics from Stanford University (1978). He came to the University of Chicago in 1978 as an Enrico Fermi Fellow and joined the faculty in 1980. From 2003 to 2006, Turner served as the assistant director of the National Science Foundation for the Mathematical and Physical Sciences, and from 2006 to 2008 as chief scientist at Argonne National Laboratory.

From 1997 to 2003 Turner was chair of the Department of Astronomy & Astrophysics at Chicago, and from 1998 to 2001 he was the first scientific spokesperson for the Sloan Digital Sky Survey. He was instrumental in establishing the Kavli Institute for Cosmological Physics at the University of Chicago in 2001. In 1983, with Edward Kolb, he established the Theoretical Astrophysics Group at Fermilab, which today is part of the larger Center for Particle Astrophysics at Fermilab. Turner is currently a member of the board of directors and the executive committee of the Fermi Research Alliance, which manages Fermilab for the Department of Energy. Since 1984 he has been on the board of trustees of the Aspen Center for Physics and from 1989 to 1993 served as its president.

Turner is a fellow of the American Physical Society, the American Association for the Advancement of Science, and the American Academy of Arts and Sciences, and he is a member of the National Academy of Sciences. Turner has been honored with the Helen B. Warner Prize of the American Astronomical Society, the Julius Edgar Lilienfeld Prize of the American Physical Society, the Halley Lectureship at Oxford University, the Klopsteg Lecture Award of the American Association of Physics Teachers, the Quantrell Award for Excellence in Undergraduate Teaching at the University of Chicago and an honorary Doctor of Science degree from Michigan State University. In 2006, he received the Distinguished Alumnus Award from Caltech, and in 2009 he will give the Biermann Lectures at the Max Planck Institute for Astrophysics in Garching.

Turner helped to pioneer the interdisciplinary field that has brought together cosmologists and elementary particle physicists to unravel the origin and evolution of the universe and to understand the unification of the fundamental forces and particles of nature. His research focuses on the earliest moments of creation, and he has made seminal contributions to inflationary cosmology, particle dark matter and structure formation, the theory of big-bang nucleosynthesis, and the nature of dark energy that is causing the expansion of the universe to speed up. He believes that cosmic acceleration is the most profound mystery in all of science today, and he coined the term “dark energy.” Dark energy is the focus of his current research.

Turner has served on and chaired numerous committees for the Department of Energy, the National Aeronautics and Space Administration, National Science Foundation, the American Physical Society, and the National Academies. The National Academy study *Connecting Quarks with the Cosmos*, which he led, identified opportunities at the intersection of astronomy and physics and

has shaped the science investment in the United States and elsewhere around the world. Turner is currently the chair of the Physics Section of the National Academy of Sciences and the chair-elect of the Division of Astrophysics within the American Physical Society.

J. ANTHONY (TONY) TYSON is Distinguished Professor of Physics at the University of California at Davis and the director of the Large Synoptic Survey Telescope (LSST). LSST will look wide, fast, and deep, scanning the entire night sky every three nights for 10 years. Its mission will be to map the mysterious “dark matter” and “dark energy” that physicists say make up 95 percent of the universe. His research interests are in cosmology, dark matter, dark energy, observational optical astronomy, experimental gravitational physics, and new instrumentation. He received his Ph.D. from University of Wisconsin in 1967 and was a member of the technical staff at Bell Laboratories from 1969 to 2003. His honors include election to the American Philosophical Society and the National Academy of Sciences, the Aaronson Memorial Prize, and fellowships in the American Academy of Arts and Sciences and the American Physical Society.

STEVEN C. WOFSY is the Abbott Lawrence Rotch Professor of Atmospheric and Environmental Sciences in the Department of Earth and Planetary Sciences at Harvard University. Dr. Wofsy holds a Ph.D. in chemistry from Harvard University. He studies a variety of atmospheric gases using instruments aboard aircraft and also on the ground at long-term measurement sites. His research interests include undertaking theoretical and modeling studies to understand depletion of stratospheric ozone in polar regions, to assess future impacts of pollutants injected into the stratosphere, and to examine ecological and historical factors affecting atmospheric concentrations of CO₂. In 2001, Dr. Wofsy received the Distinguished Public Service Medal from the National Aeronautics and Space Administration. He is a fellow of the American Geophysical Union and the American Association for the Advancement of Science.

Appendix B

Relevant National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and National Research Council Reports

On Being a Scientist: Responsible Conduct in Research, Third Edition (2009)
Committee on Science, Engineering, and Public Policy

Synopsis: Describes the ethical responsibilities of researchers, using case studies. Treatment of data is one of the topics covered. Provides an overall framework for responsible research practices that underlies this study's discussion on ensuring the integrity of data.

Models in Environmental Regulatory Decision Making (2007)
Committee on Models in the Regulatory Decision Process, National Research Council

Synopsis: Examines the use of models by the Environmental Protection Agency in the regulatory process, and recommends a life-cycle management approach to developing, testing, and revising models. Developing environmental regulations relies on both data and models. Principles outlined in the report, such as the importance of peer review and of providing accurate descriptions of a model's assumptions, are analogous to this study's principles for providing access to data and metadata.

Environmental Data Management at NOAA: Archiving, Stewardship, and Access (2007)

Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA, National Research Council

Synopsis: The National Oceanographic and Atmospheric Administration (NOAA) collects, manages, and disseminates a wide range of climate, weather, ecosystem, and other environmental data used by scientists, engineers, resource managers, policy makers, and others in the United States and around the world. The increasing volume and diversity of NOAA's data holdings—which

include everything from satellite images of clouds to the stomach contents of fish—and a large number of users present NOAA with substantial data management challenges. The report offers nine general principles for effective environmental data management, along with a number of guidelines on how the principles could be applied at NOAA. The principles and guidelines developed for NOAA are consistent with the accessibility and stewardship principles laid out in this study, and represent an example of how they apply to an agency with significant data management responsibilities in the earth sciences. The description of NOAA's data management challenges also illustrates the challenges of providing access and stewardship for large, heterogeneous datasets.

Science and Security in a Post 9/11 World (2007)

Committee on a New Government-University Partnership for Science and Security

Synopsis: Explores various aspects of science and security, including access to data and movement of students and researchers across borders. Upholds the principle that the results of unclassified basic research should not be restricted.

Surface Temperature Reconstructions for the Last 2,000 Years (2006)

Committee on Surface Temperature Reconstructions for the Last 2,000 Years, National Research Council

Synopsis: Examines the use of proxy evidence from multiple sources to reconstruct surface temperatures. In addition to its main conclusions about the reliability of multiproxy reconstructions, the report points out the differences in approaches to data availability in the fields covered, and that open access to data and methods will improve public confidence in the results of this research.

Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation, and Public Health (2006)

Committee on Intellectual Property Rights in Genomic and Protein Research and Innovation, National Research Council

Synopsis: Explores intellectual property (IP) issues related to genomic and protein research, identifies areas where emerging practices in patenting and sharing data or research resources might impede research, and recommends steps that federal agencies, research institutions, and companies should take to prevent IP protections from impeding future breakthroughs. Access to and sharing of research data are addressed in several recommendations.

Improving Business Statistics Through Interagency Data Sharing: Summary of a Workshop (2006)

Caryn Kuebler and Christopher Mackie, Rapporteurs, Steering Committee for the Workshop on the Benefits of Interagency Business Data Sharing, National Research Council

Synopsis: Describes the benefits of greater sharing of business and other data among federal agencies, the barriers (mainly the need to maintain confidentiality), and possible approaches. Covers issues of data access relevant to economics and other social sciences.

Expanding Access to Research Data: Reconciling Risks and Opportunities (2005)
Panel on Data Access for Research Purposes, National Research Council

Synopsis: Focuses on expanded access to microdata from studies conducted by federal statistical agencies under pledges of confidentiality. Describes barriers to data access that are common in the social sciences, and develops approaches to overcoming them.

Building an Electronic Records Archive at the National Archives and Records Administration (NARA): Recommendations for a Long-Term Strategy (2005)

Committee on Digital Archiving and the NARA, National Research Council
Synopsis: Develops a comprehensive long-term strategy for how the NARA should approach archiving digital data. Many of the issues and barriers identified in the report, and the recommended strategies for addressing them, are relevant to a wide range of research fields and organizations charged with stewardship of research data.

Improving Data to Analyze Food and Nutrition Policies (2005)

Panel on Enhancing the Data Infrastructure in Support of Food and Nutrition Programs, Research, and Decision Making, National Research Council

Synopsis: Examines existing data sources used to support policy making and policy evaluation in food and nutrition programs. Recommends steps to strengthen the data infrastructure in this area. A good example of an end-use-motivated inventory of open and proprietary data sources.

Electronic Scientific, Technical, and Medical Journal Publishing and Its Implications: Report of a Symposium (2004)

Committee on Electronic Scientific, Technical, and Medical Journal Publishing and Its Implications and Committee on Science, Engineering and Public Policy, The National Academies

Synopsis: Summarizes a symposium that considered the changing digital environment for scholarly publishing.

Licensing Geographic Data and Services (2004)

Committee on Licensing Geographic Data and Services, National Research Council

Synopsis: Addresses the growing practice whereby federal agencies license geographic data from private vendors for their own use and for the use of outside researchers. Provides guidelines for when and under what circumstances agen-

cies should enter such agreements, and describes complementary strategies, such as creation of a National Commons and Marketplace in Geographic Data, to maximize access to data for research and other uses. A careful examination of a field where access to private data is necessary for the advance of research. These guidelines may become applicable to other fields in the future.

***Seeking Security: Pathogens, Open Access, and Genome Databases* (2004)**
Committee on Genomics Databases for Bioterrorism Threat Agents, National Research Council

Synopsis: Examines the security implications of access to genomic data, concluding that continued open access to genomic data is the best approach. Recommends that professional societies educate researchers about the risks of research results being misused. An example of a field in which open access is the best approach to ensuring security.

***Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences* (2003)**
Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council

Synopsis: The publication of experimental results and sharing of research materials related to those results have long been key elements of the life sciences. Over time, standard practices have emerged from communities of life scientists to facilitate the presentation and sharing of different types of data and materials. But recently a concern has emerged that, in practice, publication-related data and materials are not always readily available to the research community. This report finds that the life sciences community does possess commonly held ideas and values about the role of publication in the scientific process. Those ideas define the responsibilities of authors and underpin the development of community standards: practices for sharing data, software, and materials adopted by different disciplines of the life sciences to facilitate the use of scientific information and ensure its quality. The report is a very clear and thorough exploration of standards and expectations for making data accessible in an important field. The principles developed—that authors are required to make data available as a quid pro quo for publication, that authors are obligated to provide data and other materials in a form on which scientists can build further with research, and that all members of the scientific community have equal responsibility for upholding community standards—are consistent with those recommended by this study, and represent something of a “gold standard” that other fields might try to emulate.

Government Data Centers: Meeting Increasing Demands (2003)

Committee on Coping with Increasing Demands on Government Data Centers, National Research Council

Synopsis: Describes the increasing demands on government data centers that store and provide access to environmental data, and technical approaches to ensure effective operation in the future. In the earth and environmental sciences, the federal government has a major responsibility for the stewardship of data. Provides an overview of the issues and makes recommendations for technical approaches that might be used by the centers and users. These approaches might have relevance to other fields.

Ensuring the Quality of Data Disseminated by the Federal Government: Workshop Report (2003)

Committee on Ensuring the Quality of Government Information, National Research Council

Synopsis: Summarizes discussion at a series of workshops involving agencies and researchers to discuss implementation of the Data Quality Act. Provides background on the Data Quality Act, which is an important part of the policy context for this study's discussion of the integrity and accessibility of data.

The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium (2003)

Julie M. Esanu and Paul F. Uhler, Editors, National Research Council

Synopsis: Papers from a symposium on how the scientific community can maintain and expand the public domain for scientific and technical data and information. The papers explore many aspects of the intellectual property environment for research.

Access to Research Data in the 21st Century: An Ongoing Dialogue Among Interested Parties, Report of a Workshop (2002)

Science, Technology, and Law Panel, National Research Council

Synopsis: A workshop on issues related to the Data Access Act (the Shelby Amendment) which was adopted in 2000. Points out that peer review does not detect fraud or substitute for the judgment of the scientific community as a whole; it provides advice to a journal editor about the importance of the findings and whether the reported evidence supports the author's claims. Illustrates the barriers to making data available, particularly in fields where data can be used to identify individuals. Also illustrates the pros and cons of various approaches to ensuring the accessibility of data, including that of the Data Access Act, which is modeled on the Freedom of Information Act.

Toward New Partnerships in Remote Sensing: Government, the Private Sector, and Earth Science Research (2002)

Steering Committee on Space Applications and Commercialization, National Research Council

Synopsis: Much of the remote sensing data needed for earth sciences research are now provided by private sector entities, and are made available to the federal government and university researchers through various licensing agreements and partnership arrangements. The report evaluates these arrangements and makes recommendations for how they should be structured in order to best advance science. The report explores intellectual property issues involved when private sector data is obtained for use in government and university environments. The principles developed might be useful for other fields where data generated by the private sector might be utilized to advance research.

Integrity in Scientific Research: Creating an Environment That Promotes Responsible Conduct (2002)

Committee on Assessing Integrity in Research Environments, National Research Council, Institute of Medicine

Synopsis: Provides a high-level view on research integrity and how it can be promoted. Much of the focus is on institutional approaches to education and self-assessment. Consistent with this study's findings and recommendations on institutional responsibility.

Geoscience Data and Collections: National Resources in Peril (2002)

Committee on the Preservation of Geoscience Data and Collections, National Research Council

Synopsis: Describes the importance of geoscience data and collections and the challenges of stewardship. Develops criteria for prioritizing geoscience data and collections to be preserved, and recommends a specific strategy for doing so. A case study of the tension between devoting resources to creating new data and preserving existing data. A good example of how criteria can be developed on a disciplinary basis for making these judgments.

Assessment of the Usefulness and Availability of NASA's Earth and Space Science Mission Data (2002)

Task Group on the Usefulness and Availability of NASA's Space Mission Data, National Research Council

Synopsis: Calls on NASA to devote more resources and management attention to data stewardship, including ensuring compatibility with parallel data efforts such as the National Virtual Observatory. Earth and space science examples illustrating the importance of data reuse.

Preparing for the Revolution: Information Technology and the Future of the Research University (2002)

Panel on the Impact of Information Technology on the Future of the Research University, National Research Council

Synopsis: Broad overview of information technology changes and their implications for the research university. Calls attention to the institutional role in preserving and disseminating knowledge, including data.

Transforming Remote Sensing Data into Information and Applications (2001)
Steering Committee on Space Applications and Commercialization, National Research Council

Synopsis: Examines possibilities for applying remote-sensing data to new applications and the implications for policy. Illustrates the value of data reuse while also recognizing that developing new applications may carry considerable costs. Points out the lack of standard data protocols and formats as a barrier to using data for new applications.

Issues for Science and Engineering Researchers in the Digital Age (2001)
Office of Special Projects, National Research Council

Synopsis: A broad overview of how information technology is transforming science and engineering research, and the implications for researchers. Highlights the importance of ensuring the quality of digital data and the challenges of stewardship.

Resolving Conflicts Arising from the Privatization of Environmental Data (2001)
Committee on Geophysical and Environmental Data, National Research Council

Synopsis: Defines appropriate spheres for the public and private sectors in the growing field of environmental data. Recommends that the public sector should continue to collect and synthesize data, and to provide such data at no more than the marginal cost of reproduction with no usage restrictions. The private sector would focus on value-added distribution and specific observational systems.

Improving the Collection, Management, and Use of Marine Fisheries Data (2000)
Ocean Studies Board, National Research Council

Synopsis: Describes the current system of data collection, management, and use in the marine fisheries field, and recommends improvements. Illustrates the growing need to work across sectors to improve data quality and stewardship in a “small science” field that is highly relevant to policy.

Bioinformatics: Converting Data to Knowledge: Workshop Summary (2000)
A Workshop Summary by Robert Pool and Joan Esnayra, Board on Biology, National Research Council

Synopsis: Summary of a workshop on data issues related to bioinformatics. Illustrates how the growing availability of data is transforming science and engineering.

Improving Access to and Confidentiality of Research Data: Report of a Workshop (2000)

Christopher Mackie and Norman Bradburn, Editors, National Research Council

Synopsis: Explores the challenges of improving access to data with confidentiality restrictions. The challenge of improving access to data with confidentiality restrictions goes across several fields.

The Digital Dilemma: Intellectual Property in the Information Age (2000)
Committee on Intellectual Property Rights in the Emerging Information Infrastructure, National Research Council

Synopsis: In-depth examination of copyright issues, including those related to digital archiving, in the wake of the Digital Millennium Copyright Act. Relevant to the changing environment for scientific publishing, an important aspect of the context for this study, as well as the role of libraries.

A Question of Balance: Private Rights and Public Interest in Scientific and Technical Databases (1999)

Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest, National Research Council

Synopsis: Describes the importance of scientific and technical databases in research, and standard practices for production, dissemination, and use of data in federal, nonprofit, and commercial contexts. Develops principles and guidelines for agencies, research institutions, and investigators. Explores various proposals for creating new intellectual property protection for noncopyrightable databases current at the time of the study, along with the pros and cons of these proposals. The European Union had recently created such protection. Several of the principles and guidelines are consistent with this study, including: (1) scientific and technical data owned or controlled by the government should be made available for use by not-for-profit and commercial entities alike on a nonexclusive basis and should be disseminated to all users at no more than the marginal cost of reproduction and distribution, whenever possible; (2) federal funding agencies should require university and other not-for profit researchers or their employing institutions that use federal funds, wholly or in substantial part, in creating databases not to grant exclusive rights to such databases when

submitting them for publication or for incorporation into other databases. Also provides a good overview of intellectual property issues related to data. Data itself is not copyrightable, and there are significant limitations on copywriting databases. The policy context has not changed much since the time of this report, as the United States and other nations have not followed the European Union to create new intellectual property protection for databases.

Finding the Path: Issues of Access to Research Resources (1999)

Committee on Federal Policy for Access to Research, Resources, National Research Council

Synopsis: This conference summary describes issues affecting access to a variety of research resources in the life sciences, including data and databases, materials, software, and so forth. Provides background on data access issues in the life sciences. The recommendations are largely superseded by *Sharing Publication-Related Data and Materials* (2003).

Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making: Workshop Report (1999)

Jonathan R. Davis, Vivian P. Nolan, Janet Woodcock, and Ronald W. Estabrook, Editors, Institute of Medicine

Synopsis: Describes the process for assuring the integrity of clinical trial data and suggests improvements. Background to the issues of clinical trials data discussed in this study.

Bits of Power: Issues in Global Access to Scientific Data (1997)

Committee on Issues in the Transborder Flow of Scientific Data, National Research Council

Synopsis: Outlines the needs for access to data in the physical, astronomical, geological, and biological sciences. Characterizes the legal, economic, policy, and technical factors and trends that have an influence on access to data by the scientific community. Identifies and analyzes the barriers to international access to scientific data. Recommends approaches that could help overcome those barriers. The two key challenges are the increasing quantities, varieties, dissemination modes, and interdisciplinary relevance of data, and increasing legal and economic restrictions on publicly funded data. States the principle that “full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from publicly funded research. The public-good interests in the full and open access to and use of scientific data need to be balanced against legitimate concerns for the protection of national security, individual privacy, and intellectual property.” This study would extend this principle somewhat, to include private-sector-funded data on which published research results are based.

Responsible Science: Ensuring the Integrity of the Research Process (1992)
Committee on Science, Engineering, and Public Policy

Synopsis: Broad overview and guidance on how the research enterprise should ensure research integrity. The principles and approaches developed in this study still underlie the definitions, standards, and policies related to ensuring responsible research and dealing with misconduct.

Sharing Research Data (1985)

Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf, Editors, Committee on National Statistics, National Research Council

Synopsis: Explores advantages of and barriers to sharing social sciences data. Early exploration of the idea of asking researchers to provide a data dissemination plan in their proposals, including “the time of release of data, the means by which the data would be made available and preserved for long-term use, the technical form in which data would be released, the supporting documentation that would accompany the data, what forms of access to confidential or other sensitive data would be provided, and an assessment of the policy relevance and broad research value of the data.”

Appendix C

Letters from Scientific Journals Requesting the Study



E-MAILED
5/26/06

Phone: (202) 326-6502
Fax: (202) 289-7562
E-mail: mbradfor@aaas.org

26 May 2006

Dr. Ralph J. Cicerone
President
National Academy of Sciences
500 Fifth Street, NW
Washington, DC 20001



Dear Dr. Cicerone:

Over the last year, a handful of journal editors have been discussing a possible meeting on the topic of digital data handling practices and the potential for inappropriate manipulation and misrepresentation. In an effort to include other stakeholders, we had some discussions with staff at the Office of Research Integrity and the Council of Science Editors. We recognize, however, that the scientists are the most important stakeholders. So we were pleased to learn that the National Academy of Sciences (NAS) is also grappling with how best for the scientific community to address this issue. In a discussion with Ken Fulton, he advised us to share our ideas with you and to offer our cooperation and assistance.

As you know, a few recent highly publicized cases of data manipulation and misrepresentation have raised questions about the validity of the scientific record and the role of academia, authors, reviewers, and journal editors in safeguarding against such misconduct. Much of the attention and discussion thus far has centered on detecting data manipulation and enforcement of appropriate penalties after the fact. The editors of *Cell*, *Science*, *Nature*, and *Nature Cell Biology* would like to contribute in a constructive and comprehensive way to the broader discussion of the ethics of data presentation, in which equal consideration is given to implementing mechanisms to discourage such behaviors before the submission of a manuscript.

Before we learned of the Academy's discussions on this topic, we were planning to propose a meeting that would bring together all of the relevant stakeholders—including scientists, funding agencies, journal editors, and institutional ethics and education committee representatives. The major goal for the meeting was the development of a consensus that can be adopted by funders, scientists and editors. Three potential outcomes of such a meeting would be:

- the establishment of agreed-upon guidelines for best practices in digital data handling;
- a plan of action for increased education and mentoring on the ethics of data presentation;

Headquarters
1200 New York Avenue, NW, Washington, DC 20005 USA Tel: +1 202 326 6550 Fax: +1 202 289 7562
Europe Office
Batemans House, 82-88 Hills Road, Cambridge CB2 1LQ, UK Tel: +44 1223 326500 Fax: +44 1223 326501
Published by the American Association for the Advancement of Science

Dr. Ralph J. Cicerone
26 May 2006
Page 2

- a recommendation of appropriate methods for detection of improper data manipulation and establishment of reasonable and appropriate sanctions for violations of the guidelines in submitted research manuscripts.

As these are all issues that are of core concern and relevance to the broad scientific community, we believe that the impact of such a meeting would be greatly increased if it is conducted under the aegis of the NAS.

We are attaching the agenda we were developing as it spells out more specifically the range of topics that we were considering. Included is a list of scientific societies/journal editors that have expressed interest in working with us to tackle the issue of inappropriate data manipulation. If the NAS is interested in taking the lead, we are all available for consultation or participation in any way that you would find helpful. We do hope that the NAS will be interested in taking on this important task.

We look forward to hearing from you.

Yours sincerely,

Monica Bradford
Executive Editor, *Science*

Katrina Kelner
Deputy Editor for Life Sciences, *Science*

Linda J. Miller
US Executive Editor, *Nature* and the *Nature* research journals

Bernd Pulverer
Editor, *Nature Cell Biology*

Emilie Marcus
Editor, *Cell*, and Executive Editor, Cell Press

Orla Smith
Editor, *Cell's* Leading Edge

MMB/ssk
Attachment

List of Interested Journal Editors/Scientific Societies

Dr. Frank Gannon, Executive Director of EMBO (*EMBO J.*, *EMBO Reports*, *Molecular Systems Biology*) / Les Grivell, Director of *e-BioSci* and *Oriel*

Dr. Gerald Weissmann, Editor-in-Chief, *FASEB J.*

Dr. William Hill, Editor, *Proceedings of the Royal Society B: Biological Sciences*

Dr. Gordon Pike, *Journal of Materials Research*

Dr. Marty Blume, Editor-in-Chief, American Physical Society journals

Dr. Robert Kennicutt, Editor-in-Chief, *Astrophysical Journal*

Dr. Brian Crawford, Senior Vice President, Journals Publishing Group,
American Chemical Society

We are also aware that the *Journal of Cell Biology* is also concerned about the issue of data manipulation.

Draft Agenda: Proposal A

One Day Meeting on Digital Data Manipulation Under the Auspices of the National Academy of Sciences (NAS)

Location: NAS, Washington DC
Duration: One day: 8:30 am to 5:00 pm
Date: Fall 2006
Size: 40 people total

Objective: To bring together leading members of the scientific community, journal editors, and representatives of major funding agencies and institutes to discuss standards for best practice in handling digital images for data presentation.

Goal: To derive an agreed set of guidelines for data presentation and to delineate responsibilities for monitoring adherence to guidelines

Proposed Attendees:

Leading Scientists and Institute Directors
Journal Editors
Funding Agencies
DHHS Office of Research Integrity

Proposed Schedule:

8:30 am Objective and goal of the meeting

8:45 am Round Table Breakfast

Developing appropriate guidelines for handling digital images

Using current journal guidelines as a starting point, could discuss such issues as: Which types of manipulation are acceptable, which are not? Blots and gels, microscopy and micrographs, image contrast and background, conservation of original digital datasets, accurate recording of all instrument settings, avoiding image "beautification," unique features of large complex datasets, e.g., patient datasets, astronomy datasets.

11:45 am Round Table Lunch

Education about Guidelines

Dissemination of the new guidelines: by journals, through institute training programs for postgrads and postdocs, and by inclusion of guidelines in the information for grant/fellowship applications.

Appendix / Page 3

1:15 pm **Monitoring Adherence to Guidelines**

The role of the institute director

The role of the principal investigator

The role of the researcher (including discussion of author responsibility)

The role of the journal editor

The role of the reviewer

The role of the funding agency

3:30 pm **Detecting Data Manipulation**

4:15 pm **Actions if Data Manipulation or Fraud are Suspected**

4:45 pm Concluding Remarks

5:00 pm Close

Draft Agenda: Proposal B

8:30 am Meeting introduction and plan (3-5 slides)

Defining the problem and responsibilities for various stakeholders

8:45 am Biological authors/PIs
9:00 am Physical authors/PIs
9:15 am Funding agencies
9:30 am Referees in physical sciences
9:45 am Referees in biological sciences
10:00 am Journal editors

10:15 am Break

Related considerations

10:30 am Staying true to the data when preparing beautiful images
11:00 am Ethical line: when does enhancement become manipulation?

12:00 pm Lunch – demo of image manipulation detection – slide show

When is data manipulation necessary?

12:45 pm Physicists
1:00 pm Biologists

1:15 pm Review of current guidelines from various journals (to be distributed before the meeting) for *J. Cell Biol.*, *Science*, *Nature*, *Cell*, *Proc. Natl. Acad. Sci.*, *NEJM*, *Lancet*

Development of general guidelines – draft statement

2:00 pm Applicable to all
2:45 pm Breakout groups for guidelines specific to biological or physical sciences
3:30 pm Summaries from the breakout groups compiled into draft guidelines

4:00 pm Break

Educating scientists about manipulation – Roundtable discussion

4:15 pm What institutes/graduate programs/lab chiefs can do
4:30 pm What editors can do
4:45 pm What funding agencies can do

Compliance – Roundtable discussion of possible actions if deceptive manipulation suspected

5:00 pm in the lab
5:15 pm at the agencies
5:30 pm at journals

5:45 pm Wrap-up with action points

Yale University

February 20, 2006

Dr. Ralph J. Cicerone, President
National Academy of Sciences
500 Fifth Street, NW
Washington, DC 20001

Dear Dr. Cicerone,

15.1.8

*Ira Mellman
Professor and Chairman
Department of Cell Biology
School of Medicine
Sterling Hall of Medicine
P.O. Box 30802
New Haven, Connecticut 06520-8002*

*Campus address:
Sterling Hall of Medicine
333 Cedar Street
Telephone: 203 785-4302
Fax: 203 785-4301*

Feb 21 2006

I am writing in my capacity as Editor-in-Chief of The Journal of Cell Biology (JCB), on behalf of my fellow senior academic Editors, to call your attention to a matter that we believe the National Academy of Sciences needs to address. The matter concerns establishing standards for the management of digitally processed data, and for the larger issue of scientific ethics in publication.

For the past four years, the JCB has had a pioneering program to monitor the veracity of digital images provided to us for publication by the authors of accepted manuscripts. Our procedure involves visual inspection of the image files using basic adjustments in Photoshop, and we plan to incorporate the use of detection software being developed by a collaborating mathematician at Dartmouth College.

During the course of this program, overseen by our Executive Editor, Dr. Mike Rossner, we have found that at least 25% of these manuscripts contain one or more figures that must be remade because we detect inappropriate image manipulation, that is, the manipulation violates our guidelines for image presentation, but it does not affect the interpretation of the data. Our standards have been published on our website (www.jcb.org) as well as in a highly distributed article written by Dr. Rossner and my Editorial colleague, Dr. Kenneth Yamada of the National Institutes of Health (Rossner and Yamada, 2004, JCB).

If we suspect fraudulent manipulation that does affect the interpretation of the data, Dr. Rossner contacts the authors requesting original, unmanipulated data. In the large majority of cases, original data were provided and the manipulated images simply needed to be corrected to more accurately reflect those data. In an alarming high number of cases (1% of our accepted manuscripts), however, the authors were unable to provide the original data or provided data that clearly confirmed that fraudulent manipulation had occurred. In such cases, we revoke acceptance of the manuscript and, on occasion, inform the authors' home institution regarding our findings.

The JCB is a high impact, highly selective journal publishing only 15-20% of the papers it receives. Since it is edited by practicing scientists, we have taken the issue of data manipulation very seriously. Our image-screening program helps to ensure the validity of the data we publish (and several other journals, including Science, have taken up this cause), but the prevalence of image manipulation reveals a lapse in the education of young scientists and students on the proper handling of digital image data. The transition of biomedical and biological science to a nearly complete reliance on digital data is relatively recent, but it is already time for the academic

15.1.9

community to establish standards and to educate its trainees on what is or is not appropriate. We have worked, largely without success, to get professional societies to take up this task.

The topic of image manipulation has made news headlines recently amidst the discrediting of the cloning article published in *Science* by Hwang and colleagues. Since our system would have detected at least one anomaly published by these authors, the popular press reported our efforts extensively, capped off by a front page article written in the *New York Times* "Science Times" by Nicholas Wade. We have now been inundated with requests for help from virtually all major biomedical science journals from *Nature* to the *New England Journal of Medicine*, and we are assisting these other journals to adopt our procedures and standards. *Science*, ironically, contacted us long before the Hwang case and has implemented our system, but had not done so by the time the Hwang et al. article was published. We have only recently been in contact with George Kendall, the manager of Production, Marketing, and Licensing at the Proceedings, although the Proceedings has yet to adopt any manipulation guidelines.

Even though my senior Editorial group consists of some of the most illustrious and respected cell biologists in the US and Europe, we do not feel that our efforts comprise a community-sanctioned effort to address the issue. Furthermore, there are a number of other critical problems pertaining to the general issue of ethics and publication, such as authorship and the presentation of other types of data besides images, that need to be addressed.

We feel that the goal of establishing clear and logical standards for ethics in scientific publishing is one that is appropriate to be addressed by a high level panel convened by the National Academy of Sciences. The Academy has addressed other issues of community ethics in the past, such as the sharing of published reagents, and with quite some success. We believe that it will not be at all difficult to arrive at a consensus, and doing so will have two very positive effects. First, it will assure the public that the scientific community is acting in an aggressively responsible fashion to ensure the integrity of publicly-funded science. Second, it will provide a clarion pretext to begin educating students, fellows, and even our colleagues on what is acceptable, and not acceptable, in the digital age.

I or any of my fellow JCB editors would be delighted to discuss this issue with you at any time. Thank you for your consideration.

Yours sincerely,



Ira Mellman
Editor in Chief, *The Journal of Cell Biology*
Sterling Professor of Cell Biology & Immunobiology
Chair, Department of Cell Biology
Scientific Director, Yale Comprehensive Cancer Center

15.1.10



Don Cleveland
Editor, *The Journal of Cell Biology*
Professor of Medicine Neuroscience and Cellular and Molecular Medicine
University of California, San Diego



Pier Paolo Di Fiore
Editor, *The Journal of Cell Biology*
Scientific Director
The FIRC Institute for Molecular Oncology
Milan, Italy



Elaine Fuchs
Editor, *The Journal of Cell Biology*
Rebecca C. Lancefield Professor of Mammalian Cell Biology and Development
Rockefeller University

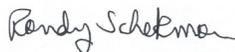


Alan Hall
Editor, *The Journal of Cell Biology*
Director
MRC Cell Biology Unit and Laboratory for Molecular Cell Biology
University College London, UK

15.1.11



Louis Reichardt
Editor, The Journal of Cell Biology
Jack D. and DeLoris Lange Professor of Cell Physiology
University of California, San Francisco



Randy Schekman
Editor, The Journal of Cell Biology
Professor of Molecular and Cell Biology
University of California, Berkeley



Kenneth Yamada
Editor, The Journal of Cell Biology
Chief, Craniofacial Developmental Biology and Regeneration Branch
National Institute of Dental and Craniofacial Research
National Institutes of Health



Junying Yuan
Editor, The Journal of Cell Biology
Professor of Cell Biology
Harvard Medical School

Index

A

- Access to research data. *See also* Open;
Sharing data
- Bermuda statement, 60
- confidentiality or privacy considerations, 5, 67
- costs of limiting access, 70-71, 77-78
- cross-discipline diversity in, 5, 27, 63-64, 88, 90, 136
- cyberinfrastructure and, 61-62, 83-84, 87, 135
- data mining tools, 20-21
- data professional's role, 90
- definition of accessibility, 26
- economic considerations, 26, 59, 69-70, 71-72, 85, 87, 89, 141
- educational considerations, 70
- federal policies, 48-49, 60-61, 62, 68-69, 74, 78-79, 82-83, 91-92, 116-117
- geographic data and services, 75-76, 133-134, 135-136, 138
- institutional and research sponsor responsibilities, 7, 87, 90-91, 116-117, 119
- and integrity of data, 5, 26, 42, 46-47, 63, 64, 70, 81-82, 89, 137
- international dimensions, 17, 35, 64, 66, 70, 75, 76-77, 79, 83-84
- journal policies, 21, 38, 43, 61, 65, 78-79, 82-83, 87, 89, 90, 91, 92, 93, 116-117, 135
- legal issues, 67, 71, 75, 80-82, 85, 141
- metadata and, 25, 26, 70, 85
- methods or programs used to derive data, 6, 63-64, 81, 85, 89
- national security issues, 5, 68-69, 83, 134, 136
- OECD principles and guidelines, 59, 62
- ownership issues, 5-6, 21, 73-79, 85-86, 134, 135-136, 137, 138
- Paris Guidelines, 88, 89
- principles for enhancing access, 6, 82, 84-86, 87, 88-90, 120
- private/commercial interests, 6, 19, 26, 48, 69, 71-73, 86, 134-135
- professional organizations and, 91, 116-117
- public policy interests, 18, 71-73, 135, 141
- publicly funded research, 6, 74, 76-77, 78, 80, 82, 83, 84, 85, 92-93, 141
- raw data, 80
- recommendations, 6-7, 87, 88, 90, 91
- and reproducibility, 81, 93
- research field responsibilities, 87, 88-90, 136
- researcher responsibilities, 7, 85, 86-87, 116-117, 136
- Air Force Office of Scientific Research, 52-53
- Amazon, 106
- America COMPETES Act of 2007, 82
- American Association for the Advancement of Science, 144-145
- American Association of Universities, 118
- American Chemical Society, 39
- American Economic Review*, 39, 65
- American Geophysical Union, 38, 39

Andrew W. Mellon Foundation, 101, 107, 112
Annals of Internal Medicine, 93
Arabidopsis Information Resource, 30
Arts and Humanities Data Service (UK), 66
arXiv publication repository, 102
Association of American Universities, 91
Association of Public and Land-Grant Universities, 91
Association of Research Libraries, 91
Astrobiology, 28
Astronomical sciences, 14-15, 16, 17, 64, 66, 95, 141
Automatic Plate Measuring Facility, 14

B

Bayh-Dole Act of 1980, 76
Bell Laboratories, 45
Biomedical Informatics Grid, 61
Biomedical Informatics Research Network, 20, 61
Biomedical research, 68, 86, 93. *See also*
Clinical
Bit Torrent, 89
Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 107, 113
Boeing, 105
Boomerang (balloon-borne millimeter-wave telescope), 14

C

Canadian Department of National Defence, 105
CC0 protocol, 76
Cell Centered Database, 20
Center for Astrophysics, 14
CERN (European Center for Nuclear Research), 13, 16, 34
Chaotic motion, 41
Charge-coupled devices, 11
Chemical crystallography, 66
China, Scientific Data Sharing Project, 84
Climate sciences, 66, 71-72, 82, 96, 134
Clinical research, 2, 4, 47, 48-49, 141. *See also*
Biomedical
Cloud computing, 61
Collaborative
data-sharing networks, 15, 17-18, 64, 84, 111

research, 17-18, 20, 35, 63, 64, 66, 96, 101
stewardship, 103, 113
Committee on Data for Science and Technology, 84
Committee on Government Relations, 118
Committee on Institutional Cooperation, 118
Condensed-matter physics, 45
Confidentiality or privacy considerations, 5, 67, 140
Consolidated Appropriations Act
of 2001, 82
of 2008, 79
Copyrights, 73, 74-75, 76, 78, 79, 85, 100, 140-141
Cornell University, 102
Cosmic Background Explorer, 14
Cosmic Background Imager, 14
Council of Graduate Schools, 55
Council of Science Editors, 37
Council on Government Relations, 91
Creative Commons, 76
Cyberinfrastructure
and access to data, 61-62, 83-84, 87, 135
and stewardship, 9, 95, 96, 99-106, 113

D

Data, defined, 23
Data Access Act of 1999, 80-81, 92
Data collections. *See* Databases
Data Explorer program, 83
Data management. *See also* Stewardship
institutional plans, 92-93, 98, 110
principles and guidelines, 88-90, 133-134
training in, 56
Data mining tools, 20-21
Data producers. *See also* Researchers
defined, 40
integrity-related responsibilities, 40, 42
Data professionals
and access to data, 92
and integrity of data, 5, 50, 57-58
researcher collaborations with, 57-58
responsibilities, 3, 50, 58
Data providers
defined, 41
integrity-related responsibilities, 41, 42
Data users
defined, 41
and integrity of research results, 41, 42

Databases and repositories. *See also*
Stewardship; *individual databases*
access to, 73, 83
defined, 29
developmental questions about, 21
disciplinary depositories, 27, 100, 102
integrity of data, 21, 46, 50
ownership issues, 73-74, 140-141
reference collections, 29, 30
research collections, 29, 30
resource collections, 29, 30
for supporting data, 89
tools for building, 100-101

Digital data. *See also* Research data; *individual disciplines*
ownership, 73
projected growth, 12, 13
quantity, 1, 11-12, 13, 29, 84-85
risks to reliability, 96-99
transfer rates, 11
trends, 14-15
units of size, 11, 12

Digital Millennium Copyright Act of 1998,
75, 140

Digital networks, 83-84

Digital technology
in astronomy, 14-15
challenges posed by, 1-2, 19, 22, 74-75
computing power, 11-12, 20
and copyrights, 74-75
education impacts, 18
and integrity of data, 1, 3-4, 21, 33-34, 37,
44, 46-50
in neurosciences, 20-21
public policy implications, 18
research impacts, 1, 12-18, 20-21, 34-35,
44, 51, 55-56, 57-58, 85, 139, 140
sensors and sensor networks, 11, 16-17
simulations and mathematical modeling, 1,
17, 20
and stewardship, 1-2, 19, 22, 27
storage devices, 11

Diversity of data
in collection types, 29, 30
cross-disciplinary, 27-28
in origins, 28-29

DNA. *See also* Genomic data
“junk,” 41

E

Earth Observing System, 17
Ecological Society of America, 39, 111, 112,
113
Economic issues, 21, 22, 26, 43, 44, 59, 69-70,
71-72, 85, 87, 89, 98-99, 105-106, 113
Economics research, 47, 50, 64, 65, 105-106
Educational considerations, 70
Employment Responses to Global Markets
data, 105-106
Environmental sciences, 17, 104, 133-134,
137, 139
European Community Directive on the Legal
Protection of Databases, 75
European Southern Observatory, 64
Experimental data, 1, 13, 16, 17, 20, 22, 23,
26, 27, 29, 35, 40, 42, 45, 64, 67, 109,
136
Extramural grants, federal policies on, 52-53

F

Federal policies
on access, 48-49, 60-61, 62, 68-69, 74, 78-
79, 82-83, 91-92
on integrity, 52-53
on stewardship, 79, 104-106, 113, 133-134,
137
Federation of American Societies for
Experimental Biology, 38, 39
Fischer, Paul, 71-72
Fluxes Over Snow Surfaces Project, 30
Food and Drug Administration, 48
Fostering Integrity in Research, Scholarship,
and Teaching, 56
Freedom of Information Act, 80

G

Galaxy Redshift Surveys, 14
Galaxy Surveys, 14
GenBank, 3, 46, 47, 52, 83, 104
Gene expression data, 20
General Accounting Office, 81
Genomic data, 16, 20, 60-61, 69, 77-78, 95,
96, 134, 136
Geographic data and services, 75-76, 133-134,
135-136
Georgia Open Records Act, 71
Geosciences, 98, 138

GlaxoSmithKline Clinical Trial Register, 48
Global Biodiversity Information Facility, 84
Global Earth Observation System of Systems, 84
Google, 106
Government Accountability Office, 82
Government-Industry Data Exchange Program, 105
Grid computing, 61

H

Health Insurance Portability and Accountability Act, 68, 110
Health records data, 49, 68
High-energy physics. *See* Particle physics
Howard Hughes Medical Institute, 77
Human Brain Dynamics Resource data, 105
Humanities research, 66
Hwang, Woo Suk, 44

I

IDEALS repository, 102-103
Information Quality Act of 2001, 80, 81, 82, 137
Institute of Electrical and Electronics Engineers, 39
Institute of Medicine, 55
Integrity of research data
accessibility and, 5, 26, 42, 46-47, 63, 64, 70, 81-82, 137
in clinical research, 2, 4, 47, 48-49, 141
collective scrutiny of data and results, 41-43, 47
contextual documentation, 3, 42-43, 45, 46-47, 63-64
data professionals' responsibilities and roles, 5, 50, 57-58, 116
decentralized approach, 57
defined, 2, 25-26
digital technology to enhance reliability, 1, 3-4, 21, 33-34, 37, 44, 46-50
economic issues, 43, 44
economics research, 47, 50
federal policies, 52-53, 116, 137
inappropriate manipulation, 3, 5, 35-36, 37, 38-39, 40, 44
journal policies, 3, 5, 35-37, 38, 39, 40, 43, 116

metadata and, 42, 46, 50
open and public reviews, 46
in particle physics, 34-35, 47
peer review, 2, 3, 7, 22 n.17, 24, 33, 35, 39, 42, 43-44, 46-47, 52, 55, 79, 133, 137
principle, 4, 51, 120
producer's role, 40, 42, 116
provider's role, 41, 42, 116
quality control measures, 35, 44-50, 138
recommendations, 4-5, 54, 57, 58
and reproducibility of research results, 2, 26, 29, 33, 45
researchers' obligations, 26, 39-40, 45, 51-54, 116, 119, 133
standards, 25, 35-36, 44, 45, 50, 54-55, 56-57, 138, 142
threats to, 2-3, 19, 33-34, 39-40, 96-99
training for researchers, 4-5, 44, 54-56
user's role, 41, 42, 116
Interagency Working Group on Digital Data, 9, 107-108
Interdisciplinary research, 1, 6-7, 9, 17, 28, 44, 57, 60, 70, 141
Intergovernmental Panel on Climate Change, 72
International Council of Scientific Unions, 84
International dimensions of access, 17, 35, 64, 66, 70, 75, 76-77, 79, 83-84
International Federation of Digital Seismograph Networks, 84
International Geophysical Year, 83
International Nucleotide Sequence Database Collaboration, 84
International Virtual Observatory Alliance, 84
Inter-University Consortium for Political and Social Research, 99-101, 102, 113
Intramural research, federal policies, 52-53
Invisible colleges, 28
Ithaka, 112

J

Journal of Cell Biology, 35-37, 39, 40, 150-153
Journals
access policies, 21, 38, 43, 61, 65, 78-79, 82-83, 87, 89, 90, 91, 92, 93, 135
confirmatory studies in, 65
copyrights, 78-79
data manipulation policies, 3, 5, 35-37, 38, 39, 40, 43
as data providers, 41

ethics and scientific misconduct policies,
38
integrity policies, 35-37, 38, 39, 40, 43, 56,
65
letters requesting this study, 143-153
open access, 46 n.16, 78-79, 83
peer-reviewed, 43
responsibilities of, 119-120
stewardship role, 21, 106, 113, 117
JSTOR, 112

L

Laboratory Management Institute, 55
Large Hadron Collider, 13, 16, 34-35
Large Synoptic Survey Telescope, 13, 14, 15
Las Campanas Redshift Survey, 14
Legal issues, 67, 71, 75, 80-82, 85
Library of Congress, 112, 113
Licensing, 75-76, 77-78, 135-136
Lick Observatory, 14
Life sciences, 23 n.18, 59, 60-61, 64, 66, 69,
77-78, 88, 95-96, 134, 136, 141
Lockheed Martin, 105

M

Manipulation of data, inappropriate, 3, 5, 35-
36, 37, 38-39, 40, 44
Mann, Michael, 71-72
Marine fisheries data, 139
Medical College of Georgia, 71
Merck, 86
Metadata
access issues, 25, 26, 70, 85
defined, 24
importance, 42
and integrity of research data, 42, 46, 50
providers, 21
standards, 24
stewardship, 95, 110
Misconduct. *See* Research misconduct
Molecular and Cellular Proteomics, 89
Molecular Dynamics Simulation Data, 105
Moore, Gordon, 11
Moore's law, 11, 96
Multidisciplinary research, 28
myGrid, 61

N

National Aeronautics and Space
Administration, 17, 52-53, 64, 105, 138
National Archives and Records
Administration, 135
National Association of State Universities and
Land-Grant Colleges, 118
National Center for Biotechnology
Information, 95-96, 104
National Commons and Marketplace in
Geographic Information, 76
National Ecological Data Center, 112
National Ecological Observatory Network,
16-17
National Institute for Technology and Liberal
Education, 103
National Institute of Standards and
Technology, 52-53
National Institutes of Health, 52-53, 54-55, 78,
79, 82-83, 90-91, 99, 100, 102
National Library of Medicine, 79, 83, 104
National Oceanographic and Atmospheric
Administration, 104, 133-134
National Optical Astronomy Observatory, 64
National Postdoctoral Association, 55
National Research Council, 9, 59-60, 64, 67-
69, 72, 75-76, 77-78, 80-81, 82, 98, 104,
108, 111
National Science and Technology Council,
107-108
National Science Board, 29, 30
National Science Foundation, 16-17, 24, 52,
82, 99, 100, 102, 104, 107
National Security Decision Directive 189,
68-69
National security issues, 5, 68-69, 83, 134, 136
National Virtual Observatory, 15, 105, 138
Natural Environment Research Council (UK),
66
Nature, 37, 38, 46, 109-110
Network for Earthquake Engineering
Simulation, 16-17
Neurosciences, 13, 20-21, 99, 105
New England Journal of Medicine, 39

O

Observational data, 2, 17, 22, 23, 26, 28, 33,
40, 42, 48, 64, 67, 106, 109

Office of Management and Budget, 22 n.17, 52, 69, 80, 81, 82
Office of Naval Research, 52-53
Office of Science and Technology Policy, 36-37
Omnibus Appropriations Act of 2009, 79
Open-access
 academic policies, 91
 corporate/private platforms, 86
 journals, 78-79, 91
 public policies, 82-83, 85, 91
 repositories, 83
Open-knowledge environments, 62-63
Open-notebook science, 63
Open reviews, 46
Open-source software, 62-63
Organisation for Economic Co-operation and Development (OECD), 59, 62, 84
Ownership issues
 academic research, 69, 77
 copyrights, 73, 74-75, 76, 78, 79, 85, 100, 140-141
 database protections, 73-74, 75, 76, 140-141
 digital technologies and, 74-75
 fair use exceptions, 74
 journals, 78-79
 licensing, 75-76, 77-78, 135-136
 patents, 6, 25, 69, 76-78, 85-86, 134
 publicly funded research, 74, 76-77, 78
 trade secrecy, 77

P

Palimpsest, 106
Paris Guidelines, 88, 89
Particle physics, 12-13, 34-35, 42, 47, 64, 131
Patents, 6, 25, 69, 76-78, 85-86, 134
Peer review, 2, 3, 7, 22 n.17, 24, 33, 35, 39, 42, 43-44, 46-47, 52, 55, 79, 133, 137
Pennsylvania State University, 71
Physics, 27, 45, 102. *See also* Particle physics
Planck survey, 14
Plate tectonics, 41
Portico, 106
Principles
 access to data, 6, 82, 120
 integrity of data, 4, 51, 120
 stewardship, 8, 109, 120
Private/commercial interests, 6, 19, 26, 48, 69, 71-73, 86, 106, 134-135

Proceedings of the National Academy of Sciences, 37, 38
Processed data
 defined, 25
 integrity, 1, 25, 33-34
Professional organizations, 91, 111-112, 117, 119-120. *See also individual organizations*
Protein Data Bank, 30, 105
Proteomics research, 77-78, 89, 134
Public Library of Science (PLoS), 78-79
Public policies on access, 52-53, 66, 78-79, 82-83, 133-134
Public policy interests, 18, 71-73, 135
Publicly funded research, 6, 74, 76-77, 78, 80, 82, 83, 84, 85, 92-93
PubMed, 60

Q

Quality control and quality assurance, 35, 44-50. *See also* Integrity of research data

R

Raw data
 access to, 15, 74, 80
 defined, 25
 stewardship, 25, 45, 106
Raytheon, 105
Reagan, Ronald, 68
Recommendations
 access to data, 6-7, 87, 88, 90, 91
 integrity of research data, 4-5, 54, 57, 58
 stewardship, 8-9, 110, 111, 113
Remote sensing data, 40, 75, 138, 139
Reproducibility of research results
 access to data and, 81, 93
 defined, 81
 integrity of data and, 2, 26, 29, 33, 45
Research Councils of the United Kingdom, 84
Research data. *See also* Digital data; *individual disciplines*
 definitions, 22-24
Research Information Network (UK), 64, 66
Research institutions and sponsors,
 responsibilities of, 7, 9, 56-57, 87, 90-91, 100-103, 110, 112-113, 117, 118-119, 139

- Research misconduct
 - access to supporting data and, 63
 - defined, 36-37
 - examples, 44, 45
 - training, 52
 - Researchers
 - access-related responsibilities, 7, 85, 86-87, 116-117, 118, 136
 - collaborations with data professionals, 57-58
 - integrity-related obligations, 26, 39-40, 45, 51-54, 116, 133
 - professional standards, 54, 56-57
 - stewardship role, 8-9, 45, 99-100, 109-112, 117
 - training in conduct of research, 4-5, 44, 54-56, 118
 - R.J. Reynolds Co., 71
 - Rockefeller University Press, 35, 38
 - Roles and responsibilities
 - assigning, 115
 - journals, 119-120
 - professional societies, 119-120
 - research institutions, 118-119
 - research sponsors, 119
 - researchers, 115-118
 - Rural Economy and Land Use Program (UK), 66
- S**
- Sage, 86
 - San Diego Supercomputer Center, 105
 - Schön, Jan Hendrik, 45
 - Science*, 38, 44, 144-149
 - Science Commons, 76
 - Seoul National University, 44
 - Sharing data. *See also* Access; Databases, Open
 - in astronomy, 15, 64, 66
 - barriers to, 1, 5-7, 19, 26, 28, 60, 63-70, 76, 85, 88, 135, 137, 139, 141, 142
 - benefits, 5, 20, 59-62, 70, 142
 - in biomedical research, 68, 86, 93
 - in chemical crystallography, 66
 - in climate sciences, 66, 71-72, 82, 134
 - collaborative efforts, 15, 17-18, 64, 84, 111
 - contextual documentation, 65, 66, 67
 - in economics, 64, 65
 - in humanities, 66
 - incentives for, 90-91
 - in life sciences, 23 n.18, 59, 60-61, 64, 66, 69, 77-78, 88, 134, 136, 141
 - norm differences of research fields, 6-7, 63-64, 66, 88, 90
 - in particle physics, 64
 - protocols and standards, 7, 83, 87, 88-90, 92, 139
 - public policies, 52-53, 66, 78-79, 82-83, 133-134
 - in social sciences, 65, 66, 68, 100-101, 135
 - UPSIDE principle, 64, 67-68, 86
 - Shelby, Richard, 80
 - Simulations and mathematical modeling, 1, 17, 20, 22, 23, 28, 40, 42, 46, 61, 105, 106
 - SkyServer Web site, 14
 - Sloan Digital Sky Survey, 3, 14-15, 46, 47, 95, 105
 - Social network modeling, 17
 - Social sciences, 22, 27, 65, 66, 68, 98, 100-101, 135, 142
 - Standards
 - for access, 7, 83, 87, 88-90, 92
 - for integrity of research data, 25, 35-36, 44, 45, 50, 54-55, 56-57
 - for stewardship, 8-9, 111
 - Stewardship of research data. *See also*
 - Databases
 - annotating data for long-term use, 8, 21, 22, 25, 95, 99, 106-107, 110
 - in astronomy, 95
 - for broad research enterprise, 107-108
 - challenges, 8, 103
 - collaborations, 103, 113
 - companies, 106
 - defined, 27, 95
 - disciplinary depositories, 27, 100, 102
 - economic issues, 7-8, 21, 22, 98-99, 105-106, 113, 119
 - ESA initiative, 112
 - exchanges of data, 105
 - federal agencies, data centers, and digital archives, 79, 104-106, 113, 117, 133-134, 135, 137
 - in geosciences, 98, 138
 - infrastructure and incentives, 9, 95, 96, 99-106, 113
 - institutions and research sponsors, 9, 100-103, 110, 112-113, 117, 119, 139
 - international participation, 111
 - journals, 21, 106, 113, 117
 - in life sciences, 95-96
 - loss and underutilization problems, 96-99

metadata, 95, 110
old vs. new data, 98, 138
oversight board, 111
ownership issues, 9, 99
principle for enhancing, 8, 109, 120
professional societies and, 111-112, 117
raw data, 25, 45, 106
recommendations, 8-9, 110, 111, 113
researchers and, 8-9, 45, 99-100, 109-112, 117
simulation-related models and software tools, 28
in social sciences, 98, 100-101
standards development, 8-9, 111
technology changes and, 1-2, 19, 22, 27
training in, 119

SumsDB, 20-21

Sustainable Digital Data Preservation and Access Network (DataNet) program, 104-105

T

Terminology

accessibility, 26
integrity, 25-26
metadata, 24
processed data, 25
raw data, 25
research data, 22-24
standardization of, 49
stewardship, 27

Trace Archive, 95, 96

Training for researchers, 54-56

Tranche, 89

Two-degree-Field (2dF) Galaxy Redshift Survey, 14

U

University of California at Davis, 55

University of California at San Diego, 105

University of Illinois at Urbana-Champaign, 102-103

University of Konstanz, 45

University of Minnesota, 56

Utility of research data, defined, 27. *See also* Access; Integrity; Stewardship

U.S. Department of Agriculture, 52-53

U.S. Department of Commerce, 52-53

U.S. Department of Energy, 52-53, 83, 105

U.S. Department of Health and Human Services, 52, 55

U.S. Department of Veterans' Affairs, 49

U.S. Environmental Protection Agency, 52-53, 133

U.S. Patent and Trademark Office, 78

UPSIDE principle, 64, 67-68, 86

V

Virtual team science, 61

Visible and Infrared Telescope for Astronomy, 14

W

WebCaret, 20-21

Wellcome Trust, 61, 83

Wikis, 46, 63

Wilkinson Microwave Anisotropy Probe, 14

World Data Center system, 83

World Intellectual Property Organization treaty, 75

Y

Yale University, 150-153