Open Science by Design: Realizing a Vision for 21st Century Research (2018)

DETAILS

232 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-47624-9 | DOI 10.17226/25116

CONTRIBUTORS

Committee on Toward an Open Science Enterprise; Board on Research Data and Information; Policy and Global Affairs; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25116.

OPEN SCIENCE BY DESIGN

Realizing a Vision for 21st Century Research

Committee on Toward an Open Science Enterprise

Board on Research Data and Information

Policy and Global Affairs

A Consensus Study Report of

The National Academies of

SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS

Washington, DC

www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by the Laura and John Arnold Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-47624-9
International Standard Book Number-10: 0-309-47624-0
Library of Congress Control Number: 2018950760
Digital Object Identifier: https://doi.org/10.17226/25116

Additional copies of this publication are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25116.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

COMMITTEE ON TOWARD AN OPEN SCIENCE ENTERPRISE

Alexa T. McCray (NAM) *(Chair)*, Professor of Medicine, Harvard Medical School
Francine Berman, Edward P. Hamilton Distinguished Professor of Computer Science, Rensselaer Polytechnic Institute
Michael Carroll, Professor of Law, American University Washington College of Law
Donna Ginther, Professor, Department of Economics; Director, Center for Science, Technology and Economic Policy, University of Kansas
Robert Miller, Chief Executive Officer, LYRASIS
Peter Schiffer, Vice Provost for Research and Professor in Applied Physics, Yale University
Edward Seidel, Vice President for Economic Development and Innovation, University of Illinois System; Founder Professor of Physics, Professor of Astronomy and Computer Science, University of Illinois at Urbana-Champaign
Alex Szalay, Bloomberg Distinguished Professor of Astronomy, The Johns Hopkins University
Lisa Tauxe (NAS), Distinguished Professor of Geophysics, Scripps Institution of Oceanography, University of California, San Diego
Heng Xu, Associate Professor of Information Sciences and Technology, College of Information Sciences and Technology, The Pennsylvania State University

Principal Project Staff

Tom Arrison, Program Director, Policy and Global Affairs Division (from November 2017)
Emi Kameyama, Associate Program Officer, Board on Research Data and Information
George Strawn, Director, Board on Research Data and Information
Ester Sztein, Deputy Director, Board on Research Data and Information
Nicole Lehmer, Senior Program Assistant, Board on Research Data and Information
Alan Anderson, Consultant
Christine Liu, Senior Program Officer (until October 2017)

*Denotes ex-officio member.

Preface

Just as society has been transformed by the digital revolution, so, too, have many aspects of the scientific enterprise. Publicly available data in federally sponsored databases serve as starting points for many research investigations. Collaborations are no longer hampered by geographic distance, and, in some cases, the majority or even all of the work is conducted by sharing digital research files, corresponding by email, and meeting virtually, with time zone differences being the only deterrent to the frequency of the meetings. Data are largely collected, stored, manipulated, and shared in electronic form. Research papers are prepared using word-processing software and are often formatted and submitted in camera-ready form to the publisher. The majority of published articles are no longer bound in print journals and disseminated by conventional postal delivery, but rather are available through the publisher's database, most often mediated by contracts with institutional libraries.

This transformation has had economic, policy, and practical implications, many of which are still in the process of being fully addressed and resolved. An increasing number of scientists have begun to question the closed world of scientific publishing and have suggested that the results of their research should be openly available for all, to benefit not only fellow scientists, but also the general citizenry. Indeed, the pursuit of "citizen science" is now recognized as a valid and useful activity. Faculty at many universities have adopted university-wide "open access" policies that ensure that, at a minimum, their research papers are available through their institution's repository.

New publishing venues have arisen, including open access journals, some of very high-quality and others not. Individual researchers, while interested in having their work broadly read and cited, are faced with competing pressures, including publishing in journals with high "impact factors," such that they are in the best possible position for promotion and tenure.

Research funders have seen the value of openly sharing the results of the research that they have supported, not just in the form of publications, but also in the form of the data that have been produced in the course of the investigation. They have begun to require that applicants prepare data management plans as part of their grant proposals.

A number of legal and policy developments have facilitated broader access to scientific research. Recognizing the potential of the Internet to broadly and equitably disseminate scientific knowledge, a collaborative effort has created a legal framework that is consistent with U.S. copyright law, and that provides guidance to researchers who would like to have greater control over how their research results are used and disseminated. Several federal policies require that publicly

funded research results, in the form of data and publications, be deposited in public access repositories. Legislation is now also pending in Congress that would strengthen these policies.

To evaluate more fully the benefits and challenges of broadening access to the results of scientific research, described as "open science," the National Academies of Sciences, Engineering, and Medicine appointed an expert committee in March 2017. Brief biographies of the individual committee members are provided in Appendix A. The committee was charged with focusing on how to move toward open science as the default for scientific research results, and to indicate both the benefits of moving toward open science and the barriers to doing so. This report presents the findings and recommendations of the committee, with the majority of the focus on solutions that move the research enterprise toward open science.

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report: Prudence Adler, Association of Research Libraries; David Allison, Indiana University, Bloomington; Geoffrey Boulton, University of Edinburgh; Anita de Waard, Elsevier; Michael Forster, Institute of Electrical and Electronics Engineers; Laura Greene, Florida State University; Heather Joseph, Scholarly Publishing and Academic Resources Coalition; Véronique Kiermer, Public Library of Science; Michael Lesk, Rutgers University; William Mobley, University of California, San Diego; Mark Musen, Stanford University; Sarah Nusser, Iowa State University; and George Schatz, Journal of Physical Chemistry.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations of this report nor did they see the final draft before its release. The review of this report was overseen by Carl Lineberger, University of Colorado, Boulder and Julia Phillips, Sandia National Laboratories (Retired). They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring committee and the National Academies.

The report would not have been possible without the sponsor of this study, the Laura and John Arnold Foundation, whom we thank for their support. The committee gratefully acknowledges all of the speakers for their informative presentations at our meeting and public symposium. They are listed in Appendix E at the conclusion of the report. The information provided during the meeting and symposium is used throughout this report and provided important perspectives that were utilized in this report's findings and conclusions.

The committee is also grateful for the assistance of the National Academies staff in preparing this report. Staff members who contributed to this effort are Tom Arrison, program director, Policy and Global Affairs; Emi Kameyama, associate program officer, Board on Research Data and Information; George Strawn, director, Board on Research Data and Information; Ester Sztein, deputy director, Board on Research Data and Information; Nicole Lehmer, senior program assistant, Board on Research Data and Information; Alan Anderson, consultant; Christine Liu, senior program officer (through October 2017); Adriana Courembis, financial officer; Marilyn Baker, director for reports and communication; and John Boright, interim executive director, Policy and Global Affairs.

Finally, I thank especially the members of the committee for their tireless efforts throughout the development of this report.

Alexa T. McCray, Chair
Committee on Toward an Open Science Enterprise

Contents

Abbreviations and Acronyms

AARNET	Australia's Academic and Research Network
AAU	American Association of Universities
ACS	American Chemical Society
AGU	American Geophysical Union
ALPSP	Association of Learned and Professional Society Publishers
ANDS	Australian National Data Service
APC	Article Processing Charges
APHIS	Animal and Plant Health Inspection Service
API	Application Programming Interface
APLU	Association of Public and Land-Grant Universities
APO	Apache Point Observatory
APOGEE	Apache Point Observatory Galactic Evolution Experiment
ARC	Astrophysical Research Consortium
ARS	Agricultural Research Service (U.S. Department of Agriculture)
ARXIV	Archive
AWSI	Amazon Web Services
BIA	Bureau of Indian Affairs
BIORXIV	Bio Archive
BLM	Bureau of Land Management
BOAI	Budapest Open Access Initiative
BOSS	Baryon Oscillation Spectroscopic Survey
BRDI	Board on Research Data and Information
CADRE	Center for the Advancement of Data and Research in Economics
CAPTCHA	Completely Automated Public Turing Test to Tell Computers and Humans Apart
CC	Creative Commons
CC BY	Creative Commons Attribution
CC BY-NC	Creative Commons Attribution-Noncommercial
CC BY-ND	Creative Commons Attribution-Noderivs License
CC BY-SA	Creative Commons Attribution-Sharealike
CC0	Cc Zero
CDL	California Digital Library
CERN	Conseil Européen Pour La Recherche Nucléaire (European Organization for Nuclear Research)
CHORUS	Clearinghouse for the Open Research of the United States
COAR	Confederation of Open Access Repositories
CODATA	Committee on Data for Science and Technology
COPE	Committee on Publication Ethics
COS	Center for Open Science

CRISPR/CAS9	Clustered Regularly Interspaced Short Palindromic Repeats/Crispr Associated Protein 9
CSIRO	Commonwealth Scientific and Industrial Research Organization
DASH	Digital Access to Scholarship at Harvard
DDI	Design, Development, and Implementation
DFIG	Data Fabric Interest Group
DMP	Data Management Platform
DO	Digital Object Architecture
DOAJ	Directory of Open Access Journals
DOE	U.S. Department of Energy
DOI	U.S. Department of Interior
DOIS	Digital Object Identifiers
DVN	Dataverse Network
EC	European Commission
EMBL-EBI	The European Molecular Biology Laboratory-The European Bioinformatics Institute
EOSC	European Open Science Cloud
EPA	U.S. Environmental Protection Agency
ESO	European Southern Observatory
ESSOAR	Earth and Space Science Open Archive
EU	European Union
EUA	European University Association
FAIR	Findable, Accessible, Interoperable, and Reusable
FASTR	Fair Access to Science and Technology Research
FDA	Food and Drug Administration
FDAAA	Food and Drug Administration Amendments Act
FOSTER	Facilitate open science training for European Research
FRED	Federal Reserve Economic Data Site
FSIS	Food Safety and Inspection Service
FWS	U.S. Fish and Wildlife Service
G7	Group of Seven
GCMS	Geologic Collections Management System
GNU	General Public License
GO FAIR	Global Open Findable, Accessible, Interoperable, and Reusable
HEP	High Energy Physics
HIPAA	Health Insurance Portability and Accountability Act
HOA	Hybrid Open Access
HOAP	Harvard Open Access Project
ICPSR	Inter-university Consortium for Political and Social Research
ICSU	International Council for Science
IDW	International Data Week
IEDA	Interdisciplinary Earth Data Alliance
IEEE	Institute of Electrical and Electronics Engineers
IGSN	International Geo Sample Number
IMDB	Internet Movie Database
IODP	International Ocean Discovery Program
IoT	Internet of Things

IP	Internet Protocol
IRIS	Incorporated Research Institutions for Seismology
IT	Information Technology
IQSS	Institute for Quantitative Social Science
IUPUI	Indiana University-Purdue University Indianapolis
IWGSC	Interagency Working Group on Scientific Collections
JIF	Journal Impact Factor
JISC	Joint Information Systems Committee
JSTOR	Journal Storage
LIBER	Library Federations
LOD	Linked Open Data
LSST	Large Synoptic Survey Telescope
MAGIC	Magnetics Information Consortium
MANGA	Mapping Nearby Galaxies at Apo
MDPI	Multidisciplinary Digital Publishing Institute
MEDOANET	Mediterranean Open Access Network
MIT	Massachusetts Institute of Technology
MPDL	Max Planck Digital Library
NASA	National Aeronautics and Space Administration
NASEM	National Academies of Science, Engineering, and Medicine
NBER	National Bureau of Economic Research
NDS	National Data Service
NIFA	National Institute of Food and Agriculture
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NLM	National Library of Medicine
NOAA	National Oceanic and Atmospheric Administration
NPS	National Park Service
NRC	National Research Council
NRCS	Natural Resources Conservation Service
NSF	National Science Foundation
NSFNET	National Science Foundation Network
NSTC	National Science and Technology Council
NUTRIXIV	Nutritional Sciences Archive
OA	Open Access
OAD	Open Access Directory
OASPA	Open Access Scholarly Publishers Association
OECD	Organization for Economic Co-operation and Development
OPENAIRE	Open Access Infrastructure for Research in Europe
ORCID	Open Researcher and Contributor ID
ORFG	Open Research Funders Group
OSC	Open Science Collaboration
OSF	Open Science Framework
OSPP	Open Science Policy Platform
OSTP	Office of Science and Technology Policy
PCR	Polymerase Chain Reaction
PEERJ	Peer-Reviewed Journal

PII	Personally Identifiable Information
PLOS	Public Library of Science
PMC	PubMed Central
PNAS	Proceedings of the National Academy of Sciences
R&D	Research and Development
RCT	Randomized Controlled Trial
RDA	Research Data Alliance
RE3DATA	Registry of Research Data Repository
REPEC	Research Papers in Economics
RNA	Ribonucleic Acid
ROARMAP	Registry of Open Access Repository Mandates and Policies
SCADA	Supervisory Control and Data Acquisition
SCICOLL	Scientific Collections International
SCOAP3	Sponsoring Consortium for Open Access Publishing in Particle Physics
SDSS	Sloan Digital Sky Survey
SESAR	System for Earth Sample Registration
SHARE	SHared Access Research Ecosystem
SLAC	Stanford Linear Accelerator Center
SPARC	Scholarly Publishing and Academic Resources Coalition
STEM	Science, Technology, Engineering and Mathematics
TCP/IP	Transmission Control Protocol/ Internet Protocol
TOP	Transparency and Openness Promotion
UC	University of California
UK	United Kingdom
UNESCO	United Nations Education, Scientific and Cultural Organization
USDA	U.S. Department of Agriculture
USFS	U.S. Forest Service
USFSC	U.S. Federal Scientific Collections
USGS	U.S. Geological Survey
UUID	Universally Unique Identifier
WAME	World Association of Medical Editors
WDS	World Data System

Summary

Openness and sharing of information are fundamental to the progress of science and to the effective functioning of the research enterprise. The advent of scientific journals in the 17th century helped power the Scientific Revolution by allowing researchers to communicate across time and space, using the technologies of that era to generate reliable knowledge more quickly and efficiently. Harnessing today's stunning, ongoing advances in information technologies, the global research enterprise and its stakeholders are moving toward a new *open science* ecosystem. Open science aims to ensure the free availability and usability of scholarly publications, the data that result from scholarly research, and the methodologies, including code or algorithms, that were used to generate those data.

BENEFITS AND MOTIVATIONS

The research enterprise has already made significant progress toward open science, and is realizing a number of benefits, with the expectation that these will expand in the future:

- **Rigor and reliability**. New standards for data and code sharing in fields such as biomedical research and psychology are making it easier for researchers to reproduce and replicate reported work, strengthening scientific rigor and reliability.
- **Ability to address new questions**. Open science allows researchers to bring data and perspectives from multiple fields to bear on their work, opening up new areas of inquiry and expanding the opportunities for interdisciplinary collaboration.
- **Faster and more inclusive dissemination of knowledge**. The proportion of scientific articles that are openly available is increasing, which accelerates the process of disseminating research and building on results. Open publication also allows broader, more inclusive participation in research and expands the possibilities of productive research collaboration within the United States and around the world.
- **Broader participation in research**. Large-scale projects in fields such as astronomy and ecology are utilizing open data and expanding opportunities for citizen scientists to contribute to scientific advances.
- **Effective use of resources**. Reuse of data in fields such as clinical research is facilitating the aggregation of multiple studies for meta-analysis and allows for more effective testing of new hypotheses.

- **Improved performance of research tasks**. New tools such as electronic lab notebooks enable more accurate recording of research workstreams and automate various data curation tasks.
- **Open publication for public benefit**. The belief that the broader public should have access to publicly-funded research and its benefits provides an additional strong rationale for open science. In the case of publicly-funded research, the ultimate sponsor is the taxpayer. The public benefits from open science as new knowledge is utilized more rapidly to improve health, protect environmental quality, and deliver new products and services.

BARRIERS AND LIMITATIONS

The benefits of open science are accruing to researchers themselves, research sponsors, research institutions, disciplines, and scholarly communicators. Yet despite the significant progress made in recent years toward creating an open science ecosystem, science today is not completely open. Most scientific articles are only available on a subscription basis. Sharing data, code, and other research products is becoming more common, but is still not routinely done across all disciplines. Several important barriers remain, as well as limitations on the extent and speed with which open science can be realized. These include:

- **Costs and infrastructure**. There are significant remaining cost barriers to widespread implementation of open publication and open data. New technological and institutional infrastructure within specific disciplines and across disciplines needs to be developed.
- **Structure of scholarly communications**. Most publications are still only available on a subscription basis, and some potential pathways to open publication may disrupt the current scholarly communications ecosystem, including scientific society publishers, or may disadvantage early career researchers, researchers working in the developing world, or those in institutions with fewer resources.
- **Lack of supportive culture, incentives and training**. Open practices such as preparing datasets and code for sharing and making preprints available are not generally rewarded and may even be discouraged by current incentive and reward systems. This may have the unintended consequence of causing a disadvantage to early career researchers.
- **Privacy, security, and proprietary barriers to sharing**. Sharing data, code, and other research products is becoming more common, but barriers related to ensuring patient confidentiality and the protection of national security information exist in some domains. Proprietary research also presents barriers. Ultimately, some parts of the research enterprise may not be open.

- **Disciplinary differences**. The nature of research and practices surrounding treatment of data and code differ by discipline and even within a discipline. The size of datasets and the nature of some data may prevent immediate, complete sharing. Safeguards to prevent misuse or misrepresentation of data will be needed.

ABOUT THE STUDY

In 2017, the National Academies of Sciences, Engineering, and Medicine launched a study aimed at overcoming barriers and moving toward open science as the default approach across the research enterprise. The Laura and John Arnold Foundation provided financial support for the study. The authoring committee, established under the Board on Research Data and Information, met in person four times and held several virtual meetings to gather information from experts and develop findings and recommendations. As part of its evidence-gathering process, the committee organized a 1-day public symposium in September 2017 to explore specific examples of open science and discussed a range of challenges focusing on stakeholder perspectives. The committee also reviewed a large body of written material on open science concerns, including literature that informed the committee on how specific solutions in policy, infrastructure, incentives, and requirements could facilitate open science. The committee was not asked to examine whether or not open science is good, but, rather, how to move it forward in ways that are beneficial to the scientific community. Also, issues related to the research use of data generated in other contexts (e.g. social media data) are not considered. The statement of task is available in Chapter 1.

AN INFLECTION POINT

The open science movement stands at an important inflection point. A new generation of information technology tools and services holds the potential of further revolutionizing scientific practice. For example, the ability to automate the process of searching and analyzing linked articles and data can reveal patterns that would escape human perception, making the process of generating and testing hypotheses faster and more efficient. These tools and services will have maximum impact when used within an open science ecosystem that spans institutional, national, and disciplinary boundaries.

At the same time, a number of organizations around the world are adopting new policies and launching new initiatives aimed at fostering open science. Public and private research funders such as the Bill & Melinda Gates Foundation, the European Commission (EC), and the Wellcome Trust have introduced mandates and support systems to ensure that the results of the research they support are open. Publishers are adopting openness frameworks and strengthening requirements to ensure that the data and methods underlying articles are available. In the United States, federal agencies have developed and implemented policies based on 2013 and 2014 memoranda from the White House's Office of Science and

Technology Policy aimed at increasing public access to the results of research funded by the federal government.

OPEN SCIENCE BY DESIGN

The central aim of this study is to provide guidance to the research enterprise and its stakeholders as they build strategies for achieving open science and take the next steps. In order to frame the issues and possible actions, the committee developed the concept of *open science by design*, defined as a set of principles and practices that fosters openness throughout the entire research life cycle (Figure S-1).

The researcher is at the center of the concept of open science by design. From the very beginning of the research process, the researcher both contributes to open science and takes advantage of the open science practices of other members of the research community. The overarching principle of open science by design is that research conducted openly and transparently leads to better science. The vision of open science by design suggests that all phases of the research process provide opportunities for assessing and improving the reliability and efficacy of scientific research. The concept visualized in Figure S-1 can be further described as follows:

- **Provocation: explore or mine open research resources and use open tools to network with colleagues.** Researchers have immediate access to the most recent publications and have the freedom to search archives of papers, including preprints, research software code, and other open publications, as well as databases of research results, all without charge or other barriers. Researchers use the latest database and text mining tools to explore these resources, to identify new concepts embedded in the research, and to identify where novel contributions can be made. Robust collaborative tools are available to network with colleagues.

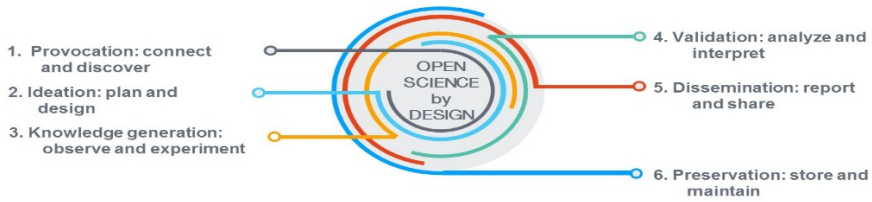


FIGURE S-1 Phases of Open Science by Design in the research life cycle. SOURCE: Committee generated.

- **Ideation: develop and revise research plans and prepare to share research results and tools under FAIR principles**. Researchers and their collaborators develop and revise their research plans, collect preliminary data from publicly available data repositories, and conduct a pilot study to test their new methods on the existing data. When applying for research funding, they develop the required data management plans, stating where data, workflow, and software code will be available for use by other researchers under FAIR (Findable-Accessible-Interoperable-Reusable) principles. In addition, in some cases, they may decide to pre-register their research plans and protocols in an open repository.
- **Knowledge generation: collect data, conduct research using tools compatible with open sharing, and use automated workflow tools to ensure accessibility of research outputs**. Researchers collect data, using tools that automate formatting and curation tasks to ensure that digital datasets are interoperable and documented. In the case of physical samples and specimens, such as rocks, ice core samples, or tissue samples, researchers develop concrete plans to archive these according to disciplinary best practices. With the availability of open software, the researcher can document approaches to cleaning and preparing data for analysis in an electronic research notebook.
- **Validation: prepare data and tools for reproducibility and reuse and participate in replication studies**. Researchers use open data techniques to analyze, interpret, and validate findings. They may present their preliminary findings at conferences and refine their methods based on relevant

comments and critiques. They may deposit their initial working paper in a preprint server and revise the paper based on the open peer review afforded by the service. They prepare their data in standard formats according to disciplinary standards and describe both data and analytical code in optimal ways for reuse and replication.

- **Dissemination: use appropriate licenses for sharing research outputs and report all results and supporting information (data, code, articles, etc.)**. Researchers select the best venue for open publication of their work, including articles, data, code, and other research products. They revise and, in some cases, substantially improve their work based on the comments of the peer reviewers. Upon acceptance and before final submission of their work, they select a public copyright license, such as the GNU General Public License for software or a Creative Commons license for other works, including scholarly articles.
- **Preservation: deposit research outputs in FAIR archives and ensure long-term access to research results**. Researchers deposit the final peer-reviewed articles in an openly accessible archive as required by their research funders. They deposit their research data and software in one or more data archives, with clear and persistent links among the article, data, and software. These FAIR data are then used by other researchers in the provocation phase of their own work.

The committee's concept of open science by design is by necessity general and idealized. Some discipline-specific nuances cannot be captured in such a broad concept. For example, there are fields where preregistration may not make sense or add value. Other challenges arise from the size or complexity of data. An important and emerging type of data are the very large datasets that capture extremely rare, time-sensitive events. Subtleties in this data and their generation may not be readily captured without detailed knowledge of how the data were collected.

Also, and importantly, open science by design is intended as a framework to empower the researcher. As expressed in other National Academies work, the principle for openness of data and other information underlying reported results is that they should be available no later than the time of publication, or when the researcher is seeking to gain credit for the work (NRC, 2003, 2009). For journal publication, any sharing prior to the point of final publication is up to the researcher, who is in full control of the decision of when to share. The committee believes that as open science by design becomes the norm, researchers will find that they benefit from sharing and collaborating early in the research process.

ACCELERATING PROGRESS

Achieving open science will require persistent, coordinated actions on the part of research enterprise stakeholders. The committee has developed findings,

recommendations, and implementation actions based on its review and synthesis of the information gathered throughout the course of the study. The complete set of findings is contained in Chapter 6 with the recommendations and implementation actions.

Building a Supportive Culture

The specific ways in which cultural barriers to open science operate vary significantly by field or discipline. Overuse and misuse of bibliographic metrics such as the Journal Impact Factor in the evaluation of research and researchers is one important "bug" in the operation of the research enterprise that has a detrimental effect across disciplines. The perception and/or reality that researchers need to publish in certain venues in order to secure funding and career advancement may lock researchers into traditional, closed mechanisms for reporting results and sharing research products. These pressures are particularly strong for early career researchers.

Initiatives such as the San Francisco Declaration on Research Assessment seek to achieve broad buy-in on the part of stakeholders to move toward evaluation systems that use other methodologies. Concrete actions, such as the National Institutes of Health (2017a) decision to encourage investigators to use and cite interim research products such as preprints in seeking funding, can have a beneficial effect.

Continued effort by stakeholders, working internationally and across disciplinary boundaries, is needed to change evaluation practices and introduce other incentives so that the cultural environment of research better supports and rewards open practices.

Recommendation One

Research institutions should work to create a culture that actively supports Open Science by Design by better rewarding and supporting researchers engaged in open science practices. Research funders should provide explicit and consistent support for practices and approaches that facilitate this shift in culture and incentives.

Implementation Actions

- Universities and other research institutions should explicitly reward the effort needed to make science open by design.
- Universities and other research institutions should partner with federal agencies in developing innovative approaches to assessing the impact of research in ways that include the impact of open science outputs. This should include, but is not limited to, the development of metrics for assessing the impact of interim research products such as preprints, with a

view toward comparing those with existing methods for measuring impact.

- Universities and other research institutions should move toward evaluating published data and other research products in addition to published articles as part of the promotion and tenure process. Archived data should be valued, just as the publications that result from them are valued.
- Researchers should make full use of the many opportunities that are available for making their research products openly available, and they should include that information in their curriculum vitae so that they can be appropriately credited and rewarded.
- In fields where this is not already common practice, research funders should encourage and reward the use of data and other research products that are available in publicly accessible databases.
- Universities and other research institutions should encourage and reward studies that focus on the replication and reproducibility of published research. Such studies should be published and made openly available.

Training for Open Science by Design

The report discusses several initiatives that emphasize training in open science and reproducibility. The emergence of data science as a recognized interdisciplinary field has highlighted the need for new educational content and approaches related to data (NASEM, 2018a).

Several federal agencies require that students or trainees supported by grants receive training in the responsible conduct of research, or RCR (NASEM, 2017b). Training and education that covers issues such as open science and reproducibility would complement the existing focus of RCR education and orient these programs toward supporting both research integrity and quality.

Recommendation Two

Research institutions and professional societies should train students and other researchers to implement open science practices effectively and should support the development of educational programs that foster Open Science by Design.

Implementation Actions

- Universities should provide training in best practices for open science and data stewardship as part of the regular curriculum in graduate and postgraduate education and should expect these practices as a default in all onboarding/orientation processes of universities, including new student orientation, new faculty orientation, library orientations, and lab training.

Course curricula should be developed and implemented to complement domain-specific courses that support open science by design.

- Research funders should support the development of training programs in the principles and practices of open science by design. Federal agencies should require this training as part of all federally funded graduate training grants (e.g., NSF research traineeships and NIH training grants) to foster open science by design.
- Library and information science schools, professional societies, and other interested organizations should develop course curricula and offer courses in the principles and practices of open science by design.
- Research funders and professional societies should create programs or contests that seek the creative and innovative integration and (re)use of open data for new and impactful research.
- The private sector and other interested parties should create innovative educational tools for open science principles and practices.

Ensuring Long-Term Preservation and Stewardship

The issues and challenges related to preservation and stewardship of research products, particularly data, code, and other non-article products, are considered in several places in the report. On the one hand, some of the technical and cost barriers to long-term data stewardship are falling, as tools for automated metadata tagging and classification become more widely used and data storage becomes cheaper over time. At the same time, the outputs of research continue to grow in volume and complexity, meaning that significant additional resources will still be required. For example, an important and emerging type of data are the very large datasets that capture extremely rare, time-sensitive events. Subtleties in these data and their generation may not be readily captured without detailed knowledge of how the data were collected.

Developing and sustaining the infrastructure required for long-term stewardship of research products will present a continuing challenge. This report does not contain a detailed cost estimate and timeline for meeting these needs. Yet several of the immediate priorities and initial steps do not, in themselves, require the expenditure of significant resources. Research communities can start by developing guidelines and criteria for determining what data and other research products should be preserved and for how long. Clearly, not everything needs to be preserved. Federal agencies that require data management plans in grant applications can better clarify guidance for compliance expectations and institutional responsibilities. The work of developing necessary standards and policies on the part of stakeholders will enable effective planning of new infrastructure and associated financing.

It is also important that approaches are flexible enough to adapt and change over time. The size and complexity of data in many fields are changing rapidly,

so that the solutions that are effective today might not be effective in a few years. At the same time, we have seen new tools and platforms continue to emerge that allow researchers to address challenges that were previously intractable.

Recommendation Three

Research funders and research institutions should develop the policies and procedures to identify the data, code, specimens, and other research products that should be preserved for long-term public availability, and they should provide the resources necessary for the long-term preservation and stewardship of those research products.

Implementation Actions

- Research institutions, professional societies and research funders should work together to develop selection guidelines and long-term stewardship best practices for the most valuable community datasets and other research products.
- Federal agencies should, consistent with the 2013 and 2014 Office of Science and Technology Policy (OSTP, 2013, 2014) memoranda for expanding public access to the results of federally funded research, continue to develop and standardize requirements for research products planning, management, reporting, and stewardship.
- Private research funders who have not already done so should adopt approaches compatible with those developed for publicly funded research products planning, management, reporting, and stewardship.
- Researchers should describe the plan for dissemination and stewardship of their research products with some specificity, consistent with the standardized sponsor requirements described above, including where their research products will be made publicly available and for what period of time.
- Research funders and research institutions should work together to resource and provide the infrastructure needed for long-term preservation, stewardship, and community control of research products. This infrastructure could be supported through direct costs or through an ear-marked percentage of each funded grant.

Facilitating Data Discovery, Reuse, and Reproducibility

As progress toward open science by design continues, it is important that the community adhere to the ultimate goal of achieving the availability of research products under open principles. Utilizing advanced machine learning tools in an-

alyzing datasets or literature, for example, will facilitate new insights and discoveries. Ensuring FAIR access should be a key consideration in deciding how to build repositories and other new resources.

As is the case with ensuring long-term stewardship, new standards should be developed by funders in collaboration with research institutions and researchers. Fields and disciplines that do not already have well-developed standards and practices for making research products available under FAIR principles will need time and help to create them. Where meeting new standards imposes costs, funders should make the necessary resources available, thereby avoiding the imposition of unfunded mandates. Specific actions enabling a transition need to be developed in a transparent manner, and avoid disrupting researchers and their work to the extent possible.

Recommendation Four

Funders that support the development of research archives should work to ensure that these are designed and implemented according to the FAIR data principles. Researchers should seek to ensure that their research products are made available according to the FAIR principles and state with specificity any exceptions based on legal and ethical considerations.

Implementation Actions

- Researchers should preferentially use open repositories that have been designed for interoperability and ease of discovery.
- Research funders should work to ensure that research products are available in repositories that allow for bulk transfer of digital objects to developers or users of automated discovery and analysis tools.
- Researchers and research funders should require that research products designated for long-term preservation and stewardship are assigned persistent unique digital identifiers.
- Professional societies and research funders should support efforts to network and federate existing repositories for improved discoverability.
- Research funders should continue to support the development of methods and tools that improve the interoperability of heterogeneous data. Metadata schemes, commonly accepted workflows for the processing and analysis of data, and other standards should be developed and used for improved data discovery.
- Research funders should commission an independent assessment of the state of university and federal data archives. The assessment should address how the FAIR principles have or have not been adhered to and make recommendations for improving accessibility to distributed or federated archives.

Developing New Approaches to Fostering Open Science by Design

There is a great deal of activity on the part of public and private research funders, research institutions, commercial and nonprofit publishers, community-organized groups and others aimed at preparing for and shaping a future research enterprise characterized by open science. Significant progress has been made, but a great deal of work needs to be done before open science by design is a reality. The committee focused on the choices facing U.S. organizations and institutions, realizing that the transition to open science by design is inherently a global process.

Effective dissemination will remain central to the advance of knowledge in the emerging open science era. Considerable resources are devoted to the publication of research results, much of them flowing to for-profit publishing companies or to nonprofit scientific societies. Many scientific societies generate surpluses through their publishing activities that support their professional ecosystems, and some would be severely challenged by some approaches to implementing open publication. At the same time, research institutions are currently experiencing difficulty in absorbing the steady increases in subscription rates of recent years.

Although scientific journals and articles will likely continue to play important roles for the foreseeable future, it is clear that the institutions and practices that support the dissemination of research will continue to evolve. Fully open publications are immediately accessible to all researchers at no cost and are available to all researchers under a copyright license that permits them to perform text and data mining or other productive reuses of the literature without the need for any negotiations or further permissions. While some subscription publishers have begun to offer researchers some forms of access for text and data mining and other productive reuses, their terms of access usually impose some restrictions on reuse.

The past several decades have seen the printed journal eclipsed by online distribution of research results. Datasets and other non-article research products will be increasingly valued and become a more significant focus of dissemination efforts. New venues for disseminating research have emerged and will continue to appear and grow.

The future evolution of research dissemination should be shaped by the changing needs of researchers and the broader enterprise, including the need to ensure openness. Issues of cost and sustainability should be considered from the standpoint of researchers. In developing new policies and support structures, research funders and research institutions should favor dissemination approaches that are responsive to community needs, and they should be transparent about their practices and costs.

Certain approaches to implementing open publication have the potential to affect the research ecosystem in significant ways, with differential impacts on different stakeholders. For example, a system that strongly favors publication approaches based on the payment of article processing charges would favor established researchers and wealthy institutions over early career researchers and

institutions with fewer resources. In planning new policies and transitions, it will be necessary to anticipate differential impacts to the extent possible, consider ways of avoiding these, and build in evaluative and corrective mechanisms to address unanticipated consequences.

Public and private funders have made significant contributions to fostering open science to this point. They should continue to support initiatives that accelerate progress, and evaluate and revise their policies as needed.

Recommendation Five

The research community should work together to realize Open Science by Design to advance science and help science better serve the needs of society.

Implementation Actions

- The federal government should revisit and update its open science policy, which is expressed in the 2013 and 2014 OSTP memoranda.
- Funders, institutions, and researchers should align policies and incentives to realize open publication, including rights-retention provisions.
- Research funders should support the establishment of a consortium of research community stakeholders to develop additional concrete methods for implementing open science by design.
- Professional societies—individually and collectively—should work to transition from current publication strategies to new ones that foster open science by design.
- Journal editors should work with publishers to transition from current business models to new ones that foster open science by design.
- Research funders should explore innovative means to support the transition from subscription-based systems to new publication strategies that enable open science by design.
- Librarians should work together with other members of the research community to promote and implement open science by design.
- The research community should develop tools and other applications that depend on the long-term availability of open research products, thereby providing new sources of revenue for the private sector, enhancing the value of research products, and leading to an acceleration of scientific progress.

1

Introduction

Organized science has always relied on the willingness of researchers to share their results, allowing others to test and build on their work. According to the Royal Society (2012), "open communication and deliberation sit at the heart of scientific practice." The digital revolution of the past several decades has greatly expanded the scope and benefits of openness by making it possible for researchers to share and access scientific articles, the data underlying reported results, the methods used to generate and analyze data such as computer code, and other products of research. Openness increases transparency and reliability, facilitates more effective collaboration, accelerates the pace of discovery, and fosters broader and more equitable access to scientific knowledge and to the research process itself.

Many consider the 2002 launch of the Budapest Open Access Initiative (BOAI)—which called for free and open online access to the scientific literature—to mark the formal beginning of the open access movement (BOAI, 2002). In the years since, the emphasis of this movement has broadened from its original focus on open access to articles, and has come to include data, code, and other research products. What we know today as *open science* comprises both principles (transparency, reuse, participation, accountability, etc.) and practices (open publications, data-sharing, citizen science, etc.) (Open Science Training Handbook, 2018).

Open science stands at an important inflection point. A new generation of information technology (IT) tools and services holds the potential of further revolutionizing scientific practice. For example, the ability to automate the process of searching and analyzing linked articles and data can reveal patterns that would escape human perception, making the process of generating and testing hypotheses faster and more efficient. In order to have maximum impact, these tools and services need to be utilized as part of an open science ecosystem that spans institutional, national, and disciplinary boundaries.

Yet, despite the significant progress that has been made to create that ecosystem, today's science is not completely open. Most scientific articles are only available on a subscription basis (European Commission, 2018a). Sharing data, code, and other research products is becoming more common, but is still not routinely done across all disciplines (Figshare, 2017). Limitations and barriers to

more rapid progress include an academic culture and researcher incentives that can work against open science, insufficient infrastructure and training, issues related to data privacy and national security, disciplinary differences in the nature of research and treatment of data, and the economic structure of the scholarly communications market.

Research enterprise stakeholders around the world are making substantial efforts to facilitate and expedite the transition to open science. The European Commission has made the creation of a European Open Science Cloud one of its policy priorities (European Commission, 2018a). Other private and public funders, such as the Bill & Melinda Gates Foundation and UK Research and Innovation (the coordinating body for the United Kingdom's public research councils), have adopted policies to support open science. Science International (2015) assessed the "boundaries of openness" and proposed 12 principles to guide the practice and practitioners of open data, while the Académie des Sciences (France), German National Academy of Sciences Leopoldina, and the Royal Society (2016, 2017) jointly issued statements on scientific publications and good practice.

There is a growing world-wide consensus in the scientific community that the transition to open science, particularly in relation to digital data and code, can best be achieved by the establishment of a globally interoperable research infrastructure. A number of evolving projects around the globe, such as the Global Open (GO) FAIR Initiative, which originated in Europe, focus on involving all networked initiatives, research disciplines, and interested Member States of the European Union to make research data findable, accessible, interoperable, and reusable (FAIR) (GO FAIR, 2018). The international Research Data Alliance (RDA) and other groups have similar goals for international scientific data management. Societies, scholarly communicators, and the library community are also adopting policies and launching initiatives aimed at fostering a transition to open science.

In the United States, the Office of Science and Technology Policy (OSTP) issued a memorandum in 2013 instructing all federal agencies that spend more than $100 million per year on research and development to "develop a plan to support increased public access to the results of research funded by the Federal Government" (OSTP, 2013, p. 5). It also directs agencies to review options and needs for data repositories in areas of research they support and to require "all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers [to] develop data management plans" (OSTP, 2013, p. 5). The memo requires investigators to specify appropriate data management processes and options for long-term data access and preservation. A full text of the 2013 memo is provided in Appendix C. Federal policy has also been developed for nondigital scientific collections in a 2014 memorandum that is included as Appendix D.

Agencies developed and are implementing plans responding to the 2013 memo. For example, the National Institute of Standards and Technology (NIST) is participating in the National Data Service project, the National Institutes of Health (NIH) has proposed a Data Commons for biomedical research data, and

the National Science Foundation (NSF) has launched the Open Knowledge Network. Overall, implementation of the OSTP memo is uneven across agencies (Kriesberg et al., 2017). A 2017 report by the Association of American Universities and the Association of Public and Land-Grant Universities points out the need for agencies to set clear, consistent requirements and for agencies and universities to work together more closely in order to avoid a situation where standards and solutions are fragmented and not interoperable (AAU-APLU, 2017).

CONTEXT FOR THE STUDY

Recognizing the importance of accelerating progress toward open science, the Laura and John Arnold Foundation requested that the National Academies of Sciences, Engineering, and Medicine (the National Academies) undertake a study on identifying and addressing the challenges of broadening access to the results of scientific research. The committee was tasked with focusing on how to move toward open science as the default for scientific research results, with specific recommendations to be implemented (see Box 1-1 for the full statement of task). While the working definition of open science provided by the sponsor of the study is described in Box 1-1, the committee envisions that open science aims to ensure the open availability and usability of scholarly publications, the data that result from scholarly research, and the methodology, including code or algorithms, that was used to generate those data. Openness and sharing of information are fundamental to the progress of science and to the effective functioning of the research enterprise. In addition, although some of the analysis and discussion in the report is relevant to the humanities or other research-based disciplines outside of science and engineering, openness as it relates to those disciplines is not explicitly addressed.

In undertaking this task, the committee builds on previous National Academies work on related issues. The National Academies' first authoritative statement on research data issues and supporting openness came in 1985, for example (NRC, 1985). The 1997 report *Bits of Power: Issues in Global Access to Scientific Data* assessed a global perspective on open science and data in the natural sciences and identified strengths and challenges in the European community. In 2003, *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences* examined key principles and recommended that open data should be the default approach for biologists, including sharing data, software, and materials related to their publications in scholarly journals (NRC, 2003). The 2009 report, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*, called on researchers to make all research data and methods publicly accessible in a timely manner (NAS-NAE-IOM, 2009). As copyright issues are closely linked to open publications, *Copyright in the Digital Era: Building Evidence for Policy* (2013a) called on federal agencies and foundations to support a broad range of empirical research studies to contribute to the comprehensive review of U.S. copyright law.

BOX 1-1
Committee Statement of Task

Wide access to scientific research results has proven to be an important tool for accelerating scientific progress. An ad hoc committee under the Board on Research Data and Information (BRDI) will conduct a study on the challenges of broadening access to the results of scientific research, described as "open science." Open science is defined, for the purposes of this study, as public access (i.e., no charge for access beyond the cost of an internet connection) to scholarly articles resulting from research projects, the data that support the results contained in those articles, computer code, algorithms, and other digital products of publicly funded scientific research, so that the products of this research are findable, accessible, interoperable, and reusable (FAIR), with limited exceptions for privacy, proprietary business claims, and national security. This study focuses on how to move toward open science as the default for scientific research results and includes the following tasks:

1. Provide a cursory overview of the extent to which scientific and engineering disciplines currently practice open science;
2. Identify the barriers to and facilitators of open science, such as cultural norms, incentives, service provider business models, policies, available infrastructure, education/training, and formal and informal data management processes, and illustrate these barriers and facilitators in at least four scientific disciplines from the biological sciences, social sciences, physical sciences, and earth sciences;
3. Describe how policies and practices of participants in the research enterprise, such as funders, publishers, journal editors, research institutions, scientific societies, researchers, service providers, and the private sector, are affecting open science;
4. Recommend specific solutions in policy, infrastructure, incentives and requirements that would facilitate open science;
5. Identify existing implementations of these solutions occurring in individual disciplines that could be extended to other disciplines (e.g., preprints), and demonstrations of proofs-of-concept that need to be brought to scale (e.g., preregistration for basic and preclinical research);
6. For potential solutions with no existing demonstrations, identify practical implementation steps, policies, and appropriate stakeholder roles to develop solutions;
7. Provide specific policy and practice options for Federal science agencies to move toward open science as the default for the research they support.

The committee will produce a consensus report with findings and recommendations that address these issues, with the majority of the focus on solutions that move the research enterprise toward open science.

The National Academies have also made significant contributions to issues related to massive data and data sharing. *Frontiers in Massive Data Analysis* (2013b) provides perspective on generating, using, sharing, and analyzing massive amounts of data. The Institute of Medicine's *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk* (2015) offers detailed analysis and recommendations on the responsible sharing of clinical trial data. Most recently, several consensus reports were released in 2017 relating to open science. *Fostering Integrity in Research* includes several recommendations responding to integrity and reproducibility concerns. Recommendation seven in the report states, "federal funding agencies and other research sponsors should allocate sufficient funds to enable the long-term storage, archiving, and access of datasets and code necessary for the replication of published findings" (NASEM, 2017b, p.8). A final report that lays out a vision for future data science education was issued in 2018, and an expert committee is assessing research and data reproducibility and replicability issues as this report is written in 2018.

STUDY PROCESS

In discussing their approach to the task, the committee acknowledged that it was not asked to examine whether or not open science is good, but, rather, how to move it forward in ways that are beneficial to the scientific community. To accomplish its task, the committee held four 2-day face-to-face meetings to gather information from experts and develop findings and recommendations. Several virtual meetings were also held. As part of its evidence-gathering process, the committee organized a 1-day public symposium in September 2017 to explore specific examples of open science and discussed a range of challenges focusing on stakeholder perspectives. During the symposium, the committee heard speakers from professional journals, the private sector, philanthropic organizations, federal agencies, academic libraries, the research community, and scientific societies, who spoke on challenges, drivers, and progress toward an open science enterprise. The committee reviewed a large body of written material on open science concerns, including literature that informed the committee on how specific solutions in policy, infrastructure, incentives, and requirements could facilitate open science.

STRUCTURE OF THE REPORT

This report is organized into an introduction, four topical chapters, a chapter that frames and discusses the committee's findings and recommendations, and appendixes. Elements one, two, and three of the committee's task are mostly addressed in Chapters 2 and 3. Elements four, five, six, and seven are largely addressed in Chapters 4, 5, and 6.

Broadening Access to the Results of Scientific Research

Chapter 2 introduces the origins and significance of open science while analyzing advantages and motivations for and barriers to open science. Open science typically refers to the entire process of conducting science, including the collaborative underpinnings of the scientific enterprise. The contemporary focus on openness in science is spurred by opportunities for sharing knowledge via the Internet, and delivers multiple benefits to society: the right of taxpayers to gain access to the results of publicly funded research and the ability of researchers and non-researchers alike to retrieve, scrutinize, and build directly on the work of investigators around the world. Although the statement of task does not call on the committee to establish that open science constitutes a superior approach, the evidence that exists so far is presented and discussed.

As for barriers, infrastructure issues, such as policy, architecture, placement, and cost, become more complex as networked computers become the standard mode of scientific communication and much of scientific performance. Finally, the open science initiative is challenged to acknowledge disciplinary differences and to avoid unintended but potentially harmful violations of privacy, intellectual property, and national security.

The State of Open Science

Chapter 3 describes the general state and current approaches to open science, focusing on open publications and open data. As part of the committee's task, the chapter includes illustrative examples drawn from the disciplines of biomedical sciences, economics, astronomy and astrophysics, and earth sciences, along with other examples from outside of those disciplines.

While the research enterprise makes steady progress toward open science, it must navigate a complex environment of socio-political, economic, and practical challenges. Individual universities must develop their own access policies, although by now there are many successful models to guide the way.

With regards to articles, methods of publication have proliferated, featuring increasing use of preprints (which have not yet been published in a journal) and open access journals (which are freely available online to readers). The economic power of for-profit publishers persists, largely because many authors prefer to publish in journals that are considered to be the most prestigious in their fields and because many for-profit journals do not charge the authors themselves, relying instead on subscription revenue. In moving toward open publications, the community must consider not only when to adopt new open models, but also how to transition from the current mixed environment of closed and open models.

In the case of data, the committee and other experts expect data to become a dominant resource in the open science ecosystem. If this expectation is borne out, questions of lifecycle, reproducibility, compliance, and sharing need to be addressed by all stakeholders. These include differing data-sharing publication

practices in big science and especially in long tail/small science. Researchers with collections of physical objects, such as geological and paleontological objects, are pressed to consider not only their physical collection and data management plans, but also accessibility, reuse, and other issues common to all data collections. As the open science movement advances at a global level, it brings a critical need to foster international cooperation.

A Vision for Open Science by Design

Chapter 4 describes how open science can be implemented "by design" by defining open science by design as a set of principles and practices that fosters openness throughout the entire research life-cycle. The reader is invited to imagine a world of complete open publication, where all steps of the research process are findable, accessible, interoperable, and reusable (FAIR). This chapter explores the steps by which a researcher can access published ideas, build on them through data mining and other techniques, find and make use of existing concepts and methods in the existing literature, develop new hypotheses or methods, seek funding for an original pilot study, and publish results in appropriate venues. The chapter also discusses the need for enabling technologies and strengthening training for open science by design.

Transitioning to Open Science by Design

Chapter 5 discusses the legal frameworks and the context for realizing open science by design shaped by the policies and requirements of research funders. The chapter also identifies possible options and transition pathways to open science by design, including paying for open science, mandates, community-based initiatives, changes in the business environment, and possible short- and long-term options.

Accelerating Progress to Open Science by Design

Recent recommendations from other organizations are reviewed, including the AAU-APLU report discussed above (AAU/APLU, 2017) and the European Open Science Cloud (EOSC) Declaration of October 2017 (European Commission, 2017a). All such recommendations call out the need for developing new infrastructure and tools that support open science and open data. The report concludes with the committee's own findings, recommendations, and implementation actions specifying agencies, universities, or other organizations to guide stakeholder efforts to fostering open science by design.

The intended audiences for this report include researchers, universities, private and nonprofit organizations, information science communities such as publishers and journal editors, scientific societies, the philanthropic community, and federal agencies interested in improving access to the results of scientific research. In other words, this report provides specific policy and practice options for all

stakeholders, not just federal scientific agencies, to move toward open science as the default for the research they support. The committee hopes that the report will help these audiences better understand the possible barriers and facilitators, desirable data policies, and infrastructure requirements that would be required to implement open science.

2

Broadening Access to the Results of Scientific Research

SUMMARY POINTS

- The concept of open science, as it has emerged over the past several decades, is tightly linked with traditional scientific values and norms. At the same time, the digital revolution makes possible a restructuring of research practices and institutions built around the openness of publications, data, code, and other research products.
- Open science is motivated by a number of actual and anticipated benefits. They include the availability of the results of publicly funded research to the public, as well as more reliable and efficient research. Openness also enables researchers to address entirely new questions and work across national and disciplinary boundaries. Open science supports expanded access to the research process itself through citizen science activities.
- Despite the advantages and motivations for open science, significant barriers and limitations remain. These barriers and limitations include aspects of research culture and incentives that work against open science, insufficient infrastructure, resource constraints, disciplinary differences, policy and legal constraints, and lack of awareness.

ORIGINS AND SIGNIFICANCE OF OPEN SCIENCE

The concept of open science, sometimes also referred to as "open scholarship," is an ambitious goal that aims to ensure the availability and usability of (1) scholarly publications, (2) the data that result from scholarly research, and (3) the methodology, including code or algorithms, that was used to generate those data. The first of these is often known as *open access.* Since the term *open access* is sometimes used in other contexts, this report will use the term *open publication* instead. Ensuring the availability and usability of data resulting from research is known as *open data*. Ensuring the availability and usability of methods, in the case of computational work, is known as *open code*, and it is related to the concept of *open source software*.

Open science typically refers to the entire process of conducting science and harkens back to the original precepts underpinning the conduct and goals of the scientific enterprise (Storer, 1966; Borgman, 2010; Neylon, 2017). Openness has been seen as a "norm" of science: "The substantive findings of science are a product of social collaboration and are assigned to the community….The institutional conception of science as part of the public domain is linked with the imperative for the communication of findings" (Merton, 1942). In addition, openness facilitates realization of the scientific norm that results are critically examined before they are accepted (Merton, 1942). The digital revolution of the past several decades has vastly increased the possibilities of openness and lowered the costs:

> Shifting from ink on paper to digital text suddenly allows us to make perfect copies of our work. Shifting from isolated computers to a globe-spanning network of connected computers suddenly allows us to share perfect copies of our work with a worldwide audience at essentially no cost. About thirty years ago this kind of free global sharing became something new under the sun. Before that, it would have sounded like a quixotic dream (Suber, 2012).

More recently, the InterAcademy Council and the National Academies of Sciences, Engineering, and Medicine have reaffirmed openness as a core value of science (IAC-IAP, 2012; NASEM, 2017b). The European FOSTER (Facilitate Open Science Training for European Research) group has argued that open science is a concept that applies to the "whole research cycle, fostering sharing and collaboration as early as possible thus entailing a systemic change to the way science and research is done" (FOSTER, 2018; Figure 2-1).

The contemporary focus on openness in science has evolved in the context of the public Internet and the communication opportunities it has afforded, as well as the broadening of the scientific enterprise to include many new institutions worldwide. Distinct, but interrelated, motivations also include: the taxpayer's right to the results of publicly funded research; the ability of any member of society to scrutinize, evaluate, challenge and reproduce scientific claims; and the opportunity for anyone, including private citizens, to build directly on the scientific investigations of others. The motivations, benefits, and challenges of open science will be explored in more detail below. These factors all influence how open science is perceived, defined, implemented, and promoted (Royal Society, 2012; Fecher and Friesike, 2014; Pomerantz and Peek, 2016; Tennant et al., 2016).

Open publication is the most developed aspect of open science and has become more widely implemented over the past decade. Open publication refers to free and unrestricted access to publications with the only restriction on use being that proper attribution and credit needs to be given to the original creator of the work, as originally advocated by the Budapest Open Access Initiative, 2002, see

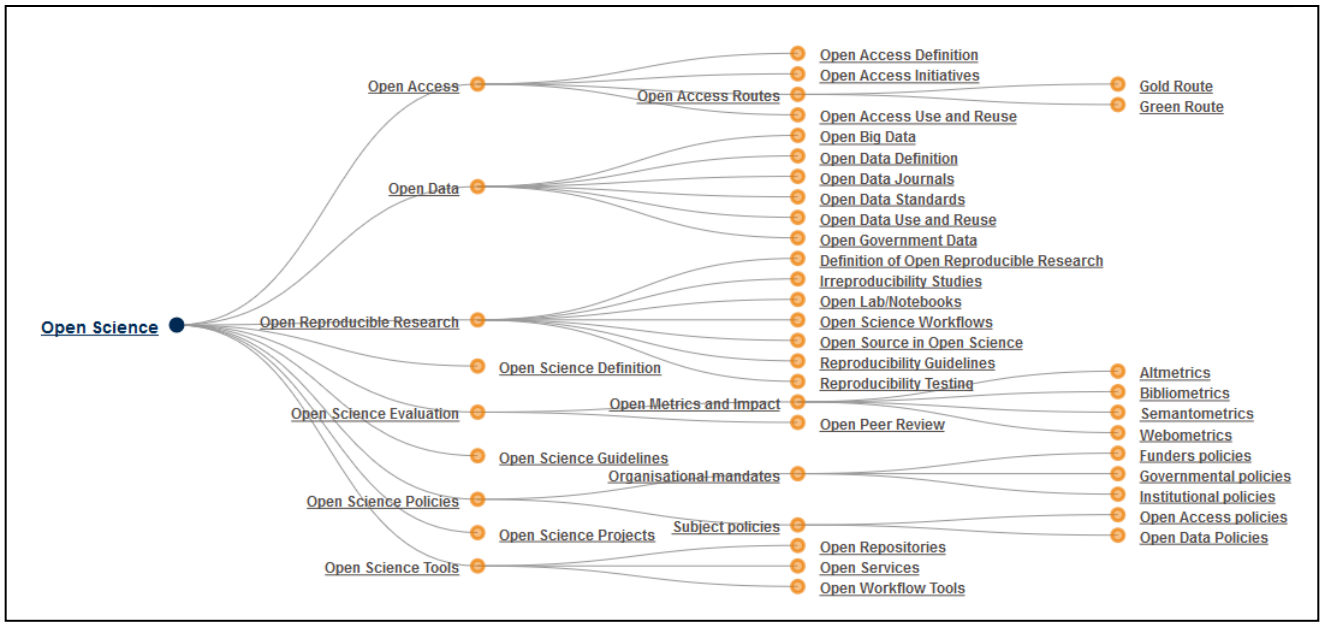


FIGURE 2-1 The FOSTER Taxonomy of Open Science. SOURCE: FOSTER (Facilitate Open Science Training for European Research) project. Online. Available at https://figshare.com/articles/Open_Science_Taxonomy/1508606. Courtesy of Attribution 4.0 International (CC BY 4.0).

Box 2-1).[1] Further, publications are to be deposited in "an appropriate standard electronic format" in at least one archive maintained by a reputable institution "that seeks to enable open access, unrestricted distribution, interoperability, and long-term archiving" (Open Access Max-Planck-Gesellschaft, 2003).

In the years since the first open access or open publication definition was put forward, open journals have emerged and traditional journals have, in some cases, revised their relevant policies. In an attempt to delineate the variation in interpretation of openness by journal publishers, the Public Library of Science (PLOS), Scholarly Publishing and Academic Resources Coalition (SPARC), and Open Access Scholarly Publishers Association (OASPA) have published the guide *HowOpenIsIt?* (Table 2-1). The guide assesses the spectrum of policies and approaches from fully open to closed along multiple dimensions. It suggests that fully open publication means that all articles in the journal are freely available to readers immediately upon publication. Immediate availability of articles at no cost to the reader beyond that required to access the Internet is known as *gold open access*. Other aspects of fully open publication in the realm of articles include generous reuse rights; the author holding copyright with no restrictions; the author being able to post any version to any repository or website with no delay; journals making copies of all articles automatically and immediately available in a trusted repository; and the full text of articles and supporting data being accessible via an application program interface (API) (SPARC et al., 2014). Less open approaches to publication include *green open access*, in which authors are able to self-archive a version of the article in an open access repository when access to the final published version requires a subscription to the journal.

BOX 2-1
The Budapest Open Access Initiative

By "open access" to [peer-reviewed research literature], we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.

SOURCE: Budapest Open Access Initiative, 2002.

[1] The Bethesda Statement on Open Access Publishing is a related statement with a focus on the biomedical research community (Bethesda Statement, 2003).

TABLE 2-1 *HowOpenIsIt?*

ACCESS	READER RIGHTS	REUSE RIGHTS	COPYRIGHTS	AUTHOR POSTING RIGHTS	AUTOMATIC POSTING	MACHINE READABILITY	ACCESS
OPEN ACCESS	Free readership rights to all articles immediately upon publication	Generous reuse & remixing rights (e.g., CC BY license)	Author holds copyright with no restrictions	Author may post any version to any repository or website with no delay	Journals make copies of all articles automatically available in trusted third-party repositories (e.g., PubMed Central, OpenAire, institutional immediately upon publication)	Article full text, metadata, supporting data (including format and semantic markup) & citations may be accessed via API, with instructions publicly posted	OPEN ACCESS
	Free readership rights to all articles after an embargo of no more than 6 months	Reuse, remixing, & further building upon the work subject to certain restrictions & conditions (e.g., CC BY-NC & CC BY-SA licenses)	Author retains/publisher grants broad rights, including author reuse (e.g., of figures in presentations/teaching, creation of derivatives) and authorization rights (for other to use)	Author may post some version (determined by publisher) to any repository or website with no delay	Journals make copies of all articles automatically available in trusted third-party repositories (e.g., PubMed Central, OpenAire, institutional) witin 6 months	Article full text, metadata, & citations may be accessed via API, with instructions publicly posted	
	Free readership rights to all articles after an embargo greater than 6 months	Reuse (no remixing or further building upon the work) subject to certain resturctions and conditions (e.g., CC BY-ND license)	_____	Author may post some version (determined by publisher) to any repository or website with some delay (determined by the publisher)	Journals make copies of all articles automatically available in trusted third-party repositories (e.g., PubMed Central, OpenAire, institutional) witin 12 months	Article full text, metadata, & ciations may be crawled without special permission or registration, with instructions publicly posted	
	Free and immediate readership rights to some, but not all, articles (including "hybrid" models)	Some reuse rights beyond fair use for some, but not all, articles (including "hybrid models")	Author retains/publisher grants limited rights for author reuse (e.g., of figured in presentations/teaching, creation of derivatives)	Author may post some version (determined by publisher) to certain repositories or websites, with or without delays	Journals make copies of some, but not all, articles automatically available in trusted third-party repositories (e.g., PubMed Central, OpenAire, institutional) within 12 months	Article full text, metadata & citations may be crawled with permission, with instructions publicly posted	
CLOSED ACCESS	Subscription, membership, pay-per-view, or other fees required to read all articles	No reuse rights beyond fair use/dealing or other limitations or exceptions to copyright (All Rights Reserved)	Publisher holds copyrights, with no author reuse beyond fair use	Author may not deposit any versions to any repositories or websites at any time	No automatic posting in third-party repositories	No full text articles available for crawling	CLOSED ACCESS

SOURCE: SPARC, PLOS, and OASPA. 2014. Online. Available at https://www.plos.org/files/HowOpenIsIt_English.pdf. Licensed under CC BY.

Note that copyright holder consent is a key requirement for making a publication openly available. Licenses designed to allow authors to retain copyright to their work have been developed by the Creative Commons organization, which allows authors to choose from one of several licenses consistent with copyright law (Carroll, 2011, 2015). The retention of copyright by authors for the purpose of making publications openly available has been one of the most contentious issues surrounding open publication, since it goes against journal publishing practices that require authors to assign the copyright to their work to the journals through copyright transfer agreements as a condition for publication.

Beyond open publication, much recent activity has been dedicated to the concept of open data, such as the availability of the data that support the research results reported in an article. Increasingly, the openness of data is seen as being critical to the progress of science, stimulating innovation, enhancing reproducibility, and enabling new research questions. Combining datasets for new insights and mining data through sophisticated machine learning algorithms are made possible by the open availability of datasets (Hrynaszkiewicz and Cockerill, 2012; Tennant, 2016). The Open Data Handbook (2018) offers this definition for open data: "Open data is data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and share alike." (Open Data Handbook, 2018). This implies that the data are available "in a convenient and modifiable form" such that there are no unnecessary technological obstacles to exercising licensed rights (Open Data Handbook, 2018).

The Panton Principles for Open Data in Science, among other points, emphasize that when publishing data, authors need to "make an explicit and robust statement" about their wishes regarding how their data can be used (Murray-Rust et al., 2010; Molloy, 2011). With a focus on data accessibility, stewardship, and reuse by humans as well as machines, the FAIR Guiding Principles were developed by an international group including individuals representing academia, industry, funding agencies, and publishers (Wilkinson et al., 2016; see Box 2-2).

It is important to note that FAIR data and open data are distinct but complementary concepts. FAIR data are not necessarily open, and open data are not necessarily FAIR. Data that are open *and* FAIR will maximize the impact of open science.

Finally, the concept of open code is fundamentally linked to open source software and the Open Source Initiative that was founded in 1998 (Open Source Initiative, 2018). Open source licenses allow users the right to modify software code and freely redistribute it. The licenses are motivated by a desire to share and improve code by participating in an engaged community of users and software developers. The recent focus on open code differs in that it has not been concerned solely with the collaborative nature of software development, but ties in with the broader goals of open science. With computation becoming an increasingly integral part of scientific research in many domains, the availability of data and computational methods for many research studies is critical to the evaluation, reproducibility, and extension of those studies. A workshop held at the American

Association for the Advancement of Science in early 2016 led to a set of recommendations to address this problem (Stodden et al., 2016). In order to allow for reproducibility, the group recommended that "data, code, and workflows should be available and cited" (Stodden et al., 2016).

The Transparency and Openness Promotion (TOP) Guidelines promulgated in 2015 are a set of recommended standards for adoption by journals to promote open practices, which encompass open data, research materials, and code (Nosek et al., 2015). The Guidelines are further described in Chapter 4.

BOX 2-2
The FAIR Guiding Principles for Scientific Data Management and Stewardship

To Be Findable:

F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To Be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2. the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To Be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To Be Reusable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1. 2. (meta)data are associated with detailed provenance
R1. 3. (meta)data meet domain-relevant community standards

SOURCE: Wilkinson et al., 2016.

MOTIVATIONS FOR OPEN SCIENCE

A vision of open science is unfolding in research communities across a wide range of scientific domains, driven by the expanding use of digital, easily shareable products of scientific research. These products range from publications to software used to produce results; from raw and/or processed data associated with research to digitized representations of physical artifacts. The rationale for opening the methods and outcomes of research is strong, multifold, and increasingly accepted by scientific, engineering, and biomedical investigators.

Published science has traditionally operated as a form of open or partially open commons or common-pool resource, subject to legal frameworks such as intellectual property rights and with a few exceptions such as those for proprietary research and research related to national security (Hess and Ostrom, 2003). Intellectual property issues are covered in Chapter 5. Researchers publish their work if they want to get credit and recognition, which sustains and advances their careers. Advances in information technology are greatly expanding the possibilities for using this resource. To the extent that science becomes more open and accessible, there should be more rapid and efficient progress in generating reliable knowledge. The more science is used, the more valuable it is. Individual researchers benefit as their own contributions become more widely known and recognized.

At the same time, there is a need to develop rules and norms to manage and cooperate in the use of this shared resource. What rules are needed to align the self-interests of the variety of stakeholders so that they contribute to the larger vision and realize the advantages of open science? Are specific efforts needed to ensure that the open science enterprise remains sustainable—that efforts to feed and replenish the commons run ahead of efforts to exploit it? What does sustainability mean in different national and disciplinary contexts? The economic analysis of open source software provides some insight on how communities can come together to create and sustain shared resources (Lerner and Triole, 2000).

This section describes the motivations for open science as well as the benefits: both those that are being realized today and those that can currently be anticipated. Chapter 3 includes more detailed descriptions of approaches to open science that are being taken in several different disciplines and their benefits. These benefits include enhancing the ability of the general public to access knowledge generated through publicly supported research, strengthening the reliability and efficiency of research, enabling researchers to address new questions, including those that cross disciplinary boundaries, and allowing a broader group of scientists to participate in the research enterprise on a global basis. The following section describes various barriers and limitations to wide implementation of open science.

Certainly, given the fact that the research enterprise as a whole is some distance from fully realizing open science, and since many of the benefits have yet to be realized, they are difficult to quantify. To that extent, this discussion is forward-looking. Many important transformations and innovations in the history of

science, and in history more broadly, have been opposed at first because of difficulty in quantifying or even imagining the benefits. For example, much of the biomedical research community was strongly opposed to the Human Genome Project when it was first proposed, believing that it diverted resources from more valuable investigator-driven work (Palca, 1992). The project and its impact look much different in hindsight. Today's advances in biomedical research, and many other fields such as archaeology, would not be imaginable without genomic mapping and analysis.

While there are undeniably significant costs associated with implementing policies and practices that support open science, realizing the benefits discussed in this section translates into a higher return on the investment of financial and human resources in research activity. Likewise, downstream societal benefits of research such as improved medical treatments and economically valuable technological advances can be realized more quickly and efficiently.

Ensuring the Reliability of Knowledge and Facilitating the Reproducibility of Results

Ensuring the reliability of knowledge and reported results constitutes the heart of science and the scientific method. Experimental research progresses by testing and refining hypotheses and building understanding based on the accumulated evidence. Throughout the history of science, there are examples of widely-accepted hypotheses being superseded or overturned due to failures to reproduce or replicate findings. Recent concerns about reproducibility and replicability in science emerged first in fields such as biomedical research and social psychology, but have become a broader issue in science (Economist, 2013).

In recent years, a number of efforts to reproduce or replicate published results have been undertaken. Several efforts in biomedical research found rates of reproducibility of fifty percent or lower (Begley and Ellis, 2012; Prinz et al., 2012). In 2015, the Open Science Collaborative attempted to replicate100 psychological studies published in leading journals (Nosek, 2015). Although 97 percent of the original studies had statistically significant results, OSC researchers were only able to replicate 39 percent of the findings. Camerer et al. (2016) replicated 18 laboratory experiments in economics and confirmed over 60 percent of the published findings. However, Chang and Li (2015) could only replicate half of the results in published economics journals using author-provided code and data because many journal data archives did not have the code and data.

John Ioannidis has highlighted issues such as underpowered studies, flexibility in study design and analysis, and publishing bias that favors articles reporting positive results as causes of irreproducibility (Ioannidis, 2005). Other causes include the use of underperforming computational tools in data analysis and cross contamination or misidentification of cell lines in biological research (Offord, 2018; Huang et al., 2017). Outright fabrication or falsification of data is also a cause of lack of reproducibility. Although there is not enough information available to estimate the percentage of published work that is fabricated or falsified,

there has been a steady stream of high-profile cases from countries around the world, and several examples of researchers in fields such as anesthesiology who have built entire careers on fraudulent work spanning 100 or more articles (NASEM, 2017b). While some level of irreproducibility is normal in research, the inability to replicate a very high percentage of scientific findings undermines the credibility of science (Wykstra, 2017).

How does open science relate to concerns about reproducibility? Certainly, open science in the form of open publication, open data, and open code supports the ability of researchers to confirm and reproduce findings. Ensuring openness and access facilitates better quality research through prevention of mistakes and more rapid and efficient discovery and correction of mistakes that do occur. Once it becomes common practice for significant and relevant portions of digital representations of scientific results to be open and shared, one can anticipate more care and attention will be paid to the process of preparing and producing the results—including their documentation—so that others can follow the process in more depth than was possible previously. Expectations and requirements for openness also allow for a more rapid discovery of fabrication and falsification of data, serving as deterrents to misconduct (NASEM, 2017b). In short, open science strengthens the self-correcting mechanisms inherent in the research enterprise.

Greater transparency is a major focus of those working to increase reproducibility and replicability in science (e.g., Munafò et al., 2017). The Reproducibility Initiative, launched in 2012 by Science Exchange, PLOS, Figshare and Mendeley, identifies and rewards high-quality reproducible research through validation of critical research findings (Science Exchange, 2018). Recent concerns over reproducibility have served to reinforce and catalyze progress toward open science in the form of new policies and practices adopted by research funders, research institutions, and publishers, as will be explored in more detail below.

Yet open science is not the only factor or solution to addressing the reproducibility issue, and open science will not automatically solve whatever problems there are. It should also be noted that some have questioned whether reproducibility is a significant issue for science (Fanelli, 2018). As this report was being completed in 2018, the National Academies of Sciences, Engineering, and Medicine was undertaking a study on reproducibility and replicability of research, that "will draw conclusions and make recommendations for improving rigor and transparency in scientific and engineering research and will identify and highlight compelling examples of good practices" (NASEM, 2018b).

Faster, More Creative, and More Efficient Knowledge Creation

In addition to improving the reliability and reproducibility of research, open science can aid the advance of knowledge in several other ways. First, open science can accelerate progress by making research more efficient. When scientific results are made openly available in digital form, they enable faster, deeper, and broader dissemination of the results to other researchers. Wider sharing and collaboration allows research communities to quickly access results and underlying

information, which, in turn, stimulates more, and more rapid, scientific discovery. New networking tools hold out the possibility of marshalling large collaborations of researchers who will be able to tackle problems more quickly and effectively than what is feasible today (Nielsen, 2011). When data resulting from clinical research on humans and on animals is reused, it maximizes the value of the contributions made by those research subjects to the advance of knowledge. It is important to note that sharing and reuse of data vary widely between disciplines. As will be explored in more detail in Chapter 3, significant data resources have been created in genomics and astronomy that demonstrate the value and logic of data sharing and reuse. In other domains, particularly those where the culture of sharing and reuse has not taken hold, benefits are not being realized (Wallis et al., 2013).

Second, open science enables researchers to ask and address entirely new sorts of questions. Semantically linked, machine-readable data can be analyzed by computers in order to reveal relationships within and between systems that would be impossible to discover otherwise (Science International, 2015). The potential for data from different disciplines being linked in this way and queried to understand complex phenomena and systems is particularly exciting. Increasingly, addressing complex problems of interest in science and society requires a multitude of methods and scientific results from different communities. This interdisciplinary work will be greatly aided by open, searchable, digital results that are made more available across communities. Without such interdisciplinary exchanges, modern problem-solving is hindered by leaving knowledge to be in effect locked inside a particular community—even when most members of a given scientific society have free access to journals and digital artifacts in a particular field. Furthermore, as search engines are able to go beyond keywords to follow scientific arguments from one paper and even community to another, interdisciplinary science has the potential to be highly accelerated.

While the above discussion implies that many benefits of this sort of work will be reaped in the future, as open science practices become more widespread, some examples can be seen today. What is needed to address complex problems is the ability to find and integrate results not only within communities, but also across communities—without paywalls or subscription barriers. Utilizing advanced machine learning tools in analyzing datasets or literature, for example, will facilitate new insights and discoveries. Further, digital platforms for extending and repurposing scientific results and connecting them across multiple communities, as well as sophisticated search engines that can follow scientific arguments from one result to another, will need to be developed and made available. Making data available under FAIR principles is critical to facilitating this acceleration in knowledge creation. For example, when data, software, algorithms, and other digital artifacts of the scientific process are made available and interoperable, they can more easily be reprocessed, modified, extended, or used for other purposes. For example, fields such as ecology and epidemiology combine disparate data from multiple sources to analyze phenomena such as oil spills and the spread of disease (Pasquetto et al., 2017).

What evidence is there that open science will deliver these benefits? Economists have studied the knowledge production process at a broad level and largely concluded that open science promotes knowledge discovery and better science. For example, Mukherjee and Stern (2009) developed an overlapping generations model that elucidates the tradeoff between secrecy and disclosure. Secrecy yields private returns whereas the private and social returns to disclosure and the benefits of open science depend on the use of scientific discovery by subsequent generations. The model shows that open science is associated with a higher level of social welfare. Another study examined the relationship between the innovative performance of biotechnology firms and their activity in academic publishing, and found that open science strategies had a positive impact on innovation (Jong and Slavova, 2014).

Economists have also studied the returns to open science in the context of publications and patents. Publications promote open science whereas intellectual property rights assigned by patents exchange public disclosure of an invention for the right of the inventor to exclusively exploit the invention for a limited time. (Chapter 5 further explores intellectual property issues related to open science.) Researchers have examined whether there is a trade-off between patenting inventions and publishing results, and found that these research activities are complements instead of substitutes (Stephan et al., 2007; Fabrizio and DiMinin, 2008; Azoulay et al., 2009). However, Murray and Stern (2007) and Fehder et al. (2014) identified publication-patent pairs and examined the impact of patenting on subsequent research. Publications appear before the patent is granted, and citations to the publication could potentially change once intellectual property rights were assigned. They found that papers were less likely to be cited after the patent was assigned, suggesting that patenting may close off inquiry and reduce knowledge creation in areas related to the patented invention. Aghion et al. (2010, 2016) studied the impact of NIH agreements that increased academics' access to patented, genetically engineered mice. They found that increased openness, measured by access to mice, prompted entry by new researchers and increased the diversity of research topics. They concluded that intellectual property rights decrease research interest and diversity. Williams (2013) examined the effect of Celera's patents on human genes on subsequent research and innovation. She found that patenting reduced research and innovation related to the patented genes by between 20 to 30 percent. The topic of how proprietary concerns may act as a barrier to openness is discussed below.

Researchers have also examined the impact of online access and open publication of scholarship on the number of citations. Online access to articles via subscription reduces search costs and likely increases citations, but the citation impact may be conflated with the quality of the journal. Evans and Reimer (2009) found that open publication increased citations to multidisciplinary journals by 20 percent. However, McCabe and Snyder (2015) showed that this estimated increase resulted from a specification error and disappeared when time effects were included in the model. They concluded that the citation benefit of open publication

in the previous literature was attributable to omitted variable bias from not controlling for journal quality. McCabe and Snyder (2015) found that JSTOR (an article repository) increased citations to economics and business journals by about 10 percent, but Elsevier's Science Direct appeared to provide no citation boost. Both JSTOR and Science Direct provide online access but are subscription-based, not open. McCabe and Snyder (2014) found that open publication increased citations to science journals by about 8 percent.

Eysenbach (2006) demonstrates that open articles have higher citations in PNAS than subscription access articles. Gaule and Maystre (2011) revisited this question and found no significant citation effect. Davis et al. (2008) and Davis (2010, 2011) conducted an experiment where submissions to 11 American Physiological Society journals were randomly assigned to open publication or subscription access. They found that open articles were more likely to be downloaded but received the same number of citations as subscription access articles one and three years after publication. McCabe (2013) concluded that the citation impact of open publication may have been overestimated by open access supporters. On the other hand, Wagner (2014) summarized a large, annotated bibliography on the topic with the conclusion that open access articles have a persistent citation advantage that varies by discipline.

How can we reconcile the findings of Aghion et al. (2010) and Williams (2013) which show that intellectual property rights were associated with less diversity in science, with the conclusions of Davis et al. (2008) and McCabe and Snyder (2015), which found limited impact of online and open publication on citations? First, genetically engineered mice and genetic tests patented by Celera are high-impact scientific discoveries. Limiting access to these discoveries closed down some productive avenues of inquiry. However, not all published articles are of the same quality. McCabe and Snyder (2013, 2014) found that open publication increased citations to the highest quality articles and decreased citations to the least-cited articles.

Expanding Access to Knowledge and to the Research Enterprise

Open science also expands access to knowledge and to the research process itself. One important justification for expanded access is the public support for a large portion of the research activity that leads to reported results. The federal government invested $121 billion in research and development (R&D) spending in fiscal year 2015. About $34 billion of the total is allocated to university R&D, resulting in datasets, publications, and other outputs (Rosenbloom et al., 2015; Edwards, 2017; NSB, 2018). Federal spending on intramural research totaled about $36 billion in 2015 (NSB, 2018). Over the past several decades, the belief that knowledge whose creation has been supported by the public should be accessible to the public has gained considerable ground. For example, disease advocacy organizations and consumer groups played an important role in support of NIH's policy of requiring that publications based on NIH-funded work be made available to the public following an embargo period (Albert, 2006). As will be explored

in more detail below, support for open science is growing among researchers, although attitudes are ambiguous (Odell et al., 2017). In 1997, the National Research Council recommended 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 (NRC, 1997).

The proposition that research data created through public funding should be publicly accessible as a default position has been advocated as an international standard. According to Science International, "if this social revolution in science is to be achieved, it is not only a matter of making data that underpin a scientific claim intelligently open, but also of having a default position of openness for publicly funded data in general" (Science International, 2015).

The strongest early practical rationale for this position came from biomedical research; the idea was that the public should be able to see and utilize the latest research relevant to promoting health and curing disease. This rationale spurred policy makers to support the development of the National Library of Medicine's PubMed interface to MEDLINE, NLM's database of citations to the literature, in the 1990s and to PubMed Central, NLM's full text article repository, in the 2000s (Varmus, 2009). Knowledge of biomedical research has helped communities facing health crises, such as AIDS activists, to better pursue their goals (NASEM, 2016). Health literacy and broader science literacy can help individuals, communities, and entire societies to benefit from research in areas such as popular epidemiology and participatory environmental monitoring (NASEM, 2016).

Open science may also contribute to a democratization of knowledge and a better informed citizenry (Arza and Fressoli, 2017). The proposition that scientific knowledge is a global public good raises an international dimension to this particular benefit of open science (NRC, 1997; Science International, 2015). Expanded international use of publicly-funded research may deliver positive benefits without disadvantaging the researchers who originally performed it or the national government that supported it. Developing country researchers are often enthusiastic users of open science resources (Swan, 2012). An estimated 80 percent of active journals in Latin America are open access (Science International, 2015). There are several open data initiatives in Africa, including the African Open Science Platform, which aims to "promote the development and coordination of data policies, data training and data infrastructure" across the continent (CODATA, 2016).

It may also be the case that the impacts of data-enabled science and technology on individuals and societies are so profound and potentially disruptive that deeper engagement with society is necessary both in solving existing problems

and legitimating emerging technologies (NASEM, 2017a). One-way communication of science to society is not enough. In many domains, science needs actively to engage with other societal actors as knowledge partners in jointly framing questions and jointly seeking solutions. The unprecedented ubiquity and diversity in modes of modern digital communication lend themselves to this task.

An additional reason for supporting broader access to scientific knowledge and the research process is that this access may speed scientific progress. The involvement of the broader public in the research enterprise, which is also called citizen science, has become more prominent in recent years, largely due to the progress of digital technologies and open science practices (Smith et al., 2017). For example, Zooniverse is a citizen science web portal that hosts projects in which volunteers assist professional researchers (zooniverse.org). There are many examples of citizen contributions to research in areas such as data gathering and environmental monitoring (Arza and Fressoli, 2017).

Although the benefits of open science are increasingly being realized and recognized, there are significant barriers to a research enterprise and environment where access to research products is routinely expected. These barriers as well as approaches to overcoming them will be discussed in the next section.

BARRIERS TO OPEN SCIENCE

Some barriers to open access to research products may be addressed through the development of new tools and institutions. While some barriers can only be lowered through thoughtful changes in the policies and practices of research enterprise stakeholders, others are interrelated in complex ways. Some barriers are more relevant to one component of open science than to others (i.e., open publications, open data, or open code). This section will provide an overview of the major barriers, including information on how difficult change is likely to be.

Economic Barriers

Some of the most challenging barriers to open science are the incentives of market participants and the structure of the market for scholarly communication, particularly in the area of open publication. The scientific article, which is peer reviewed and compiled with other articles within a journal, is the traditional approach to disseminating new research. Scientific journals emerged during the 17th century (Fyfe et al., 2015). Traditionally, journals have been distributed to institutions (e.g., university libraries) and individuals via subscription. Since World War II, there has been a global expansion of research activity, leading to rapid growth in the number of articles published.

Publishers perform many important functions as a key component of the research enterprise. These functions include organizing the peer review process, developing and implementing policies in areas such as responsible conduct of research; addressing authorship problems; performing an array of technical tasks such as format migrations; and managing relations with authors, vendors, and the

media (Anderson, 2016). Journal publishers also maintain the information technology infrastructure that supports and controls access to content as well as the development of new infrastructure and platforms. Publishers of scientific journals have included a range of for profit and nonprofit entities, many of the latter being scientific societies. Robert Maxwell's UK-based Pergamon Press worked to make journal publishing a profitable business starting in the 1950s by launching new journals and recruiting top scientists to edit and contribute to them (Buranyi, 2017). Pergamon and other commercial publishers also took on the task of publishing the journals owned by some scientific societies. Profits increased with the number of journals, as libraries would simply add new journals requested by faculty to their subscription lists. From the 1970s on, scientists began to pay more attention to the prestige and visibility of the journals in which they published. The advent of the journal impact factor, described in more detail below, contributed to this focus on prestige. Publishing in a "high-impact" journal came to be seen as essential to career progress in many fields (Buranyi, 2017). Annual subscription prices rose as well.

The 1990s brought a wave of consolidation among scientific publishers, as Netherlands-based Elsevier acquired Pergamon, leaving it in control of over 1,000 journals (Buranyi, 2017). Further increases in subscription prices and the advent of "big deal" agreements between publishers and libraries followed in the late 1990s. Under these agreements, publishers agree to provide online access to a bundle of their journals, including all back issues, priced at a discount to the sum of the individual journal subscriptions (Bergstrom et al., 2014). Despite paying lower per journal prices, total outlays by libraries increased to the point where this has been called the "serials crisis" (Panitch and Michalak, 2005). In 2015, Larivière et al. found that the five most prolific publishers, including Reed-Elsevier, Taylor & Francis, Wiley-Blackwell, Springer, and Sage, control over one-half of all the scientific journal market, and that the profit margins of these companies have been in the range of 25 to 40 percent in recent years (Larivière et al., 2015). According to one economist who studies the industry, this situation "demonstrates a lack of competitive pressure in this industry, leading to so high profit levels of the leading publishers that they have not yet felt a strong need to change the way they operate" (Björk, 2017a).

Unlike some other intellectual property-based businesses such as recorded music, the incumbent firms in commercial scientific publishing have been able to navigate technological and other changes while maintaining a profitable business model based largely on subscription revenue. In contrast to music or other parts of commercial publishing, where firms pay creators for content, authors of research articles are not paid by the publishers. Research is supported by public and private funders and by the performing institutions.

Nonprofit publishers also occupy an important place in the scholarly communications ecosystem. The most prominent of these are scientific society publishers, although university presses and other nonprofit organizations, such as the Public Library of Science (PLOS, described in more detail in Chapter 3), also participate. Publishing has long been a core activity of many societies. The size

and relative importance of society publishers varies considerably by discipline and according to the specific society in question. For example, the American Chemical Society publishes 50 peer-reviewed journals and is one of the top five publishers of articles in chemistry (ACS, 2018; Larivière et al., 2015). By contrast, in the social and behavioral sciences, society publishers play a smaller role in overall scholarly communication than in disciplines such as physics and chemistry (Larivière et al., 2015).

Society publishers undertake publishing activities as part of their overall mission of providing service to their members and disciplines. They have traditionally used a business model centered on subscription income. For some societies, publishing operations generate a surplus that they use to subsidize other activities, such as education programs or meetings (Collins et al., 2013). Available information indicates that there is a considerable variation among disciplines and individual societies regarding the size of the surplus (if any) generated by publishing and the extent of the society's dependence on that income. For example, in 2011 subscriptions and manuscript charges accounted for 53 percent of the revenues of the Ecological Society of America and journal publication accounted for 43 percent of expenses, with society revenue and expenses each totaling over $6 million (Collins et al., 2013).

Over the past several decades, as technological change has transformed scientific publishing and for-profit publishers have increased their overall share, society publishers have faced the challenge of investing in digital production and distribution systems and responding to changes in markets and author preferences. For example, in the life sciences, where the number of journals offered by for-profit publishers has increased rapidly, some society journals have faced increased competition for manuscripts. Whereas 20 years ago an author whose manuscript was rejected by, say, Nature might then submit it to a society journal, today the author is more likely to submit to Nature Microbiology or another disciplinary journal offered by a for-profit publisher (Schloss, et al., 2017). Some societies have entered into partnerships with for-profit publishers, in which the company performs most non-editorial functions and includes the society's journals in its own subscription bundles, paying the society a fee in return. The American Geophysical Union's partnership with Wiley-Blackwell is a good example (AGU, 2012).

Competition from self-publication and open science have not seriously affected the market share of commercial and nonprofit publishers of high-prestige journals. Exploring the incentives of stakeholders gives some insight into why this may be the case:

- *Researchers:* Researchers have the incentive to maximize the visibility of each scientific discovery. These incentives are reinforced by the academic promotion and tenure processes at universities and by funders. Promotion and tenure requirements incentivize researchers to maximize the prestige of the journal in which their papers are published. Funders also require proposals to include publications, and journal impact factors are used as

proxies for the quality of science (Ginther et al., 2018). Researchers both consume and produce scholarship. Researchers prefer to read and cite high-quality work (McCabe, 2013). Researchers have no market power when it comes to publishing their research, and they prefer to publish work in a widely read journal. Researchers provide free labor to journals in addition to production of research articles in the form of editing and peer review (Bergstrom, 2001). Researchers also do not typically bear the costs of subscribing to journals if they are affiliated with an institution. Finally, researchers may bear the cost of open publication through article processing charges, while publishing an article in a traditional subscription journal is generally without cost to the researcher. Of course, researchers who are working at institutions that cannot afford subscription fees and cannot themselves afford to pay the article processing charges levied by open publication journals do not enjoy legal access to the system. To reduce the knowledge gap across the globe, Research4Life, a public-private partnership of international organizations, universities, and 175 international publishes, provides developing countries with affordable access to research and scholarly information (Research4Life, 2018).

- *Universities:* Universities seek to maximize the visibility and productivity of their faculty. Because university administrators and tenure review committees may not be subject matter experts, they rely on signals of quality for their research faculty. These include the number of publications, the prestige of the journals where faculty publish, and their success in research funding. All of these outcomes are linked to scholarly publication. Universities also purchase journals for their students and faculty at fees increasing faster than the rate of inflation, especially from commercial publishers (Bergstrom et al., 2014).
- *Research funders:* Federal research funders are held accountable by Congress. The peer review process is designed to allocate funding to the "best" science. Past accomplishments in terms of the prestige of publishing venues are used to forecast whether the current research proposal is of sufficient quality to be funded. Thus, research funders also use journal publications as proxies for quality (Ginther et al., 2018).
- *Scientific societies and other nonprofit publishers:* Scientific societies promote the scholarship of their disciplines for their members. They typically publish journals, and journal revenues may in turn support the activities of the association (Willinsky, 2004). Other nonprofit publishers such as university presses also seek to maximize the readership of their journals and cover their costs via subscription fees. Publishers pursuing open access business models are discussed in more detail in Chapter 3.
- *Commercial publishers:* Typically, publishers bundle journal subscriptions as a way of cross-subsidizing lesser journals by including high profile journals in the bundle.

Given these incentive structures, it becomes easier to understand the market structure of scholarly publication. Economists have studied the scholarly communication market structure in order to understand why for-profit publishers continue to have market-pricing power in the face of competition from self-publication and open access journals. Furthermore, while there are significant "first copy" costs, the marginal cost of providing online access to journal content is essentially zero. This situation persists because many of the incentives of researchers, universities, and funders create a powerful motivation to leave the current system in place: when the contribution of an idea is difficult to measure, institutions use signals of quality (e.g., citations, prestige of the journal) to infer quality (Bergstrom, 2001).

Varian (1994) argued that marginal cost pricing is not profit-maximizing for information goods such as scholarly publications. Thus, publishers have an incentive to engage in first-degree price discrimination, where they sell the same bundle of journals at different prices to different consumers. Bergstrom et al. (2014) examined the prices paid by public university libraries for "big deal" journal bundles from commercial and nonprofit publishers. They found significant price discrimination by commercial publishers by the research-intensiveness of the university, and a lesser amount of price discrimination by nonprofit publishers.

The "big deal" pricing strategies of journal publishers have played a major role in shaping the market for research journals. First, publishers recognized that demand for the journals was inelastic and priced subscriptions to maximize rents. Second, the shift from a physical journal to online access meant that libraries effectively "rented" access to the current journal as well as the older volumes of the journal. "Big deal" bundle pricing may have also made it difficult for new journals to enter the market given that university library budgets were being squeezed (McCabe 2013). McCabe (2013) argued that the cost pressures on libraries associated with "big deal" pricing led to the open access business model. This business model shifts the costs from subscribers (university libraries) onto the researchers. The Public Library of Science (PLOS, the largest and most highly cited open access journal publisher) charges publication fees ranging from $1,595 for *PLOS ONE* to $3,000 for *PLOS Biology* (PLOS, 2018). McCabe, Snyder and Fagin (2013) argue that the current pricing structure of open access journals may dissuade publication. The higher publication fees distort the market, leading to fewer submissions and potentially reducing the volume of publications. Further, Poynder (2018) argues that national open access "big deals" of the type that publishers conclude with higher education bodies in some European countries allow publishers to protect their market positions. These agreements combine subscription fees with discounts on the APCs paid to the journals by researchers at institutions covered by the agreement. One important aspect of these and other large subscription agreements is that they generally include non-disclosure agreements, so that purchasing organizations are not able to discern the prices that others are paying.

In response to competition from open access journals, some subscription-based publishers are offering a hybrid open access model, where authors can pay

a publication fee and the article is freely available. Mueller-Langer and Watt (2014) examined the impact of hybrid open access (HOA) pilot agreements between commercial publishers and the University of California system, the Universities of Hong Kong and Goettingen, all universities in the Netherlands, and the Max Planck Institutes. They found that HOA has no significant impact on citations after controlling for institution quality and citations to preprint versions of the article.

Society publishers are also responding to these trends. As discussed above, the size and importance of publishing activities varies by discipline and society. Societies have adopted new policies and expressed varying perspectives on trends in scholarly communication and open publication in particular. Some societies with large publishing operations have adapted their approaches to the movement toward open publication. For example, ACS offers a range of HOA (hybrid open access) options for authors, with the APC to be charged varying according to the license desired, the length of the embargo period to be followed, whether ACS is responsible for depositing the final published article in a designated repository or whether the author is responsible for depositing the accepted manuscript, and so forth (ACS, 2018). ACS has also launched its own open access journal and a preprint service.

Society publishers have expressed a range of perspectives in their public statements and policy positions as well. They are generally supportive of open publication in principle, but are skeptical about the imposition of funder mandates that require gold open access at the time of publication, or green open access with embargo periods of less than one year (Collins et al., 2013). The American Physical Society "supports the principles of Open Access to the maximum extent possible that allows the Society to maintain peer-reviewed high-quality journals, secure archiving, and the Society's long‐term financial stability, to the benefit of the scientific enterprise" (APS, 2009).

It is important to remember that scholarly communications involves real costs, and that the current state of the subscription journals market is the result of choices made by publishers, institutions, researchers, and funders over many years. Some experts argue that moving away from traditional publishers operating on a subscription model would entail forgoing the benefits of significant investments in digital infrastructure that publishers are making, and would constitute a short-sighted "race to the bottom" (Anderson, 2018). As noted above, journal revenues play an important role in supporting the programs and activities of scientific societies that advance individual disciplines and science as a whole. Some pathways to open publication, such as mandates that specify immediate gold open access or eliminate embargo periods for green open access, would be problematic for many societies and their ability to sustain their professional infrastructure.

Yet the issue is complex. Some might question why research library budgets that have been under considerable pressure should be expected to generate surplus funds to support the professional activities of societies. Others are more skeptical about the ultimate value provided by commercial publishers in particular, given

their large profit margins (discussed above), arguing that they benefit from publishing research that is funded by other sources, and that writing, reviewing, and some portion of editing tasks are performed by volunteers (Conley and Wooders, 2009). Publishing journals as a profit-maximizing business is certainly as legitimate as it is for other distributors of digital content based on intellectual property protections. The research enterprise and its stakeholders are responsible for the future of scholarly communication. Chapters 5 and 6 will cover the issues and choices facing the research enterprise in moving forward.

Academic Culture and Misaligned Incentives

One important set of barriers to open science springs from the fact that many of the benefits redound to research communities and the broader research enterprise itself, yet researchers are recognized and rewarded largely based on their individual production and accomplishments. The culture of open science is seen as being about advancing the public interest—when research products are broadly available and discoverable, they benefit more people and drive more innovation than when they are not. Research also has some characteristics of a public good in economic terms, in that use by one individual does not reduce availability to others. However, researchers can be excluded from using publications and other research products.

Getting Scooped

Barriers related to culture and incentives operate at several levels. At one level, researchers might be concerned about being "scooped" by other researchers if data are shared openly and reused by others before the researchers who generated them are able to fully exploit them in multiple publications (EC, 2018b). In some fields and disciplines, particularly those where acquiring data involves considerable effort or expense, such as collecting specimens from remote areas, or undertaking epidemiological studies that require a number of complicated steps, delays in sharing data underlying the first publications may be an accepted practice (Pearce and Smith, 2011). Whether or not the risk of being scooped is overstated, some adjustments in rewards and expectations may be necessary to address this concern in the fields where it exists in order to facilitate more rapid and complete data sharing. For example, institutions and disciplines might work to ensure that the first person to share research outputs receives appropriate credit, and that researchers who generate valuable and widely reused datasets receive proper attribution. Ultimately, the solution to ensuring that data are shared quickly and lessening the perceived need for delays motivated by career interests is ensuring that those who create valuable data are recognized and rewarded, but restructuring reward systems is not straightforward or easy. The rationale that sharing data quickly will deliver public health benefits and perhaps even save lives may not win out over the desire to hold data closely in order to ensure that one's postdocs and graduate students are able to author publishable work based on this data. Note

also that the same rules should apply to all as efforts are made to appropriately reward data creation and sharing. If some researchers practice open science and others do not, the ones who do not may enjoy competitive advantage. When funders and other stakeholders require openness of publications and data as a consequence of receiving funding, a more level playing field can be created.

Exposure of Errors

Another concern that might make researchers reluctant to share data and methods is that such sharing would expose their errors to the community. New research workflows in which reporting results and sharing research products takes place within a process where community review helps to uncover error will improve the reliability of results, as described above. Preregistration of studies can help to uncover mistakes in analytical approaches before data are collected. Journals such as *PeerJ* and *Open Science*, the latter published by the Royal Society, have instituted open peer review, another mechanism aimed at improving the quality of research (McKiernan et al., 2016). It may take time for research communities to transition to open practices that enable wider review and scrutiny of research. Psychology is a current encouraging example. Concerns about reproducibility led many inside and outside the field to critically examine practices and standards, and new open practices such as preregistration and replication studies are being tried and refined (Winerman, 2017).

At the same time, some experts have raised concerns in recent years about the nature of scientific disputes in the context of changing standards related to transparency or reproducibility. The rise of blogs, social media, and venues for post-publication comment and review has greatly expanded opportunities to correct, criticize, raise questions, and make accusations against researchers, often anonymously (NASEM, 2017b). Disciplines where standards and practices are being reexamined, such as psychology, have seen intense disputes over the validity of widely heralded results as well as over the tone and personal nature of the critiques. While some prominent leaders in the discipline have identified the harsh nature of criticism itself as a significant issue, others argue that raising concerns over tone diverts attention and focus away from the substance of critiques (Singal, 2016). It is important for errors or misconduct to be identified and corrected; it is also important that small errors or legitimate differences in analytical choices not be cast as malfeasance. In order to maximize the value of greater openness and transparency, disciplines and the research enterprise itself may need to devote some attention to developing new norms around the pursuit of accuracy and related issues (Gelman, 2018).

Career Considerations

In addition to concerns arising from relatively short-term potential impacts of sharing specific research products, longer-term career considerations may also explain reluctance on the part of some researchers to adopt open practices.

Achieving the vision of open science requires scientists to make results publicly accessible and to engage in sharing data with the community as an expected practice. Researchers are motivated by the possibility of gaining career advancement, support, and recognition for their work in addition to curiosity and the desire to advance their fields (EC, 2017b). Career prospects in science are increasingly challenging especially for early-career researchers because of the scarcity of permanent academic positions and the difficulty of getting funded (Stephan, 2012a). Individual researchers may not perceive that taking the steps necessary to make their own work accessible will be in their best interests. Data sharing requires a focus on data preparation and infrastructure for stewardship, preservation, and broad use. In the absence of clear requirements to do so, scientists who take the time to make sure that software is robust, data are sufficiently described, and data stewardship and preservation meet good practice and community standards may not be rewarded by higher education institutions (e.g., through promotion and tenure or infrastructure support) or recognized within their disciplines. Preparing data and code for deposit involves considerable time costs. Researchers may suffer if they prioritize their open science work that benefits the community at the expense of publishing more journal articles.

Some aspects of current research evaluation practices may contribute to concerns about how openness and open practices affect the career prospects of researchers. The most salient issue is the importance of bibliometric indicators such as the Journal Impact Factor (JIF) in evaluating research and researchers (Declaration of Open Research Assessment, DORA, 2013; Casadevall and Fang, 2015). Developed in the 1960s by the Institute for Scientific Information (and now a product of Clarivate Analytics), JIF measures the yearly average number of citations to recent articles in a particular journal (Cross, 2009). The ability to digitally index articles, which allows JIF and other indicators to be automatically tracked and calculated, has enabled the development and wide use of JIF and other bibliometric indicators.

The use of bibliometric indicators in research evaluation affects researcher rewards and incentives both directly (in hiring or promotion) and indirectly (as a factor in funding or publication decisions). It is widely perceived around the world that the JIF of the journals that researchers have published in plays an outsized role in hiring and promotion decisions in research institutions (Abbott et al., 2010; Casadevall and Fang, 2015). JIF was not developed as a tool for evaluating research or researchers, and there are numerous reasons why using it in this way is inappropriate. These reasons include: (1) citation distributions within journals are highly skewed, meaning that JIF may not accurately track the citation profile of individual articles; (2) there are wide differences between fields in typical citation patterns, so researchers in fields where influential articles may take several years to be heavily cited are disadvantaged; (3) JIF and other indicators can be gamed by journal editors, research institutions, and individual researchers; and (4) JIF is not transparent, as the data and methodologies underlying it are proprietary (DORA, 2013; Wilsdon et al., 2017).

Some experts argue that the misuse of JIF and other bibliometric indicators may even cause broader harm to researchers and to the research enterprise itself. The contention is that apparent imbalances within some parts of the science and engineering workforce and low rates of success in research funding proposals to U.S. federal agencies have helped to create an environment of hypercompetition that discourages risk taking, shortchanges quality control, and dissuades researchers from sharing (Alberts et al., 2014; Fang and Casadevall, 2015; NASEM, 2017b; Stephan, 2012b). Such hypercompetition may directly discourage open practices such as sharing data and other research products if researchers are primarily concerned with maintaining an advantage. Vale and Hyman (2016) argue the heightened competition between scientists in high-profile journals has strained the peer-review system; however, "the need for a system of validation has only become more pronounced as the volume of scientific work has increased" (p. 4). Researchers in a hypercompetitive environment might also prioritize publishing their work in journals with the highest possible JIFs, regardless of whether publication in such journals is consistent with making research products available under open principles. No researcher's career has been harmed by publishing in high-impact journals.

Countervailing Factors and Efforts to Address Barriers Related to Culture and Incentives

All of the barriers to open science discussed above related to culture and incentives are likely higher and more challenging for early career researchers than they are for their senior colleagues (Eveleth, 2014; The Guardian, 2018). Although some of these barriers may take considerable time and effort to address, there are some encouraging signs of positive change. First, the potential negative effects of open practices on careers, including anxieties about being "scooped," may be shrinking over time as advantages become more apparent. As discussed above, open publication may confer an advantage in terms of citations (Hitchcock, 2018; Wang et al., 2015). This merits continued study. There is also evidence that media coverage and social media discussion of openly published research is greater than that for traditionally published work (Wang et al., 2015). Further, there are indications that JIFs of indexed open access journals may be increasing compared with those of traditional, subscription journals (McKiernan et al., 2017). Moreover, more subscription journals are allowing authors to deposit preprints or postprints that are openly available (sometimes in response to funder mandates) or offering an open publication option for purchase by the author. The benefits and downsides of these options are discussed in more detail below.

In addition to encouraging progress toward open practices within the context of conventional reward and incentive systems, the participants in the research enterprise can also take steps to change cultures and incentive systems in ways that explicitly encourage and reward open practices. For example, a number of prizes and funding programs launched in recent years have recognized and supported open

science (McKiernan et al., 2017). Funder, institutional, and publisher policies mandating open policies also contribute to changing culture and incentives.

New efforts to publicly track the extent to which researchers follow open practices are also being developed. One well-known example is the initiative led by the Center for Open Science (COS) and several journals to assign badges to accompany published articles where authors have shared data or materials, or preregistered their studies (COS, 2018a). While this initiative has yielded encouraging results, further work is necessary to separate the impact of badges from other editorial changes supportive of open practices introduced at the same time, and to confirm other results of introducing badges (Kidwell et al., 2016; Bastian, 2017). At a broader level, funder and journal openness mandates may generate data that can be utilized by community compilation and reporting efforts aimed at improving transparency. For example, FDAAA Trials Tracker is a website launched in 2018 that gathers information on compliance with U.S. Food and Drug Administration requirements that all clinical trial results be reported and makes the information available in an accessible format (FDAAA Trials Tracker, 2018). Box 2-3 describes additional requirements related to open access to clinical studies.

Another approach is to modify researcher evaluation criteria and tools in order to avoid discouraging open practices or even to explicitly reward them. Preventing the misuse of JIF and other bibliometric indicators in the evaluation of research and researchers is one possible approach. The 2013 San Francisco Declaration on Research Assessment is one prominent effort that has gained many signatories among institutions, funders, and journals (DORA, 2013). The 2015 Leiden Manifesto for Research Metrics is a parallel effort (Hicks et al., 2015). Both of these statements emphasize the importance of expert judgement in the evaluation process.

Efforts are also ongoing to take advantage of the capabilities of information technologies and the explosion of online interactions to develop new measures of research impact that would address some of the negative aspects of the JIF and enable a broader consideration of the value of articles and other research products. Taken together, these new measures have been labeled *alternative metrics* or *altmetrics*. For example, efforts are underway to develop substantially new citation-based indicators based on transparent metric calculations that are open to scientifically based oversight (Hutchins et al., 2016). Others are developing metrics that go beyond citation-based indicators, incorporating information on downloads, mentions on social media, and other online reader behavior (NISO, 2014; Howard, 2013). Developing new indicators to evaluate research and researchers and facilitating their use will require a better understanding of technical and institutional prerequisites to their use—such as standards for digital author identifiers—and how these might be put in place. Indeed, the open science movement itself can provide the impetus to the improvement and wide use of high-quality metrics, and these metrics can play an important role in recognizing and rewarding open practices (Wilsdon et al., 2017).

BOX 2-3
Clinical Research

Access to information about clinical studies is important to researchers, health care professionals, and patients. For many years, patients seeking information about clinical studies were dependent on their clinicians to know about and recommend relevant studies. While their clinicians might have been aware of the clinical trials being conducted at their own institutions, there was no easy way to find out whether there was a suitable study elsewhere, even at a neighboring institution. Patient advocacy groups and others argued that information about clinical trials should be readily available to members of the public and that such availability should be required by law.

At the same time, because clinical trials are the cornerstone of evidence-based practice, many investigators had called for better reporting of clinical trials research (Meinert, 1988; Haynes, 1998). Meta-analyses and systematic reviews depend on the most comprehensive information possible for making recommendations about changes in medical practice. One author (Chalmers, 1990) went so far as to say that it is "scientific misconduct" not to report the results of one's research.

In late 1997, a section of the Food and Drug Administration (FDA) Modernization Act required the creation of a database of information about clinical trials (FDA, 1997). The law directed the Secretary of Health and Human Services through the National Institutes of Health (NIH) to establish, maintain, and operate a "registry of clinical trials (whether federally or privately funded) of experimental treatments for serious or life-threatening diseases and conditions."

The law required that for each clinical trial listed in the registry there be at least a description of the purpose of the experimental treatment, the eligibility criteria for participation in the trial, the location of the trial, and, most importantly for patients, a point of contact for enrollment. Beginning in early 1998, a working group comprising members from the NIH and the FDA began planning the implementation of the registry, and the National Library of Medicine, which had extensive experience in developing biomedical databases, took on the task of developing what became known as ClinicalTrials.gov. Standard data elements, standard methods for labeling and transmitting the data, use of standard vocabularies, and use of standard web technologies all played a role in the design of the system. ClinicalTrials.gov was launched in February of 2000 (McCray, 2000; McCray and Ide, 2000). In addition to interactive searching, the data can be freely downloaded and reused according to specified terms and conditions. As of 2017, there are several hundred thousand trials from around the world registered in ClinicalTrials.gov and an increasing number of these include detailed results data.

The legislative requirements for making clinical trials data available were critical both for the original development of ClinicalTrials.gov as well as for its continued significant expansion and growth. The initial 1997 law was amended a decade later to require submission of not just a description of the protocol design and eligibility criteria, but also the results of completed trials

(Continued)

2-3 Continued

(FDAAA 801, 2007). The final rule for implementation of this amendment was issued in 2016 and includes guidance for assessing compliance. Perhaps equally important for the extraordinary growth of the database was a joint statement by the editors of prominent medical journals in 2004 (ICMJE, 2004) that advised authors of clinical trials reports that a condition for publication would be deposit in a public registry at the inception of the trial.

References

Chalmers, I. 1990. Underreporting research is scientific misconduct. *The Journal of the American Medical Association* 263(10):1405-1408.

FDA (Food and Drug Administration). 1997. PUBLIC LAW 105–115—NOV. 21, 1997. Food and Drug Administration Modernization Act of 1997.

FDAAA (Food and Drug Administration Amendments Act) 801, 2007. PUBLIC Law 110-85 – Sept. 27, 2007. Food and Drug Administration Amendments Act of 2007.

Federal Register. 2016. Clinical Trials Registration and Results Information Submission. 42 CFR Part 11. Docket Number NIH – 2011-003. RIN 0925-AA55. 2016. National Institutes of Health, Department of Health and Human Services.

Haynes, B., and A. Haines. 1998. Barriers and bridges to evidence based clinical practice. *The BMJ* 317:273-276.

ICMJE (International Committee of Medical Journal Editors). 2004. Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. Online. Available at http://www.icmje.org/news-and-editorials/clin_trial_sep2004.pdf. Accessed March 30, 2018.

IOM (Institute of Medicine). 2015. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington, DC: The National Academies Press.

McCray, A. T. 2000. Better access to information about clinical trials. *Annals of Internal Medicine* 133(8):609-614.

McCray, A. T., and N. C. Ide. 2000. Design and implementation of a national clinical trials registry. *Journal of the American Medical Informatics Association* 7(3):313-323.

Meinert, C. L. 1988. Toward prospective registration of clinical trials. *Controlled Clinical Trials* 9:1-5.

Finally, broader efforts are underway to rethink research evaluation practices and develop new approaches that place less emphasis on JIF and other bibliometric indicators and more emphasis on other contributions of researchers, including adherence to open practices. For example, the Peer Reviewers' Openness Initiative proposes that peer reviewers commit to withholding comprehensive review of submissions where data or materials are not openly available (Morey et al., 2016). A 2017 European Commission (EC) report describes a new approach to evaluating researchers and their career contributions where open practices are

central (EC, 2017b). Some experts advocate a fundamental rethinking of approaches to peer review characterized by openness, with scholarly communications organized around network or library concepts rather than fixed journal articles (Kriegeskorte et al., 2012; Kennison and Norberg, 2015).

Privacy and Security Concerns

Privacy Concerns

As described above, open science is critical for addressing the reproducibility challenge in scientific research while facilitating future research that validates or builds on previous results. An unintended and potentially harmful consequence of publicly sharing research data, however, is the possible effect on privacy. Researchers have long recognized the privacy implications of publicly sharing research data, especially when such data involve human subjects, such as patients in a clinical trial. The tension between privacy protection and scientific openness is longstanding. For example, many studies in the area of public health pertain to health care records and medical history, which makes it extremely difficult, if not impossible, to maintain patient privacy while openly sharing all the information necessary to reproduce or replicate a published study (O'Neill et al., 2016).

Traditionally, researchers rely on anonymization, or "de-identification," methods to strike a balance between open data and human subject privacy. The idea is that once all personally identifiable information has been removed from a published dataset, an individual would no longer be associated with any record in the dataset. Participants in research studies expect that the data collected about them will be handled with care and that, unless they have given explicit consent to have their personal information shared, their data will be safeguarded. The federal government has provided specific guidance through its HIPAA legislation, which provides standards for the electronic exchange, privacy, and security of health information.[2] The intent of the legislation is to safeguard personally identifiable information, known as PII. HIPAA's "safe harbor" defines 18 specific attributes (e.g., name, phone number, medical record number) as "protected health information" in need of suppression (CDC, 2003).

In recent years, however, it has become clear that even anonymized data can reveal private information about the human subjects. The key challenge here is that even attributes that are not labeled as personally identifiable may still contain sensitive information that associates an individual, and that by linking those data to other publicly available resources, individuals can be reidentified. (Sweeney, 1997, 2002, 2003, 2009; Malin and Sweeney, 2001). In a case study of a state-released dataset containing 2.8 million hospital records, investigators showed that even after removing from the dataset all information except the pro-

[2]The Health Insurance Portability and Accountability Act of 1996 (HIPAA), Public Law 104-191. See https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations.

cedures received by a patient, the percentage of patients with a unique set of procedures is still 42.8 percent; in other words, as the investigators state, "an adversary would have about a 42.8 percent chance of linking the anesthesia record to the hospital database, thereby discovering the patient's sensitive information." (O'Neill et al., 2016)

In August 2016, after AOL Research released 20 million search queries issued by its users (with no user identifier or personal information attached), a reporter from *The New York Times* was still able to locate an individual from the anonymized search records by cross referencing the *contents* of the queries with phonebook listings (Barbaro and Zeller, 2006). Similarly, researchers were able to re-identify individuals in an anonymized version of Netflix's movie preference database for a contest that challenged researchers to try to improve its recommendation engine. By comparing rental dates and ratings in the Netflix database with reviews posted on the Internet Movie Database, the researchers were able to discover individuals' entire rental histories, potentially revealing sensitive information about them (Narayanan and Shmatikov, 2008). As a result of this re-identification, a class-action lawsuit was filed against Netflix, and, as part of the settlement, Netflix cancelled a second planned contest.

After making numerous attempts to develop better mechanisms for disassociating individuals from a published dataset (Sweeney, 2002; Machanavajjhala et al., 2006), researchers in the field of data privacy realized a fundamental issue with many of the then-existing techniques: these techniques rely on assumptions of "*adversarial background knowledge,*" i.e., the external sources of information an adversary has access to beyond the dataset being released. Examples of background knowledge include phonebook listings in the AOL example and hospital databases in the health care example. One can see that such background knowledge is plentiful and hard to enumerate in practice, leading to privacy violations even after anonymizing the data.

Recent advances in data privacy aim to address this issue by developing techniques that are agnostic to adversarial background knowledge. A notable example is the concept of *differential privacy* (Dwork, 2008), which is a uniform privacy guarantee no matter what background knowledge an adversary possesses. A wide variety of techniques has been developed to achieve differential privacy, mostly by inserting random noise into the data being released or to the query answers being generated from the dataset. In spite of these advances, there are still significant challenges facing the wide adoption of differential privacy in the research community. A notable one is how to validate previous research results or establish new findings from data that have already been perturbed with random noise. While one might be tempted to simply rerun the original research workflow over the perturbed data, research has shown that doing so may lead to statistically invalid results that require complex, task-specific procedures to correct (Gaboardi et al., 2016; Rogers et al., 2016). As such, the proper balance between open data and privacy protection of human subjects is still a major ongoing challenge. Several repositories have been developed as emerging solutions to these issues, in-

cluding Genotypes and Phenotypes (dbGAP) for genotype-phenotype relationships (Mailman et al., 2007; dbGap, 2018), the Yale University Open Data Access (YODA) project for clinical trials (The YODA Project, 2018), and the forthcoming Vivli platform for clinical research (Vivli, 2018). However, these repositories are expensive to set up and manage, and should be part of the infrastructure that is developed to support open science.

National Security Concerns

Openness of research results has been a source of tension in security research and practices for years. For example, the "export" of cryptographic technology was severely restricted in the United States until 1992, after such export control was already challenged by individual level openness efforts such as PGP,[3] which was released in 1991. A key argument in discussions of the effect of openness on national security is that providing open access to data and methodology might have the unanticipated outcome of aiding malicious individuals and organizations. Specifically,

- Adversaries might use openly available data or methods to make the design and implementation of their attacks easier. For example, an adversary might directly adopt an open-source machine learning algorithm to bypass CAPTCHA challenges commonly used by web security applications.
- Adversaries might also leverage the openly shared knowledge of a mission-critical system to find bugs or vulnerabilities to defeat the system itself. For example, by examining publicly available data on the Supervisory Control and Data Acquisition system used by a power station, an adversary might be able to design more effective attacks on the power network.

Both of these points reflect long-standing debates in security-related research. For the first concern, one can draw an analogy to the debate of whether researchers should be allowed to publish the computer security vulnerabilities they identify for, say, an encryption algorithm, or if such flaws should be kept behind closed doors to prevent adversaries from taking advantage of them (Cavusoglu and Raghunathan, 2007). As in the computer security case, while there might be perceived costs from the adversarial usage of open data and methods, what outweighs such costs is the effect open data and methods have on informing and incentivizing defenders to strengthen their defenses (Pond, 2000), easing the design and implementation of defensive systems, and eventually ensuring progress in the research fields critical to national security. In other words, openness benefits both attacker and defender, and, arguably, more the defender than the attacker.

[3]PGP (Pretty Good Privacy) is freely available software for the encryption of electronic mail and other data (Zimmerman, 1995).

The second concern reflects the debate between security-through-obscurity and security-by-design. The former tries to maintain security by hiding knowledge of the system design from attackers, with the premise that, without knowing how a system is designed, an adversary would not be able to effectively attack it. Security-by-design, on the other hand, recognizes that hiding system design from attackers rarely works in the long run, as an attacker can accumulate knowledge of the system design over time by using the system, observing its behavior, and other methods. Thus, security-by-design assumes the system design to be public knowledge, and aims to make the design inherently secure even when an adversary knows how it works. The progress of computer security research in the last few decades has repeatedly shown that security-by-design is the only viable long-term approach (Cavoukian and Chanliau, 2013).

Insufficient Infrastructure

Infrastructure provides the engine that supports the vision of open science. If articles, data, code, and other research products constitute the content that is to be available under FAIR principles, open science infrastructure consists of the tools and metadata through which research products are created, shared, and assessed, including "data about the research process itself, such as reference lists and funding information" (Peters, 2017).

As noted earlier, the foundation that enables open science is the spectacular improvement in the capacity and performance of information technologies that has occurred in accordance with Moore's law and related formulations (Moore, 1965). For example, computing power has increased exponentially as chip densities have grown from a thousand transistors in 1970, to a million transistors in 1990, to a billion transistors in 2010. At the same time, network bandwidths have increased from thousands of bits per second in the 1980s to millions of bits per second in the 1990s to billions of bits per second in the 2000s. The capacity of storage devices (electromechanical disks and electronic flash memory) has grown from millions of bytes to billions of bytes to trillions of bytes. For example, a terabyte capacity disk now costs less than 100 dollars. Because of this great increase in capacity, we can now store more data than we can effectively and efficiently process. Ongoing data science research will contribute to the advance of open science as well as data processing techniques.

As discussed above, FAIR data is a requirement of open science. An example of FAIR data for human use is provided by public webpages. Search engines have made many such pages findable and they are usually either immediately accessible or accessible via a paywall. Since these pages are designed for human readers, they are made (more or less) interoperable by the readers' knowledge of the language and the subject matter. Pages are often reusable by cut-and-paste document editing tools. Open science data should also be FAIR for software agents. This requires that both a wider array of data be available and that knowledge about the data be "machine readable." That is, machine-readable

metadata should be available for software agents to support automated interoperability and reusability.

Other attributes of data that are important for open science include trustworthiness and citability. Techniques for assessing and rating trustworthiness are essential to enable proper reuse of data (and to avoid harmful reuse). And citability is an important step towards rewarding scientists for publishing important data. The definition and use of DOIs (digital object identifiers) is a related example of a useful technique for uniquely identifying journal articles.

The Semantic Web is a vision for how data and knowledge might be stored online in a machine-accessible form. The Semantic Web offers a set of standardized computer languages for representing data and knowledge ("recommendations" of the World Wide Web Consortium), and one of these languages, the Resource Description Framework (RDF), is well suited for representing the metadata needed to make online datasets FAIR. In fact, the Center for Expanded Data Annotation and Retrieval (CEDAR), a standards-based metadata authoring system developed under the NIH Big Data to Knowledge Program, uses RDF for precisely this purpose (CEDAR, 2018).

A second architecture, the Digital Object Architecture (DO), has also been under development for several decades. The DO addresses the interoperability of heterogeneous data in a manner similar to how the Internet addressed the interoperability of heterogeneous networks: that is, a new layer of abstraction is introduced. In the case of the Internet, TCP/IP (Transmission Control Protocol/Internet Protocol) defined a virtual network that interconnected physical networks at computers. In the case of the DO, a digital object is a virtual data object that references a data object at a lower level of abstraction. Just as an Internet message has a header that contains the necessary metadata to transmit the message, a DO digital object has a "landing page" containing the necessary metadata to understand and manipulate the digital object. As an additional point of similarity, just as each computer has an Internet address (its IP number), so each digital object has a "handle," which is used to reference its digital object.

There is a need for infrastructures that semantically link research objects to each other, such as persistent identifiers for research objects (PIDs), and standard ways of collecting, expressing as metadata and semantically linking PIDs. Some groups are developing such services and integrations, including ORCID, the Data Citation Implementation Pilot (DCIP) project, and FREYA under the European Commission's Horizon 2020. These efforts are described in Chapter 4.

The distributed location of data repositories is an issue that is mitigated as network performance continues to improve. That is, distributed data is increasingly understood to be the norm for data processing activities. And in some cases, the dataset is too large to move efficiently and "processing is brought to the data" rather than data being brought to a processing center. In fact, the location of the data has increasingly been pushed into the background by the emergence of cloud computing. Cloud computing is sometimes called the industrialization of IT, much as the electric power grid was the industrialization of local generation of electricity. This industrialization of the underlying infrastructure for open science,

among other things, turns IT capital costs into operating costs and could thereby accelerate the emergence of open science.

The question of "who pays" remains important, especially for science, which has seldom been overfunded. If open science infrastructure remains an unfunded mandate, the movement towards an open science enterprise will be significantly slowed. Proposals on both sides of the Atlantic have been made to address this problem. In the United States, NIH has considered calling for the establishment of a "data commons," which would be financially supported by grant funds earmarked for data infrastructure. In Europe, consideration is being given to a similar earmarking of some research funds.

Appealing again to the Internet experience, the NSFnet was supported from the start with a mixture of NSF funds, university funds, and private sector investment funds. When the NSFnet was retired in 1995, universities began shouldering most of their networking costs themselves. And after the Federal Next Generation Internet program awarded research universities grants to connect to (what became) Internet2, virtually all network costs were borne by the universities themselves. In Europe, however, university networking costs continue to be partially supported by government.

Making data, code, and other research outputs available under FAIR principles involves both a number of specific short-term and long-term costs that need to be covered. As discussed in the next chapter, many "big science" projects in astronomy, high-energy physics, and genomics are funded and undertaken with the starting assumption that the resulting data are a central output. The hardware, software, and other resources needed to enable long-term access to data are included in the budget and built as part of the project itself. Likewise, in the case of smaller projects, the costs of cleaning and formatting data, ensuring that adequate documentation and metadata are attached, and other short-term costs may be supported by the grant.

As for long-term costs, some disciplinary communities have built institutions and repositories that are responsible for keeping smaller community datasets, such as the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan. However, the cultures of some disciplines might lack a shared understanding that data should be curated and made available on a long-term basis. Berman and Cerf (2013) used the example of sensor data that are made available for several years after the research is concluded, paid for by grant funds, but where there is no funding to support longer-term access. What if research a decade later would benefit from access to and reuse of this sensor data?

Erway and Rinehart (2016) reviewed various possible funding strategies for long-term data management, noting that funders are increasingly advocating that institutions accept responsibility for data management as a library preservation function. They found that institutions are mainly supporting data management services through their library budgets, but that some are exploring more diversified sources of funding. In taking on a larger role, institutions might need to be more involved in working with researchers to decide how and when data may be released, ensure data quality, and meet requirements for protection of private information.

Performing these functions would help support rigor and protect the institution's reputation, but would also require additional resources and capabilities.

Much important work remains, with key tasks and decisions facing all the participants in the research enterprise. Ensuring that resources for management and long-term stewardship of data and other research products are available—including highly trained data scientists, tools, and data standards—will require significant long-term effort on the part of stakeholders working across disciplines, sectors, and national boundaries. As will be discussed below, researchers in several fields have made significant progress and have created numerous examples and models that hold the potential for wider deployment.

Disciplinary Differences in the Nature of Research and Data

Differences in the nature of research and the types of data collected may create special barriers or limitations in sharing data, reusing data, or ensuring the long-term availability of data. The privacy and national security barriers discussed above are examples. Other challenges arise from the size or complexity of data generated by "big science" projects, such as those in some areas of physics. An important and emerging type of data are the very large datasets that capture extremely rare, time-sensitive events. Subtleties in this data and their generation may not be readily captured without detailed knowledge of how the data were collected. Safeguards may be needed to prevent misuse or misrepresentation of certain types of data. The challenges of making such data available for sharing and reuse, and providing for long-term curation, are considerable.

For example, seismology illustrates issues related to the reproducibility and replicability of research results, discussed above. While it is impossible to replicate a given unique natural phenomenon such as a seismic event, it is possible to reproduce an analysis of the data collected on an event (i.e., analyze the same data using the same software). The seismology community around the world maintains a network of regional data archives that facilitate study and understanding of earthquakes and other seismic phenomena. For example, the Southern California Earthquake Data Center, founded in 1991, operates the Seismological Laboratory at the California Institute of Technology and serves as an archive of seismological data for southern California. It links to other seismological data archives around the world.

Long-term curation and stewardship of data is another general challenge that affects disciplines differently. For some "big science" fields, funds to support data sharing and archiving are included in the overall project budget, but stewardship may be difficult or impossible to sustain once the project or experiment ends. Even in fields where the size or complexity of data do not present particular challenges, communities may not have well-developed standards for deciding which datasets are of long-term value and how or where they should be curated. Chapter 3 discusses several specific examples of challenges related to data stewardship.

The Laser Interferometer Gravitational Wave Observatory (LIGO) is an example of a project that is generating very large, complex datasets and that illustrates the challenges of imagining a route to complete open data that would allow an outsider to carry out credible analysis of the data streams from three sites (LIGO, 2018). Caltech and the Massachusetts Institute of Technology operate LIGO with support from the National Science Foundation. LIGO achieved the first direct observation of gravitational waves in 2015. The LIGO detectors collect very large amounts of data on astronomical events that occur erratically or evolve slowly, such as the collision of black holes many light years away from earth. A great deal of knowledge about the detectors themselves, the analytical software, and other aspects of the experiment is required to use the data effectively. In working to overcome these challenges of size and complexity, LIGO supports data sharing and reuse through the LIGO Open Science Center (LOSC, 2018). In addition to providing access to LIGO data packages on specific events, the LOSC site includes video tutorials and extensive data usage notes.

Open Science and Proprietary Research

This report focuses on transitioning to open science mainly in the context of published research. In most cases, the principles, practices, and expectations for openness in published research should not vary according to whether the funder is a federal agency, private foundation, or profit-making company, or whether the performer is a university, government laboratory, or corporate researcher. When a company performs research that produces an invention for which intellectual property protection should be secured and where results are publishable, it can choose to file a patent application before the relevant research article is published. If open science requirements such as data sharing would expose information about the research that the company does not wish to publicize, it can choose not to publish an article and protect the invention through patenting or trade secrecy. This principle is seen in clinical research, where requirements for preregistration and data sharing are being codified and enforced regardless of funding source or performer (FDA, 2007; Taichman et al., 2017).

Open science does have implications for proprietary research in some areas where the need to publish and stay on the cutting edge overlaps with interest in developing products. Some research methods and technologies fall into this category. For example, many advances in biomedical research techniques involving zinc-finger proteins and zinc-finger nuclease have been patented, leading to a complex intellectual property landscape that affects how research and product development progresses in academic and corporate settings (Chandrasekharan et al., 2009). Open science data and materials options have been developed to work around some barriers caused by proprietary data and materials (Chandrasekharan et al., 2009).

It will be important to see how the relationship between proprietary research and open science evolves in the future. It is possible that companies will find that participating in an open science ecosystem is beneficial and advances innovation.

It is also possible that companies will find it more difficult to manage intellectual property risks in an open science world, which might constitute a disincentive to performing research.

Research Underlying Regulations

The above discussion illustrates that transitioning to open science will involve addressing a number of complex issues involving how the research enterprise operates and how it relates to the broader society. This process will necessarily require time, development of new approaches, and a certain amount of trial and error. This report develops a vision for moving forward and identifies priority tasks. However, several important issues that lie largely outside the scope of the study will remain.

One example is the implementation of open science practices in research relevant to policymaking and regulation in areas such as environmental health. An Environmental Protection Agency (EPA) proposal for new requirements for openness that would cover research underlying some regulations spurred spirited debate at the time this study was being completed (EPA, 2018). Opponents of the proposal argued that it would unduly restrict the scientific basis for regulations, while proponents argued that the change would improve transparency and that the concerns were overblown (985 Scientists, 2018; Hahn, 2018).

Although the issues raised here are outside the scope of this study, the example does illustrate that implementing requirements for open science in certain policy contexts will raise difficult questions, and may become politicized. Issues of data access and quality have been subject to political debates in other areas, such as climate change, and will continue to be (NASEM, 2009). There will be cases where data and code cannot be made completely open, but where the results should not be simply rejected out of hand. Ensuring that efforts to expand openness and transparency are consistent with other priorities will be a key challenge in realizing the benefits of open science.

3

The State of Open Science

SUMMARY POINTS

- Despite the barriers discussed in Chapter 2, open science has made steady progress over the past several decades. More and more research products are available on an open basis. Still, this progress has been uneven, and the research enterprise remains some distance from achieving complete open science.
- Several significant trends have expanded the possibilities for publishing articles on an open basis. These trends include the emergence of open publishing venues, author self-archiving through institutional repositories and preprint servers, and open publication mandates adopted by funders and institutions. However, a large percentage of the world's scientific literature is still only available via subscription. Achieving universal or near-universal open publication in a way that serves the research enterprise and its stakeholders remains a challenging, pressing task.
- In the area of data, code, and other research products, there has also been significant progress toward developing practices and infrastructure that would support openness under FAIR principles. There are wide disparities by discipline, with some coming close to the expectations of open data and others quite far away. Different disciplines face different challenges in fostering open data related to cost and infrastructure. For example, some disciplines lack well-developed metadata standards, researchers may not have the incentives or resources to prepare data according to FAIR principles, and repositories that support FAIR data might not be available.

GENERAL STATE OF OPEN SCIENCE

In the 15 years since the Budapest Open Access Initiative (BOAI) issued its declaration, there have been numerous efforts to promote and realize open science. A growing number of public and private research sponsors around the world are mandating open publication, open data, or both, on the part of grantees, with some variety in the specifics of their policies, including the National Institutes of Health, the National Science Foundation, the Bill & Melinda Gates Foundation,

the European Commission (EC), and the Wellcome Trust. The University of Southampton maintains a repository of open science policies adopted by funders and research organizations (Figure 3-1; ROARMap, 2018). Supportive tools and infrastructure have been developed, including discovery platforms (e.g., Science-Open and 1Science) and browser-based extensions (e.g., Open Access Button, Canary Haz, and Unpaywall) (Piwowar et al., 2018). Academic social networks, such as ResearchGate and Academia.edu, provide an increasingly popular but controversial solution to author self-archiving (Van Noorden, 2014). At the same time, some articles are shared in copyright-violating pirate sites, such as Sci-Hub and LibGen, provoking debate over the efficiency and ethics of traditional models of scientific publishing (Björk, 2017b; Piwowar et al., 2018). The open science movement has catalyzed new investment, prompted controversy, and had a significant impact on the global research enterprise and its stakeholders. While underscoring the impact of existing policies and progress made, Figure 3-1 also reveals the speed of change and puts in perspective the need for additional efforts.

Several entities have monitored and analyzed the progress and status of open science. Most of these efforts focus on open publication. For example, Science-Metrix, a Canadian science data analytics company, found that as of 2013 over half the articles published during the period 2007–2012 were available for free download (Science-Metrix, 2014). Using oaDOI technology, an open online service that determines open publication status for 67 million articles, it is estimated that at least 28 percent of the literature is open (green or gold, 19 million articles in total) and that this proportion is growing, driven particularly by growth in gold and hybrid open access adoption (Piwowar et al., 2018). Piwowar et al. (2018) also suggested that the most common mechanism for open publication is not gold, green, or hybrid open access, but rather an under-discussed category of articles made free-to-read on the publisher website, without an explicit open license (Piwowar et al., 2018). In December 2017, Web of Science, a large bibliographic database, began to release more detailed data on the availability of publications than were available previously, categorizing open articles as "gold," "green accepted," or "green published" (Bosman and Kramer, 2018; Library Research News, 2018). Most recently, Science-Metrix (2018), analyzed three bibliographic databases (1Science database, Scopus, and Web of Science) to measure the availability of open publications, finding that at least two-thirds of the articles published between 2011 and 2014 and having at least one U.S. author could be downloaded for free as of August 2016 (Science-Metrix, 2018).

Using newly available open publication status data from oaDOI in Web of Science, Bosman and Kramer (2018) explored year-on-year open access levels across research fields, countries, institutions, languages, funders, and topics by relating the resulting patterns to disciplinary, national, and institutional contexts. They find that openness varies significantly by discipline, with the highest levels (over 50 percent) in some life sciences/biomedicine and physical sciences/technology fields and lower levels (under 20 percent) in social sciences and arts/humanities (Bosman and Kramer, 2018). Within the broad category of social

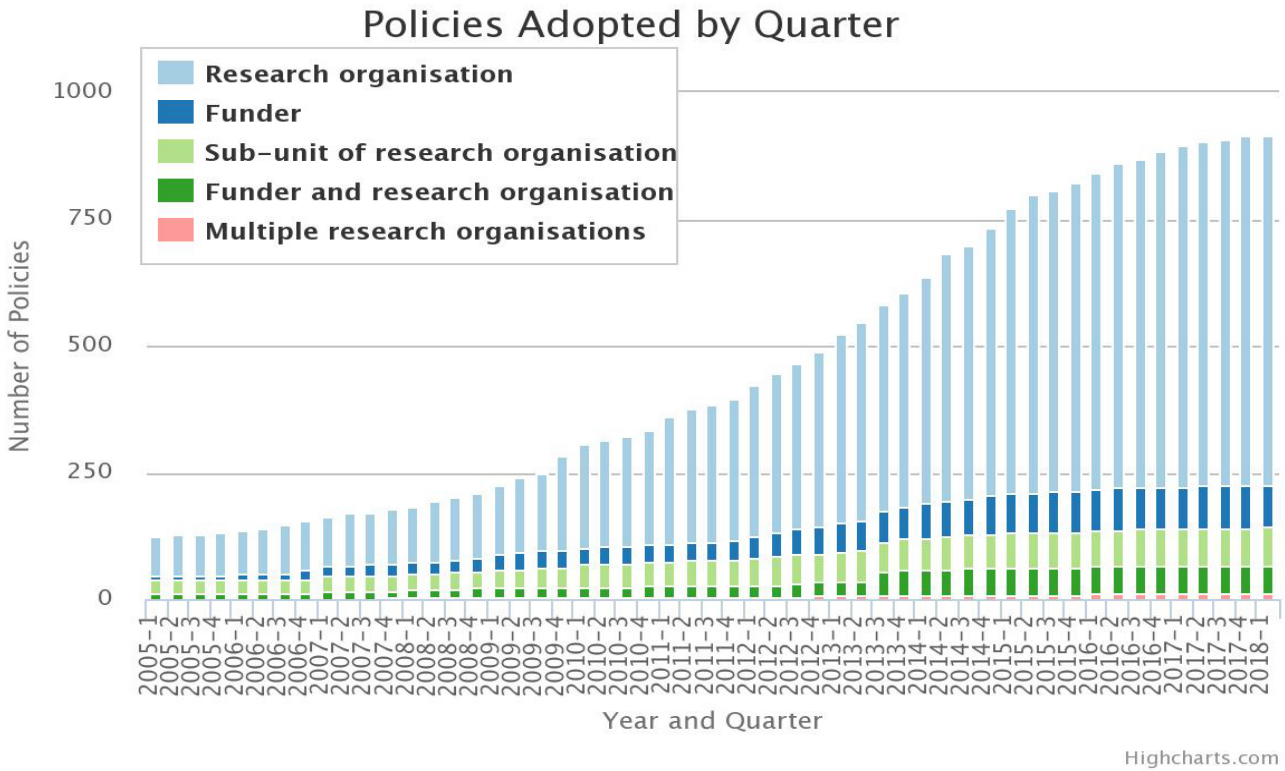


FIGURE 3-1 Open science policies adopted by research funders and research organizations around the world. SOURCE: ROARMap, University of Southampton.

sciences, psychology registers the highest levels of open publication, possibly because its publication culture is more similar to life sciences/biomedicine than to the other social and behavioral sciences. Similarly, Piwowar et al. (2018) found that over half of the papers are freely available in biomedical research and mathematics, while less than one-fifth of the publications in the disciplines of chemistry and engineering and technology are freely open (see Figure 3-2). The figure demonstrates that green open access is popular in physics and mathematics, while hybrid articles are common in mathematics and biomedical research. Authors in biomedical research, mathematics, health, and clinical medicine often publish in gold journals. Regarding specialties within disciplines, over 80 percent of publications in astronomy and astrophysics, fertility, and tropical medicine were open. On the other hand, more than 90 percent of publications are hidden behind a paywall in pharmacy, inorganic and nuclear chemistry, and chemical engineering (Piwowar et al., 2018). Different fields of science have different cultures, and common issues are availability of infrastructures, policies and standards, and culture. Astronomy has had a culture of sharing, for example, in part because of limited access to the equipment to conduct observations and experiments (NASEM, 2018c). There is a need for raising awareness within different disciplines about the value of open science. Examples of disciplinary approaches are described in the boxes throughout this chapter, including biological sciences such as genomic research and precision medicine; astronomy and astrophysics; earth sciences; and economics. Regarding funders, the proportion of open publications that are based on research supported by NIH and the Wellcome Trust is high and increasing, which is understandable given their mandates requiring deposit in PubMed Central or Europe PubMed Central (PMC) within 12 and 6 months after publication respectively for all research funded (Bosman and Kramer, 2018; Open Access Oxford, 2018).

The United Kingdom and Austria, through the Universities UK and the Austrian Science Fund respectively, have conducted quantitative studies to monitor the transition to open publication. Universities UK, the representative organization for the United Kingdom's universities (2017), recently found that the proportion of journals published globally with immediate open access increased from under 50 percent in 2012 to over 60 percent in 2016, while the proportion of subscription-only journals has fallen (Universities UK, 2017). The global proportion of articles accessible immediately on publication rose from 18 percent in 2014 to 25 percent in 2016; and the global proportion of articles accessible after 12 months increased from 25 percent to 32 percent (Universities UK, 2017). The Austrian Science Fund—Austria's main public funder of basic research—actively monitors compliance with its open publication mandate (ASF, 2018). The 2017 assessment found that 92 percent of all peer-reviewed publications listed in final reports of ASF-funded projects were openly available (Kunzman and Reckling, 2017).

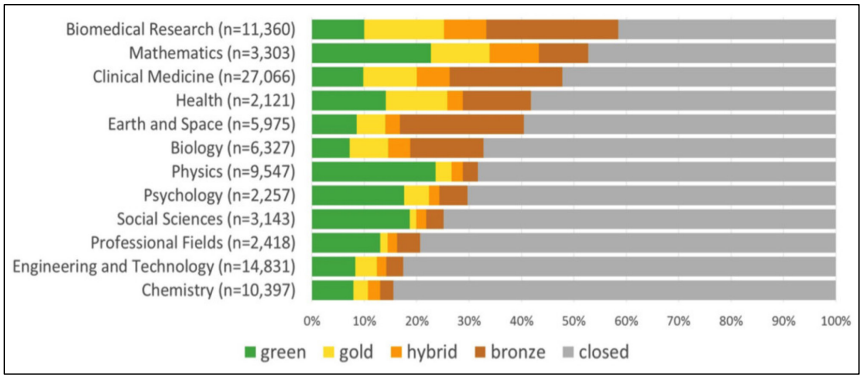


FIGURE 3-2 Percentage of different access types of a random sample of WoS articles and reviews with a DOI published between 2009 and 2015 per NSF discipline (excluding arts and humanities). SOURCE: Piwowar, H., J. Priem, V. Larivière, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J. West, and S. Haustein. 2018. The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ 6:e4375. DOI 10.7717/peerj.4375. Courtesy of Attribution 4.0 International (CC BY 4.0).

Status and trends related to open data and open code are more difficult to track than those related to open publication. In October 2017, Figshare, an open access repository that is part of the Holtzbrinck Publishing Group, released its second *State of Open Data* report (Figshare, 2017). The report includes perspectives from leaders in the open data field and results of a survey of researchers. The survey discovered that 82 percent of nearly 2,300 respondents are aware of open datasets and that 74 percent of their respondents are curating their data for sharing (Figshare, 2017). A global online survey of 1,200 researchers, conducted by the Leiden University and Elsevier in 2017, found that less than 15 percent of researchers share data in a data repository and most (>80 percent) researchers only share data with direct collaborators (Berghmans et al., 2017). In 2017, the International Development Research Centre launched the State of Open Data project, which includes a plan to "critically review the current state of the open data movement" and produce a core reference publication during 2018 (State of Open Data, 2018).

CURRENT APPROACHES TO OPEN SCIENCE

This section explores various approaches to open science, focusing on open publication and open data. Part of the committee's task was to provide illustrations from several scientific disciplines within the biological sciences, social sciences, physical sciences, and earth sciences. The section includes examples drawn from biomedical sciences, economics, astronomy and astrophysics, and earth sciences, along with other examples from outside of those disciplines. A comprehensive assessment of open science within individual disciplines or across disciplines is

beyond the scope of the study. Nonetheless, this overview and the illustrative examples provide insight on how policies, practices, and resources that support open science can be developed and implemented.

Open Publications

Open Access Journals

Open access journals are freely available to readers online "without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself" (Suber, 2015). In contrast to traditional subscription models of scientific publishing, open access publishers typically charge an article processing charge (APC), which is paid by the author or the author's home institution. Open access facilitates free and unrestricted access to articles for everyone immediately after publication (*gold open access*). As described in Chapter 2, less open approaches to publication include *green open access*, in which authors are able to self-archive a version of the article in an open access repository when access to the final published version requires a subscription to the journal. Open publication may also be provided following an embargo period. A list of open access journals in all fields and languages is available in the Directory of Open Access Journals (DOAJ), a community-based online directory launched in 2003 in Sweden with 300 open access journals (DOAJ, 2018). As of March 2018, this number has increased to over 11,100 open access journals, with nearly 2,982,000 articles in 124 countries (DOAJ, 2018).[1]

Although the majority of open access journals do not require APCs, these journals account for a minority of the open access articles published worldwide, and only 18 percent of the open access articles published in the United States (Crawford, 2018). A wide range of APCs is charged by open access journals. For example, F1000 Research charges $150 to $1,000 depending on word count (F1000 Research, 2018). F1000 Research gives discounts or waivers to its referees, advisory board members, and authors from institutions in some developing countries (F1000 Research, 2018).

A successful case of open access publishing is the Public Library of Science (PLOS), a nonprofit scientific organization founded in 2001. PLOS launched its first journal, *PLOS Biology,* in 2003 (see Box 3-1). PLOS publishes several peer-reviewed journals, providing free and unrestricted access to research and an open approach to scientific assessment (PLOS, 2017a). *PLOS One,* a multidisciplinary peer-reviewed journal launched in 2006, had been the largest journal in the world in terms of articles published until 2017, when it was passed by Scientific Reports (Davis, 2017).

[1]DOAJ does not include "hybrid" journals that contain open access and subscription access articles.

BOX 3-1
Public Library of Science (PLOS)

The Public Library of Science (PLOS) is a nonprofit publisher with a mission to accelerate progress in science and medicine by leading a transformation in research communication (Heber, 2017). In 2001, PLOS founders Harold Varmus, Patrick Brown, and Michael Eisen circulated an open letter urging scientific and medical publishers to make published research available through free online public archives, such as the U.S. National Library of Medicine's PubMed Central. Nearly 34,000 scientists from 180 nations signed the letter (PLOS, 2017). In 2001, PLOS became a nonprofit entity and officially became a publisher in 2003, making published scientific and medical articles immediately and freely available online across the globe without restriction.

PLOS rapidly became a key component of the open science movement. In 2003, PLOS launched its first open access journal, *PLOS Biology*. Since then, the organization has introduced six additional peer-reviewed journals, including *PLOS Medicine* in 2004; community journals, *PLOS Computational Biology*, *PLOS Genetics*, and *PLOS Pathogens* in 2005; *PLOS ONE,* the first multidisciplinary open access journal in 2006; and the fourth community journal *PLOS Neglected Tropical Diseases* in 2007. PLOS became financially self-sufficient in 2010 based on the Article Processing Charge model (PLOS, 2017). PLOS also introduced new communications tools, including *The PLOS Blogs Network, PLOS Collections,* and *PLOS Currents,* while publishing over 165,000 articles from authors in 190 countries (PLOS, 2017). PLOS currently partners with protocols.io, in the development of practical tools for PLOS authors to address reproducibility and to gain recognition and credit for their work (Heber, 2017; PLOS Blogs, 2017). PLOS has also been actively engaging early career researchers with social media and live blogging at scientific conferences.

References

Heber, J. 2017. Advocating Open Science at PLOS. Presentation to the National Academies of Sciences, Engineering, and Medicine's Committee on Toward an Open Science Enterprise, Public Symposium. September 18, 2017.

PLOS. 2017. Who We Are. Online. Available at https://www.plos.org/who-we-are. Accessed December 1, 2017.

PLOS Blogs. 2017. Protocols.io Tools for PLOS Authors: Reproducibility and Recognition. Online. Available at http://blogs.plos.org/plos/2017/04/protocols-io-tools-for-reproducibility. Accessed December 4, 2017.

Several entities provide guidelines for assessing the quality of open access journals. DOAJ, in collaboration with the Committee on Publication Ethics (COPE), Open Access Scholarly Publishers Association (OASPA), and World Association of Medical Editors (WAME), identifies principles of transparency and best practice for scholarly publications according to several criteria, such as peer review process, governing body, copyright, ownership and management,

conflicts of interest, revenue sources, etc. (DOAJ, 2018). Publishers or journals that do not meet these criteria will not be included in their publisher's list. Additionally, the Open Access Directory (OAD) provides guidelines, best practices, and recommendations for open access journals (OAD, 2017), while COPE offers resources in the current debates related to promoting integrity in research and scholarly publication (COPE, 2017). OASPA has strict criteria for becoming a member of its organization. The Think, Check, and Submit website provides a checklist for selecting trusted journals (Think, Check, and Submit, 2017).

Some journals exhibit questionable marketing schemes via spam e-mails, perform only cursory peer-review procedures, lack transparency in publishing operations, and imitate legitimate journals (Beall, 2016; Pisanski, 2017). Researchers who are eager to publish or scientists who lack sufficient time to investigate a publisher may submit their papers without verifying a journal's reputability. Beall recommends that scholars read the available reviews and descriptions, and then decide whether they want to submit articles, serve as editors, or serve on editorial boards.

Open Access Repositories

An open access repository is "a set of services that provides open access to research or educational content created at an institution or by a specific research community. Repositories may be comprehensive or may focus on publications or data. They may be institutionally-based or subject-based collections" (COAR, 2015a, p. 3). Lynch (2003) defined the institutional repository as "a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members" (Lynch, 2003, p. 2).

While institutional repositories were developed as a new strategy for universities to accelerate changes in scholarly communication, disciplinary repositories have been established since the early 2000s, often focused on preprints and rapid dissemination of research results. To improve the visibility and impact of research, the majority of open access policies and laws require or request authors to deposit their articles into an open access repository, which has become a key infrastructure component to support these policies. Networked open access repositories enable funders and institutions to track funded research output across repositories, deliver data usage, host collections of academic journals, and link related content across the network (COAR, 2015a). The Confederation of Open Access Repositories (COAR) has developed a roadmap to identify key trends to identify priorities for further investments in interoperability (COAR, 2015b). PubMed Central, managed by the National Library of Medicine, is one of the largest and best-known public access repositories of publications in the biomedical sciences (See Box 3-2).

BOX 3-2
PubMed Central

"As we all know, scientists want their work to be found, read, and cited" (Varmus, 2008).

PubMed Central (PMC), founded in 2000, is a free digital archive of full-text biomedical and life sciences journal articles housed at the U.S. National Institutes of Health's National Library of Medicine (NLM) (NLM, 2018b). The motivation for PMC is to maximize the public investment in NIH-supported research. Articles are submitted to PMC by publishers or directly by authors. PubMed Central is distinct from PubMed, NLM's database of some 27 million citations to the biomedical literature (NLM, 2018a).

In response to a Congressional mandate in 2008 (the Consolidated Appropriation Act of 2008, P.L. 110-161), NIH implemented its Public Access Policy (NIH Public Access Policy, 2016). Since April of that year, authors of NIH-funded research have been required to deposit, or have deposited for them, their final accepted peer-reviewed manuscripts in PMC, with an allowable embargo period of up to 12 months (NIH Public Access Policy, 2016; Varmus, 2008). Francis Collins, responding to a request from Congress in 2011, noted that the public access policy is a "prudent and beneficial" policy for several reasons: It applies 21st century information technology to the NIH investment in the promotion of science and health; it allows NIH to make strategic reasons about its portfolio; and it ensures more rapid progress in science and medical treatments (NIH, 2011).

PMC provides free access to the articles in its database but the majority of the articles, with the exception of those that are already in the public domain, are protected by copyright law. This means that users of the database are subject to the fair use principles of copyright law and cannot, for example, download the entire database for text mining or other purposes. PMC identifies those articles that are open access and provides a service for downloading them, including a filter for the subset of articles that have a CC-BY or CC-0 license. As of January 2018, there are 4.6 million full-text articles from several thousand journals archived in PMC, and some 39 percent (1.8 million) of these are fully open access (NLM, 2018b).

References

NIH (National Institutes of Health). 2011. Francis Collins letter to the Honorable Joseph R. Pitts. Online. Available at https://publicaccess.nih.gov/Collins_reply_to_Pitts121611.pdf. Accessed March 29, 2018.

NLM (National Library of Medicine). 2018a. PubMed. Online. Available at https://www.ncbi.nlm.nih.gov/pubmed. Accessed March 29, 2018.

NLM. 2018b. PubMed Central. https://www.ncbi.nlm.nih.gov/pmc. Accessed March 29, 2018.

NIH Public Access Policy. 2016. NIH Public Access Policy Details. Online. Available at https://publicaccess.nih.gov/policy.htm. Accessed March 29, 2018.

Varmus, H. 2008. Progress toward public access to science. *PLOS Biology*, Apr 8;6(4):e101.

University Open Access Policies

Open access policies have become increasingly adopted in academia. Since 2008, faculties of over 70 universities, schools, and departments have established open access policies to make their publications and research more accessible to policy makers, educators, scholars, and the public (Columbia University, 2017). In 2008, the Harvard Faculty of Arts and Sciences voted unanimously to grant the university a nonexclusive, irrevocable right to disseminate their scholarly articles for non-commercial purpose (Harvard Library, 2017). By June 2014, the remaining eight Harvard schools, including the law school and medical school, adopted similar open-access policies. Scholarly articles provided by Harvard faculty and researchers are stored, preserved, and made available in the Digital Access to Scholarship at Harvard (DASH), a free open access repository available to anyone with internet access. Similarly, Massachusetts Institute of Technology (MIT) faculty voted unanimously in 2009 to make their scholarly articles available free online through DSpace, the open source software created by Hewlett-Packard and the MIT Libraries. Faculty authors may opt out on a paper-by-paper basis (MIT Libraries, 2009). The faculty of the University of California (UC) adopted an open-access policy in 2013. The policy was amended in 2015 to include all researchers employed by the UC. The UC open access policies require that UC faculty and other employees provide a copy of their scholarly articles for inclusion in the eScholarship.org repository, or provide a link to an open version of their articles elsewhere.

A number of guidelines are available to facilitate open access to faculty research and improve scholarly communication. For example, *A SPARC Guide for Campus Action* includes suggestions related to understanding rights as an author and making informed choices about publication venues (SPARC, 2012). Recommendation 4.2 of the 10-year anniversary statement of the Budapest Open Access Initiative (2012) states, supporters of open access "should develop guidelines to universities and funding agencies considering OA [open access] policies, including recommended policy terms, best practices, and answers to frequently asked questions" (BOAI, 2012). As part of the BOAI recommendation, the Harvard Open Access Project (HOAP) released a comprehensive guide, *Good Practices for University Open Access Policies* in 2012 and 2015, based on policies adopted at Harvard University, Stanford University, MIT, and the University of Kansas (Shieber and Suber, eds., 2015). The guide has been endorsed by 15 organizations and projects in the U.S., Europe, and Australia. Similarly, open tools and resources for data management have been promoted in the research library world in the "23 Things: Libraries for Research Data" overview (23 Things, 2018) by the Libraries for Research Data Interest Group of the Research Data Alliance. The overview has been widely disseminated and translated from English into 10 languages.

According to the guide, there are at least six types of university open access policies. Among those types, Shieber and Suber recommend a policy that "provides for automatic default rights retention in scholarly articles and a commitment

to provide copies of articles for open distribution" (Shieber and Suber, eds., 2015., p. 6). To be consistent with copyright law, the guide recommends a policy that "grants the institution certain nonexclusive rights to future research articles published by faculty. This sort of policy typically offers a waiver option or opt-out for authors. It also requires deposit in the repository" (Shieber and Suber, eds., 2015, p.7). However, compliance involving deposits in a repository requires time, which necessitates education, assistance, and incentives. The guide suggests "when the institution reviews faculty publications for promotion, tenure, awards, funding, or raises, it should limit its review of research articles to those on deposit in the institutional repository" (Shieber and Suber, eds., 2015, p. 22). Indiana University-Purdue University Indianapolis (IUPUI) has become one of the first institutions to include open access as a value in its promotion and tenure guidelines, through librarian-facilitated efforts (Odell et al., 2016).

While an effective open access policy can build support for open access, institutions considering adopting their own open access policies are able to refer to the current Harvard model policy (see Box 3-3), which incorporates the latest recommended practices described in their 2015 guide (Shieber and Suber, eds., 2015). To date, over 60 organizations worldwide have adopted a version of the Harvard policy for the development and promotion of open access (Harvard Library, 2017). Internationally, the Registry of Open Access Repository Mandates and Policies (ROARMAP) lists over 200 open access mandates and policies adopted by universities, research institutes, and research funders across the globe (ROADMAP, 2017). In addition to the policy guidelines published by the United Nations Education, Scientific and Cultural Organization (UNESCO) (Swan, 2012) and Mediterranean Open Access Network (MedOANet, 2013), the European University Association (EUA) provides a practical guide for universities in the context of current European open access policies (EUA, 2015).

Preprints

A preprint is defined as "a complete written description of a body of scientific work that has yet to be published in a journal" (Bourne et al., 2017). Preprints can be the complete and original manuscripts of scientific documents, including a research article, review, editorial, commentary, and a large dataset. that are not yet certified by peer review. Preprint servers can also host other objects such as posters presented at scientific meetings. The purpose of preprint distribution is "to share the results of recent research freely and openly before they are certified by peer review, in a manner that permits immediate discovery and discussion of the results and feedback to authors from the research community at large" (Inglis, 2017).

Providing preprint services is not without costs. For large services such as arXiv and bioRxiv, extensive hardware and software infrastructure is required. Although articles are not peer reviewed, they are screened and categorized, which requires staffing. Costs are typically covered by the host institutions and by foundation grants.

BOX 3-3
A Model Open Access Policy

The Faculty of <*university name*> is committed to disseminating the fruits of its research and scholarship as widely as possible. In keeping with that commitment, the Faculty adopts the following policy: Each Faculty member grants to <*university name*> permission to make available his or her scholarly articles and to exercise the copyright in those articles. More specifically, each Faculty member grants to <*university name*> a nonexclusive, irrevocable, worldwide license to exercise any and all rights under copyright relating to each of his or her scholarly articles, in any medium, provided that the articles are not sold for a profit, and to authorize others to do the same. The policy applies to all scholarly articles authored or co-authored while the person is a member of the Faculty except for any articles completed before the adoption of this policy and any articles for which the Faculty member entered into an incompatible licensing or assignment agreement before the adoption of this policy. The Provost or Provost's designate will waive application of the license for a particular article or delay access for a specified period of time upon express direction by a Faculty member.

Each Faculty member will provide an electronic copy of the author's final version of each article no later than the date of its publication at no charge to the appropriate representative of the Provost's Office in an appropriate format (such as PDF) specified by the Provost's Office.

The Provost's Office may make the article available to the public in an open-access repository. The Office of the Provost will be responsible for interpreting this policy, resolving disputes concerning its interpretation and application, and recommending changes to the Faculty from time to time. The policy will be reviewed after three years and a report presented to the Faculty.

SOURCE: S. M. Shieber, 2015.

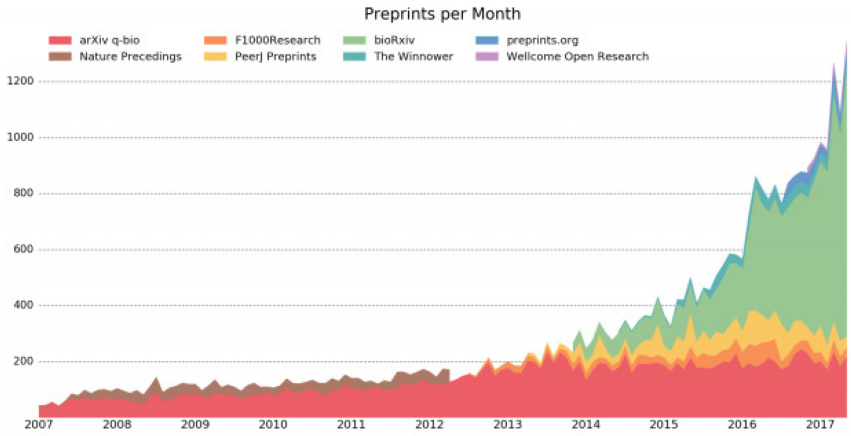


FIGURE 3-3 Biology preprints over time. SOURCE: http://asapbio.org/preprint-info/biology-preprints-over-time. Courtesy of Attribution 4.0 International (CC BY 4.0).

Preprints are gaining momentum among the scientific community. Since 1991, researchers in disciplines such as physics (and later mathematics, computer science, and quantitative biology) have been able to access preprints through arXiv, a repository of electronic preprints of scientific papers. arXiv is operated by the Cornell University Library and currently contains over 1.3 million preprints (Cornell University Library, 2017). In 2013, bioRxiv was launched as a repository of life science preprints covering all of the life sciences, clinical trials, epidemiology, as well as science communication and education (see Figure 3-3). Operated by the Cold Spring Harbor Laboratory, bioRxiv is modeled conceptually on arXiv but uses different technology, and offers somewhat different features and functions (Inglis, 2017). Economics has a long history of utilizing preprints, which are called working papers in that discipline (See Box 3-4).

Preprint services are being launched in a growing number of disciplines, as indicated in Table 3-1. For example, the American Chemical Society (ACS) and its global partners launched ChemRxiv, a preprint server for chemistry-related information. The Center for Open Science (COS) has launched PsyArXiv (psychology), AgriXiv (agriculture), SocArXiv (social sciences), engrXiv (engineering), and LawArXiv (law), with the most recent additions including NutriXiv (nutritional sciences) and SportRxiv (sport) (COS, 2017; Luther, 2017). In 2017, the American Geophysical Union and Atypon announced the development of Earth and Space Science Open Archive (ESSOAr). This preprint server will join the existing EarthArXiv as preprint servers for the earth and space science community (Voosen, 2017).

There are other services that provide preprint functions. For example, the Social Science Research Network (SSRN) was created in 1994 as a tool for rapid dissemination of scholarly research in the social sciences and humanities. The SSRN, bought by Elsevier in 2016, facilitates the free posting and sharing of research material, including preprints, conference papers, and non-peer-reviewed papers in social science research (Gordon, 2016). F1000Research is "an open research publishing platform for life scientists that offers immediate publication and transparent peer review" (F1000Research, 2018). An article submitted to F1000Research also requires data and code deposition, either in an F1000 approved repository or in an institutional repository.

Bourne et al. (2017) described a number of advantages of preprint submission from the standpoint of both individual researchers and the broad community. Preprints are free to post and to read, which provides accelerated transmission of scientific results. Researchers can evaluate new findings and their reliability without the delay introduced by journal peer review. Some funders are now providing incentives to those who submit preprints (Inglis, 2017). However, there are challenges associated with managing preprints, including anxieties about "scooping" (other researchers using the preprint to publish work in advance of those submitting a preprint) and reluctance to use open licenses (Inglis, 2017; INLEXIO, 2017). There is a need for more education and discussion regarding the choice of licenses and ways to prevent unattributed use of the results. NIH is working with

an international group of research funders to examine the feasibility of establishing a central service of preprints to encourage sharing of preprints in the life sciences (NIH, 2017b).

BOX 3-4
Working Papers in Economics

The National Bureau of Economic Research (NBER) issued its first working paper (preprint) in 1973 as a way of disseminating research more quickly than waiting for lengthy editorial review at the Bureau. The papers were originally mailed to libraries, research institutes, journalists, and other interested parties on a subscription basis; over time, print distribution has given way to electronic dissemination. As of October 2017, approximately 24,000 working papers have been issued by the NBER. These working papers reside behind a pay wall for 18 months and then are provided freely to the international research community (green open access). For residents of developing countries, journalists, and government employees, there is no pay wall, even for new papers. NBER research associates are leading economics researchers and not necessarily representative of the entire economics profession. Today, there are nearly 1,500 NBER-affiliated researchers. Only NBER research associates and conference participants are allowed to release NBER working papers. Nevertheless, as the thought leader in the profession, NBER created a culture of openness for the economics profession that has had a lasting impact.

Economists outside of the NBER recognized the need to disseminate research prior to publication. In 1993, the Economics Working Paper archive was opened at Washington University in St. Louis. In 1997, Research Papers in Economics (RePEc) was created to facilitate the sharing of economic research (http://repec.org). RePEc is a "decentralized bibliographic database of working papers, journal articles, books, books chapters and software components, all maintained by volunteers." According to RePEc, 1,900 archives from 93 countries have contributed 2.3 million research pieces to the archive. Although economics journals have pay walls, the free availability of working papers means that almost all economics research is open. Any economist can register and maintain an author profile at RePEc and as of 2017, more than 50,000 authors have registered worldwide.

Economics journals have supported replication and hosted data archives for almost 30 years. As a condition of acceptance, the *Journal of Human Resources* required authors to preserve data for 3 years after publication in order to promote replication starting in 1989. *The Journal of Applied Econometrics* has data archives for most papers starting in 1995 (http://qed.econ.queensu.ca/jae). *The American Economic Review* and other American Economic Association journals required data archiving starting in 2004. *The American Economic Review* has hired a data editor to ensure the proper archival of datasets and software programs, and to consider exceptions to the data archival policies for restricted use datasets. In 2007, the American Economic Association launched four field journals in part to reduce the influence of for-profit journals in the profession.

TABLE 3-1 Preprint Servers

Name	Fields	Start Year	Owned/Operated by	Submissions in 2016
Selected preprint services				
arXiv	Physics, mathematics, computing, quantitative biology, quantitative finance, statistics	1991	Cornell University Library	113,308
bioRxiv	Life sciences	2013	Cold Spring Harbor Laboratory	4,712
PeerJ Preprints	General	2013	PeerJ	~1,000
Preprints (MDPI)	General	2016	Multidisciplinary Digital Publishing Institute (MDPI)	~1,000
SocArXiv	Social sciences	2016	Open Science Framework (OSF)	633
PsyArXiv	Psychology	2016	OSF	191
engrXiv	Engineering	2016	OSF	35
ChemRxiv	Chemistry	2017	ACS	N/A
AgriXiv	Agriculture	2017	OSF	N/A
EarthArXiv	Earth Sciences	2017	OSF	N/A
LawArXiv	Law	2017	OSF	N/A
NutriXiv	Nutritional Sciences	2017	OSF	N/A
Sport RXiv	Sport science	2017	OSF	N/A
Services with preprint functions				
Social Science Research Network (SSRN)	Social sciences	1994	Elsevier	66,310
Figshare	General	2012	Figshare	Unknown
Zenodo	General	2013	OpenAire/CERN	318
F1000Research	General	2013	F1000Research	215
Authorea	General	2013	Authorea	Unknown

SOURCE: https://www.inlexio.com/rising-tide-preprint-servers; https://researchpreprints.com/2017/03/09/a-list-of-preprint-servers.

European Commission Open Research Publishing Platform

The European Commission (EC) has proposed to fund the EC Open Research Publishing Platform for Horizon 2020 beneficiaries to comply with the Horizon 2020 open access mandate and to increase open access peer reviewed publications in Horizon 2020 (EC, 2017c). The platform will provide an easy, fast, and reliable open access publishing venue free to Horizon 2020 grantees on a voluntary basis, including preprints support, open access, open peer review, and innovative research indicators most appropriate for individual disciplines and/or national context. Building on the best practices of other funders, such as the Bill & Melinda Gates Foundation and the Wellcome Trust, the commission hopes that the platform will contribute to a more diversified and competitive open access publishing market. One contractor or a consortium led by one contractor will be selected to run the platform with a 4-year initial contract. The contractor will be required to commit to a minimum number of preprints and articles to be published during the initial 4-year period and to develop a plan for sustainability of the service beyond the 4 years. While some experts such as Jacobs (2018) interpret this movement as "a sign of increasing frustration on the part of research funders and institutions at the pace and cost of the change to open access," the success of the platform will depend on the quality of the scientific publication service provided (EC, 2017c). Current international approaches to open science are described further in the final section of this chapter.

Pay It Forward Initiative

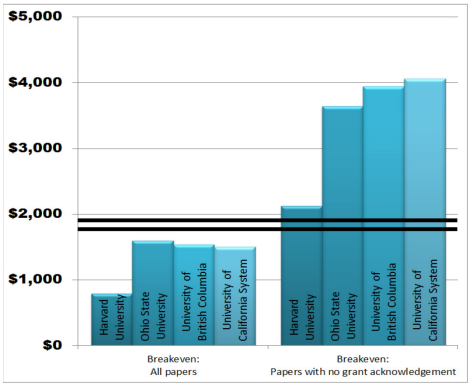
The University of California (UC), Davis and the California Digital Library (CDL) conducted a study in 2015 and 2016 to examine the economic implications of large North American research institutions converting to an entirely article processing charge (APC) business model. With support from the Andrew W. Mellon Foundation, the study was conducted in partnership with Harvard University, Ohio State University, the University of British Columbia, and University of California Libraries, along with the Association of Learned and Professional Society Publishers (ALPSP) and the private sector, including Thomson Reuters (Web of Science) and Elsevier (Scopus). These large North American research institutions would assume the large part of the financial burden in an APC-driven open access model, the predominant open access business model of gold open access publishers (Anderson, 2017). The study involved a number of qualitative analyses based on academic author surveys and publisher surveys, as well as quantitative analyses based on data for a 5-year period (2009-2013), including library subscription expenditures, university publishing output, and potential APCs (UC Libraries, 2016; Anderson, 2017).

The final report, *Pay It Forward: Investigating a Sustainable Model of Open Access Article Processing Charges for Large North American Research Institutions* (2016), has the following three major findings:

1. The total cost to publish in a fully APC-funded journal will exceed current library journal budgets for the most research-intensive North American research institutions;
2. This cost difference could be covered by grant funds, already a major source of funding for publishing fees; but
3. Ultimately, author-controlled discretionary funds, such as research grants and personal research accounts that incentivize authors to act as informed consumers of publishing services, are necessary to introduce both real competition and pricing pressures into the journal publishing system (UC Libraries, 2016).

To establish these findings, the study examined the level of APCs each institution could afford, based on its current subscription spending. The study discovered that the average APC for partner institution publications in full open access journals is $1,892 (Figure 3-4). While research-intensive institutions would be unable to convert to the APC model if they had to rely solely on their existing subscription budgets, the study found that those institutions could afford a transition to APC, if grant funds were applied to the cost (Anderson, 2017). This is not an entirely novel idea, as many authors are already using grant funds for APCs. A key strategy could be a multi-payer model involving library subsidies, together with grants, startup packages, and discretionary research funds (Anderson, 2017). For example, the Wellcome Trust notes that its APC payments, which cover both full open access and high-cost hybrid journals, consume less than 1 percent of its overall research budget (UC Libraries, 2016; Anderson, 2017). According to the report, incorporating grant and discretionary funds into the financial flow for a full APC business model may be a viable direction for both research-intensive institutions and their funders.

The report emphasizes that it is essential to introduce competition for authors to ensure that APCs remain affordable in the future. This can be accomplished by giving authors some financial responsibility in deciding where to publish, using funds that they control directly. Additionally, the report acknowledges that the information available on current APCs is almost entirely derived from STEM fields, which historically have higher subscription costs than social science and humanities disciplines (Crotty, 2016; UC Libraries, 2016). Because the report provides APC estimations based on available data, it likely overestimates costs for non-STEM fields, and additional analysis may be needed for other disciplines. There is also a need to monitor global developments on an ongoing basis to assess opportunities for collaboration with European countries toward more immediate, large-scale transition to an open science enterprise.



← **$1,892:** Average APC for partner institution publications in full open access journals

FIGURE 3-4 APCs are affordable for large research-intensive institutions if grant funds are applied. SOURCE: Presentation by Ivy Anderson, California Digital Library, Committee on Toward an Open Science Enterprise public symposium, September 18, 2017.

A recent report from the Max Planck Digital Library (MPDL) has claimed that a large-scale open access transformation is possible without financial risk (MPDL, 2015). Yet this is a contentious issue. Some argue against efforts to promote publishing models based on gold open access enabled by APCs, and instead advocate for a combination of green open access mandates and community efforts to create and sustain new institutions for publishing and expert review (Shulenberger, 2016).

As a recent development, the UC Libraries released a new report, *Pathways to Open Access,* in February 2018 that identifies the current state of open access approaches, a set of strategies to achieve those approaches, and possible next steps to assist UC campus libraries and the California Digital Library to pursue a large-scale transition to open access (UC Libraries, 2018). An accompanied published chart summarizes those approaches and strategies identified in the report, including green open access, gold open access-APC based, gold open access-non APC based, and universal strategies.

Private Foundation Initiatives

Open access publishing has increasingly become part of the business process among the philanthropic community. For example, the Bill & Melinda Gates Foundation has one of the most stringent open-access policies. After a 2-year transition period for policy compliance, the foundation's Open Access Policy has been fully operational as of January 1, 2017, with no exceptions to the policy (Bill & Melinda Gates Foundation, 2017; Hansen, 2017; Adams, 2018). Under its policy, the foundation requires grantees to make their research papers and data available immediately upon publication without any embargo period and allow for their unrestricted use under the Creative Commons Attribution Generic License

(CC BY 4.0) or an equivalent license (Bill & Melinda Gates Foundation, 2017; Hansen, 2017; Adams, 2018). The foundation will pay reasonable fees in order to publish on its open access terms. Launched in July 2016, the web-based service Chronos tracks the impact of research while simplifying research publishing. As a new initiative, Gates Open Research was launched in late 2017, with a model used by the Wellcome Trust in the United Kingdom (Wellcome Open Research), to provide their grantees with an open research platform for open peer review and rapid author-led publication (Butler, 2017; Open Research Central, 2017; Van Noorden, 2017; Bill & Melinda Gates Foundation, 2018). As one of the most influential global health philanthropic organizations, the foundation emphasizes that "the free, immediate, and unrestricted access to research will accelerate innovation, helping to reduce global inequity and empower the world's poorest people to transform their own lives" (Bill & Melinda Gates Foundation, 2017). Because of a rapidly changing landscape in scholarly communications, the Wellcome Trust will conduct its first review of its open access policy and a result will be announced by the end of 2018 (Wellcome Trust, 2018).

While a growing number of funding organizations are committing to open sharing of research, the funder community is building effective partnerships in an effort to meet current and future open science challenges. One major effort is the creation of the Open Research Funders Group (ORFG)[2] in December 2016, following a forum of open access stakeholders convened by The Robert Wood Johnson Foundation and the Scholarly Publishing and Academic Resources Coalition (SPARC) in late 2015. The ORFG develops actionable principles and policies that encourage innovation, increase access to research articles and data, and promote reproducibility (ORFG, 2018). While many organizations have expressed an interest in developing their own open policies, a significant challenge is the lack of clarity about an effective policy. In an attempt to describe the variation in interpretation of openness by funding organizations, ORFG has published a guide, *HowOpenIsIt? Guide to Research Funder Policies* (2017), building on the success of *HowOpenIsIt? Guide for Evaluating the Openness of Journals* described in Chapter 2 (see Table 2-2). During recent infectious diseases outbreaks in 2016, the publishing community largely agreed, at the prompting of WHO and funders such as the Bill & Melinda Gates Foundation and the Wellcome Trust, to adopt open science practices, including early publication of data and preprints and open access publication (PLOS, 2016). Such agreements applied in times of international public health emergencies underscore the benefits of an open science approach.

[2]As of January 2018, ORFG members include the Alfred P. Sloan Foundation, American Heart Association, A Charitable Fund of Peter Baldwin and Lisbet Rausing (ARCADIA), the Bill & Melinda Gates Foundation, Eric & Wendy Schmidt Fund for Strategic Innovation, James S. McDonnell Foundation, John Templeton Foundation, Laura and John Arnold Foundation, Leona M. and Harry B. Helmsley Charitable Trust, Open Society Foundation, Robert Wood Johnson Foundation, and Wellcome Trust. Additional information can be found at http://www.orfg.org/members.

Publisher and Society Initiatives

Publishers and professional societies are exploring options for expanding open access to accelerate scientific discovery. The American Geophysical Union (AGU), which consists of 60,000 members from 137 countries, is the largest society publisher in the discipline of Earth and space science with 20 peer-reviewed scholarly journals and over 6,000 published papers in 2016 (Stall, 2017). The AGU produces four open access journals, including *Journal of Advances in Modeling Earth Systems*, *Earth's Future*, *Earth and Space Science*, and *GeoHealth*, with content currently representing nearly 100,000 articles (AGU, 2017a). Articles published in those journals become freely available immediately online upon publication, and authors can select one of several Creative Commons (CC) licenses. AGU allows a draft or the author's version of the accepted manuscript to be posted to any nonprofit preprint server to encourage community engagement. Through its publishing partner Wiley, AGU offers discounts or waivers on fees from researchers in developing countries to increase access to research. Additionally, AGU is part of the innovative Research4Life program, which provides over 5,000 institutions in low- and middle-income countries free or low-cost access (AGU, 2017a; Research4Life, 2018). In addition to these gold open access options, AGU also makes all publications open after a 2-year embargo period. (See Chapter 2 and above for more explanation on gold and green access.)

Open Data

Most research data in repositories today is not available under FAIR principles. Realizing this availability will entail significant costs and complexities. The wide variety of types and sizes of research datasets means that developing effective tools and practices will require significant and sustained community input. Long-term curation of data and research software will require standards for the types of data that should be stored and how long they should be stored. This section considers several examples and potential lessons.

Big Science Data

Open data is largely the norm in fields such as high-energy physics and astronomy, as funding for these projects is significant, and as such data distribution is well thought out and closely monitored by the respective federal agencies. Good examples include the Large Hadron Collider, and some of the large scale astrophysical archives (Hubble Legacy Archive, Sloan Digital Sky Survey, etc.). They typically started in areas where the data were far removed from any financial impacts. More recently data from other areas, like genomics (Human Genome Project, 1000 genomes, etc.) and material science (Material Genome Initiative) are also heading towards data sharing in large open archives. Such a transition for a given field typically requires a decade of focused effort by the community, and

a substantial federal investment. Boxes 3-5 and 3-6 illustrate examples of open practices in the fields of astronomy and astrophysics as well as genomics research, respectively. With the size and complexity of datasets continually increasing, yesterday's "big data" appears less big today, today's "big data" will appear small in five or ten years, and so forth.

BOX 3-5
Astronomy and Astrophysics

The Sloan Digital Sky Survey (SDSS) has been one of the largest, most detailed, and most often cited surveys in the history of astronomy. The SDSS has provided world-leading datasets for a wide range of astrophysical research, including the study of extragalactic astrophysics, cosmology, the Milky Way, and stars (ARC, 2012). All SDSS data are released to the public under open science principles. The SDSS project has revolutionized the interactions between a telescope, its data, and its user communities (NAS-NAE-IOM, 2009).

There was a desire to develop large-scale (petascale) computing and storage to enable greater access and better usability of information by the astronomy and physics community. The Astrophysical Research Consortium (ARC) was formed in 1984, and a pioneering 2.5-meter telescope was created at Apache Point Observatory (APO) in New Mexico that maps the sky to examine the structure of the universe (NAS-NAE-IOM, 2009). To accelerate discoveries in astronomy, the SDSS was initiated to "digitally map about half of the Northern sky in five spectral bands from ultraviolet to the near infrared" (Szalay, 2000). However, the data challenge in this field was the integration of disparate types of data about astronomical objects (stars, galaxies, quasars), including images, spectroscopy data, and astrometric data, along with the large volumes of data (2 to 4 TB per year) (NRC, 2008).

After nearly a decade of design and construction, the SDSS entered routine operations in 2000. With funding from multiple sources and countries, the SDSS has been releasing data annually at the American Astronomical Society Meeting. The data obtained from the project are available at SkyServer, an SDSS-managed public database designed and built at Johns Hopkins University, for both astronomers and for science education (Gatlin, 2013). Anyone with a web browser can navigate through the sky using the SkyServer website. Teachers are encouraged to adapt the projects for use in their classroom.

Since 2000, SDSS has progressed through the following phases with multiple surveys:

- SDSS I (2000–2005), including deep multicolor imaging over 8,000 square degrees and measured spectra of more than 700,000 celestial objects.
- SDSS II (2005–2008), including the Sloan Supernova Survey. SDSS II completed the original survey goals of imaging half the northern sky and mapping the 3-dimensional clustering of one million galaxies and 100,000 quasars.

(Continued)

BOX 3-5 Continued

- SDSS III (2008–2014), including the Apache Point Observatory Galactic Evolution Experiment (APOGEE) and Baryon Oscillation Spectroscopic Survey (BOSS) using the largest-ever, three-dimensional map of distant galaxies.
- SDSS IV (2014–2020), including the extended BOSS (eBOSS), APOGEE-2, and Mapping Nearby Galaxies at APO (MaNGA) (SDSS, 2017).

While SDSS has recorded a total of 25 TB data during the first (2000–2005) and second surveys (2005–2008) combined, the amount of data to be saved at the end of the third survey (2008–2014) is 100 TB due to the multiple reprocessing versions of the data (Singh and Kumar, 2016). The SDSS is distinctive within the astronomical community for its participatory, bottom-up scientific research planning process, currently involving over 50 contributing institutional members in the collaboration. For the first time in the collaboration's history, the current fourth phase of SDSS (SDSS-IV) partners with a sister telescope located in the Southern hemisphere in Chile to observe regions of the sky that are not visible from the Northern hemisphere (Alfred P. Sloan Foundation, 2017). In keeping with previous SDSS policy, the SDSS-IV provides regularly scheduled public data releases, and the current version is Data Release 14. The website for each of the SDSS I, II, and III is still available but no longer updated.

All SDSS data are available through public archives and used extensively by the community for research and teaching. For example, there are more than 7,000 refereed papers published, with well over 350,000 citations (Szalay, 2014). Citizen-science projects, such as Galaxy Zoo, invite the general public to help classify millions of galaxies in the SDSS data via the Internet (Lincott et al., 2008; Khullar, 2017), and led to the discovery of a unique celestial object by a Dutch school teacher. Next generation large astronomical surveys, such as the Large Synoptic Survey Telescope (LSST) and Panoramic Survey Telescope and Rapid Response System (Pan-STARRS), have also used the SDSS experience to develop their own data management infrastructure and services (Szalay, 2014). The SDSS has contributed to the globalization of scientific innovation through open science. The SDSS is managed by the Astrophysical Research Consortium for the participating institutions of the SDSS collaboration. Funding for the current SDSS IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the participating institutions (SDSS, 2017).

References

Alfred P. Sloan Foundation. 2017. Sloan Digital Sky Survey. Online. Available at https://sloan.org/programs/science/sloan-digital-sky-survey. Accessed November 13, 2017.

ARC (Astrophysical Research Consortium). 2012. Principles of Operation for SDSS-IV. Online. Available at http://www.sdss.org/wp-content/uploads/2014/11/principles.sdss4_.v4.pdf. Accessed November 15, 2017.

(Continued)

BOX 3-5 Continued

Gatlin, L. 2013. Johns Hopkins astronomer awarded $9.5M to create 'virtual telescope.' Johns Hopkins University. Online. Available at https://hub.jhu.edu/2013/11/01/szalay-grant-skyserver. Accessed November 15, 2017.

Khullar, G. 2017. The Sloan Digital Sky Survey: A Legacy. Online. Available at https://astrobites.org/2017/02/03/the-sloan-digital-sky-survey-a-legacy. Accessed November 13, 2017.

Lintott, C. J., K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg. 2008. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389:1179-1189.

NAS-NAE-IOM (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine). 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington, DC: The National Academies Press.

NRC (National Research Council). 2008. *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security*. Washington, DC: The National Academies Press.

NSF (National Science Foundation). 2014. SciServer: Big Data infrastructure for science. Online. Available at https://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=133526. Accessed November 16, 2017.

SDSS (Sloan Digital Sky Survey). 2017. The Sloan Digital Sky Survey: Mapping the Universe. Online. Available at http://www.sdss.org. Accessed November 13, 2017.

Singh, M. K., and G. D. Kumar. 2016. *Effective Big Data Management and Opportunities for Implementation*. Hershey, PA: IGI Global.

Szalay, A. S. 2017. From SkyServer to SciServer. *The Annals of the American Academy of Political and Social Science* 675(1):202-220.

Szalay, A. S., P. Kunszt, A. Thakar, J. Gray, and D. Slutz. 2000. The Sloan Digital Sky Survey and its Archive. Online. Available at https://arxiv.org/abs/astro-ph/9912382v1. Accessed November 16, 2017.

A major consideration is what happens to data from a major research facility, which often takes hundreds of millions of dollars and decades of effort, once the facility is shut down (e.g., BaBaR at SLAC[3]). The legacy value of the investments made remain in the data, which need to be preserved and curated for at least several additional decades. This preservation phase of the data lifecycle requires skills different from those needed for capturing and analyzing data from an active instrument. Several major facilities are getting closer and closer to this point.

[3]BaBaR is a large-scale particle physics experiment conducted at the SLAC National Accelerator Laboratory and designed to study fundamental questions about the universe, including the nature of antimatter, the properties and interactions of the particles known as quarks and leptons, and searches for new physics. For more information, see http://www-public.slac.stanford.edu/babar.

Maintaining and reinventing the data curation for each project in isolation will be very inefficient, and the task requires economies of scale. The expertise for curation will require active involvement by librarians and archivists, augmenting the legacy and corporate knowledge of the individual projects.

The Long Tail of Science

The long tail of science is increasingly gaining attention in the open science community. While big data tend to comprise homogeneous, standardized, and regulated data, long-tail data can be relatively small and heterogeneous individually but very large in the number of datasets (Heidorn, 2008; Borgman, 2015; e-IRG, 2016; see Table 3-2). Data heterogeneity includes differences in the size, structure, format, and complexity of research data.

Long tail data exist across all disciplines, mostly only in individual computers or personal websites with minimal or no attached metadata or documentation, resulting in issues such as irreproducibility of research, duplicate research, and, potentially, innovation loss (e-IRG, 2016). For example, environmental science research involves enormous complexity of its datasets, including physical, chemical, and biological data that reside in small files (e.g., spreadsheets and tables) collected in laboratories (Szalay, 2014). Other challenges associated with long-tail data include data quality due to varying technology across disciplines, difficulty of discoverability in diverse repositories, and lack of incentives for researchers to deposit their data. Mostly, the demands for metadata are simply too cumbersome for normal scientists, who feel that the relatively small amounts of data to be published do not justify the effort that needs to be spent to add the required extra information for the publishing process. Part of the reason for the balkanization of long-tail data is its isolation/geographic segregation. Most of such data sit on tens of thousands of personal computers, or personal websites. If all data could be stored on the same "science cloud," where it would take a mouse click to upload and link new information, a complex network of interrelated datasets could rapidly be built. It is quite likely that the relationships between datasets would resemble the network graphs of co-authorship. The technology to do automatic discovery of a wider context from data tables on the web is already here (Cafarella et al., 2008). A substantial amount of data currently resides in "Supplementary Information" accompanying journal articles—in front or behind paywalls, but mostly in formats that do not lend themselves to text- or data-mining. Several publishers are currently moving towards ensuring at least one copy of article-related datasets is available in open repositories (e.g., Dryad, Figshare), as well as in the journal record (COPDESS, 2015; Byrne, 2017).

BOX 3-6
Genomic Data

The Human Genome Project was a large-scale project to determine the sequence of the human genome. The project successfully created a human reference genome, together with the complete sequences of five model organisms (The Human Genome Project Completion). The work was coordinated by the National Institutes of Health and the U.S. Department of Energy and involved a large interdisciplinary team, with participating laboratories in the U.S. and abroad (Collins et al., 1998; Lander et al., 2001; Hood, 2013). The goals of the project were first set forth in 1988 by a committee of the U.S. National Academy of Sciences (NRC, 1988). Among the goals articulated by the Academy report and in subsequent publications by the leaders of the effort was a significant focus on open data sharing: "Considerable data will be generated from the mapping and sequencing project. Unless this information is effectively collected, stored, analyzed, and provided in an accessible form to the general research community worldwide, it will be of little value" (NRC, 1988, p. 7), and "Collection, analysis, annotation, and storage of the ever increasing amounts of mapping, sequencing, and expression data in publicly accessible, user-friendly databases is critical to the project's success" (Collins et al., 1998, p. 688). The National Library of Medicine's National Center for Biotechnology Information (NCBI) was founded in 1988, and since then it has built and maintains numerous publicly available genomic databases for use by scientists and the interested public (NCBI). The Human Genome Project has fostered not only an interdisciplinary culture, involving collaborations among computer scientists, engineers, mathematicians, and biologists, but also a culture in which data and computational code are openly and freely shared (Lander et al., 2001; Hood, 2013; Cook-Deegan, 2017).

The Personal Genome Project (PGP) was founded in 2005 and is dedicated "to creating public resources that everyone can access" and to a "highly participatory approach to research-participant communication and interaction" (Church, 2005; Harvard Personal Genome Project, 2014; Ball et al., 2014). The project enrolls volunteers who are interested in publicly sharing their genomic, health, and trait data for the benefit of scientific progress. Acknowledging that it is not possible to guarantee privacy, confidentiality, and anonymity of genetic data when the explicit goal is to share those data, the project has developed a novel "open consent" framework. Because the PGP aims to have all of its participants both engaged and informed, potential participants are given a study guide that provides a primer on genomic science and discusses the risks of participating, after which they must pass an exam testing their understanding of the material (Angrist, 2009). As of January 2018, the project has enrolled more than 5,000 participants.

(Continued)

BOX 3-6 Continued

The NIH updated its genomic data sharing policy in late 2014. The policy details the agency's expectations for the sharing of both human and non-human genomic data generated by studies supported by the NIH (NIH, Genomic Data Sharing). Data generated from human studies must be submitted to the NIH generally within 3 months after generation, and the NIH may allow another 6-month embargo period before public release. In addition, the policy requires that investigators obtain participants' consent to share their data broadly for future research purposes.

References

Angrist, M. E. 2009. Wide open: The personal genome project, citizen science and veracity in informed consent. *Personalized Medicine* 6(6):691-699.

Ball, M. P., J. R. Bobe, M. F. Chou, T. Clegg, P. W. Estep, J. E. Lunshof, W. Vandewege, A. W. Zaranek, and G. M. Church. 2014. Harvard Personal Genome Project: Lessons from participatory public research. *Genome Medicine* 6(2):10-16.

Church, G. M. 2005. The Personal Genome Project. *Molecular Systems Biology* 1(1):0030.

Collins, F. S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters. 1998. New goals for the U.S. Human Genome Project: 1998-2003. *Science* 282(5389):682-689.

Contreras, J. L. 2015. NIH's genomic data sharing policy: timing and tradeoffs. *Trends in Genetics* 31(2):55-57.

Cook-Deegan, R., R. A. Ankeny, and K. Maxson Jones. 2017. Sharing Data to Build a Medical Information Commons: From Bermuda to the Global Alliance. *Annual Review of Genomics and Human Genetics* 18:389-415.

Hood L, and L. Rowen. 2013. The Human Genome Project: Big science transforms biology and medicine. *Genome Medicine* 5(9):79.

Lander et al. and the International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860-921.

NCBI (National Center for Biotechnology Information). Online. Available at https://www.ncbi.nlm.nih.gov. Accessed March 30, 2018.

NIH (National Institutes of Health). 2010. The Human Genome Project Completion. Online. Available at https://www.genome.gov/11006943. Accessed March 30, 2018.

NIH. NIH Genomic Data Sharing. Online. Available at https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing. Accessed March 30, 2018.

NRC (National Research Council). 1988. Report of the Committee on Mapping and Sequencing the Human Genome. Washington, DC: The National Academies Press.The Harvard Personal Genome Project. Online. Available at https://pgp.med.harvard.edu/about. Accessed March 30, 2018.

TABLE 3-2 Big Data vs. Long-Tail Data

	Big Data	Long-Tail Data
1	Homogeneous	Heterogeneous
2	Large	Small
3	Common standards	Unique standards or no standards
4	Regulated	Not regulated
5	Central curation	Individual curation
6	Disciplinary repositories	Institutional, general or no repository

SOURCE: e-IRG, 2016.

Discovering, transforming and reusing data collected by others has become a major part of science today, yet the process is still painful. The Research Data Alliance (RDA) and the National Data Service (NDS) are leading the way in the path towards establishing a universal, easy-to-use data publishing and management framework, but this is an area that will require consistent long-term attention before it can be said that the problem has been solved (See Box 3-7). Clearly, scientists can learn from best practices in industry, but those techniques need to be carefully tailored to the specific needs of science (assessing data quality, refereeing process, relation to publications, easy attribution, tracking provenance).

A number of initiatives address challenges involved in managing long-tail data. For example, the RDA's Long Tail of Research Data Internet Group, launched in 2013 with over 90 members from around the world, has developed a set of good practices for managing research data archived in the university context (RDA, 2017a). The European Library Federation (LIBER) released 10 recommendations for libraries to get started with research data management (LIBER, 2012); the Confederation of Open Access Repositories (COAR) issued the repository Interoperability roadmap (COAR, 2014); and the Open Access Infrastructure for Research in Europe (OpenAIRE) links literature to data. Additional work is needed to establish a relevant, operational ecosystem for the long tail of science during the implementation of international, national, and local e-infrastructures, possibly using automated techniques to extract the metadata needed for discovery and indexing (Cafarella, 2008). While reuse of data remains highly dependent on discipline- and data-specific metadata, which have long been recognized as critical for reuse (Brazma et al., 2001), support for researchers willing to invest time and efforts in establishing such standards is also critical.

Scientific Collections and Sample Preservation

While much of this report focuses on digital research products, a significant percentage of research effort continues to involve collection, analysis, and use of physical specimens and materials. Metadata about specimen collections may or may not be available in digital form online.

BOX 3-7
The National Data Service

The National Data Service (NDS) is "an emerging vision for how researchers and scientists across all disciplines can find, reuse, and publish data" (NDS, 2017a). While many scientific communities are increasingly developing discipline-specific data services, the U.S. and international communities lack a unified open framework for storing, sharing, and publishing data that can be used across disciplinary boundaries (NDS, 2017b). Building on existing infrastructure for data archiving and sharing within specific communities, NDS aims to provide open, shareable tools that will support cross-disciplinary research and new discoveries to help transform education, society, and economic development. NDS focuses on innovations that bring domain specific data management components into cross-disciplinary use, as well as projects that seamlessly integrate disparate services.

To advance this vision, the NDS Consortium has been established as a coalition of stakeholders, and its inaugural workshop was held in June 2014 in Boulder, CO. The Consortium links together National Science Foundation DataNet projects (e.g., DataONE, SEAD), Data Infrastructure Building Blocks (DIBBs) projects (e.g., NCSA Brown Dog, Whole Tale), the National Science Foundation Big Data Innovation Hubs, and other major community initiatives (e.g., EarthCube, ICPSR, MagIC); Major Research Equipment and Facilities Construction (MREFC) projects; National Institute of Standards and Technology's (NIST) Material Measurement Laboratory; universities, libraries, civic organizations and municipalities (e.g., City of Chicago, ThinkChicago), and national organizations and the services that connect them (XSEDE, Globus, ESIP); publishers (e.g., Elsevier); and international efforts (e.g., RDA, GO FAIR, GODAN) (NDS, 2017b). Towards a world where it is easier to search, publish, link, and reuse data of all disciplines, the NDS Consortium is advancing discovery by enabling open sharing of data, increasing collaboration within/across fields, providing large-scale data service interoperability, and facilitating an incubator of data technologies, projects, and pilots (McHenry, 2017). Additionally, the consortium launched the NDS Labs Workbench (Willis, 2017), a scalable platform for research data access, education, and training to promote data tools (NDS, 2017a). The consortium will drive impact toward an open framework that will revolutionize data sharing through effective partnerships between the U.S. and international research organizations and publishers.

References

McHenry, K. 2017. Enabling Open Science Without Impeding Open Science. Presentation to the National Academies of Sciences, Engineering, and Medicine's Committee on Toward an Open Science Enterprise, Public Symposium. September 18, 2017.

NDS (National Data Service). 2017a. Online. Available at http://www.nationaldataservice.org. Accessed December 13, 2017.

(Continued)

BOX 3-7 Continued

NDS. 2017b. A vision for accelerating discovery through data sharing, Online. Available at http://www.nationaldataservice.org/NDS-Summary.pdf. Accessed December 14, 2017.

Willis, C., M. Lambert, K. McHenry, and C. Kirkpatrick. 2017. Container-Based Analysis Environments for Low-Barrier Access to Research Data. Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact 58. doi:10.1145/3093338.3104164.

Historically, scientists (especially natural scientists) have kept their collections either in museums or in central locations in their university departments, but also as personal collections in their own laboratories for their use and that of their research groups. These samples had collection data with varying levels of specificity associated with them; however, neither these data nor the physical samples were easily accessible by others. The preservation of scientific collections and data acquired with public and/or private funding, and their wide accessibility now and in the future as a public good, is supported and encouraged by professional scientific societies (e.g., AGU, 2016; GSA, 2018). McNutt et al. (2016) stated that "access to data, samples, methods, and reagents used to conduct research and analysis, as well as to the code used to analyze and process data and samples, is a fundamental requirement for transparency and reproducibility" (McNutt et al., 2016, p. 1024).

The Role of the U.S. Government

The U.S. government has supported the creation of scientific collections and their long-term management and use as far back as the early 19th century (Sztein, 2016). Federal spending comprises a high percentage of the total amount of money spent on research.

In the last two decades, there has been a drive to make scientific samples that were obtained or generated with support provided by taxpayer dollars more readily available to different actors in the scientific community. Two important reports on this topic have been published by the National Research Council (2002) and the Interagency Working Group on Scientific Collections (IWGSC) (2009, known as the "Green Report"). Reasons for preserving physical collections include: (1) preserved collections allow the replication of the original experiments; (2) samples are sometimes used as standards; (3) samples may be irreplaceable or too expensive to recollect; (4) samples can be sources of ideas and can be used for education and training; (5) samples can be used for future analysis or experimental use; (6) scientific collections can be used for purposes unforeseen when the collection was created; and (7) reprocessing of old samples with new technology allows for the generation of new knowledge.

The IWGSC was created in 2006 by the White House National Science and Technology Council to focus attention and planning for federal/federally funded collections management (IWGSC, 2016). It is managed by the White House Office of Science and Technology Policy (OSTP) and co-chaired by the U.S. Department of Agriculture and the Smithsonian Institution. Fifteen federal agencies have scientific collections and/or granting programs. The variety of physical collections is considerable. Some collections include rocks, minerals, meteorites, cellular and tissue samples, fossils, soils, and water, rock, soil, and ice cores. Others include type specimens of plants, microbes, and animals. Scientific collections can also include living organisms, such as type culture microorganism collections, seed banks and plant germplasm repositories, and other biological resource centers (IWGSC, 2009). An IWGSC survey to identify the scope and range of federally held scientific collections conducted a decade ago (IWGSC, 2009) revealed that, of the 291 responses received, cellular/tissue scientific collections represented 22 percent (held in 10 of the 14 responding agencies), geological collections comprised 21 percent of the collections (held in eight agencies); paleontological collections represented 14 percent (held in four agencies), and vertebrate and botanical collections each represented 12 percent and 11 percent, respectively (each held by seven agencies).

The Green Report contained several recommendations, including the need for the development of budgeting information for collections and assessing and projecting costs; the identification and dissemination of policies and best practices on organization, management, physical and online access, and long-term preservation; and issues related to data and metadata accessibility, especially the need to document physical objects and make collection information available online, and develop an online clearinghouse for information on contents and access to federal scientific collections.

The OSTP issued a Scientific Collections memo in March 2014 (OSTP, 2014; see Appendix D), where object-based scientific collections are defined as "sets of physical objects, living or inanimate, and their supporting records and documentation, which are used in science and resource management and serve as long-term research assets that are preserved, catalogued, and managed by or supported by Federal agencies for research, resource management, education, and other uses" (OSTP, 2014, pp. 2-3). The memo asks each agency to develop plans to manage their physical scientific collections "to improve management of and access to scientific collections," and to function as "an essential base for developing scientific evidence and … resource for scientific research, education, and resource management."

The end goal of this effort is the "systematic improvement of the development, management, accessibility, and preservation of scientific collections owned and/or funded by Federal agencies." This initiative is only for long-term institutional, archival collections, not for short-term project collections. The agencies were to include, among other requirements, consideration of legislative and regulatory requirements, clarification on who has the responsibility to carry out policies, projection of the costs of developing, preserving, and managing scientific

collections, agency requirements and standards for long-term preservation, maintenance, accessibility for public use, strategies to provide online information about physical collection contents and access to objects and digital files, unless limited by law or to protect national interests, definition of the process to de-access, transfer, dispose of collections, assignation of resources within each agency to implement policy, consistency with the 2013 Open, Machine-Readable Data OSTP memo (White House, 2013), and a request to agencies to work together and coordinate through the IWGSC (GSA, 2018).

The registry of U.S. Federal Scientific Collections is a curated source of information about object-based science collections owned or managed by U.S. federal departments and agencies (USFSC, 2018). The registry is a collaboration among the IWGSC, Scientific Collections International (SciColl), and the Smithsonian Institution, which manages the registry. At the time of this writing, 485 institutions are involved in this initiative, which includes 148 institutional and project collections. The main goals of this registry are to *improve access to information* about U.S. Federal scientific collections and the institutions that maintain them; and to *improve interoperability among databases* by providing an authority file of unique codes and machine-readable identifiers for institutions and their collections (OSTP, 2014).

The IWGSC compiled a list of the status of scientific collection policies by federal agencies (IWGSC, 2018). Of the 15 federal agencies, eight have scientific collections policies: the National Aeronautics and Space Administration, the Smithsonian Institution, the U.S. Department of Agriculture, the U.S. Department of Defense, the U.S. Department of Health and Human Services, the U.S. Food and Drug Administration, the National Institutes of Health, and the U.S. Environmental Protection Agency. The U.S. Department of Interior has Interior-wide Museum collection policies, and agencies within the department, such as the U.S. Geological Survey (USGS), are developing their own scientific collection policies.

For example, USGS is developing its policies based on comprehensive documents such as the USGS Geologic Collections Management System (USGS, 2018), a process to help determine the best fate for a given collection. The management of these collections and data is done through the National Geological and Geophysical Data Preservation Program (USGS, 2018). USGS provides some funds for intramural collection management and grants to State Geological Surveys and other Department of Interior agencies. The National Science Foundation's data sharing policy states “Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing” (NSF, 2018a).

A good physical scientific collection is properly documented, well preserved, and curated. The metadata attached should include field number, geographic location, collector, date collected, sample type, reason for collection, project name, other important data, and include analyses and derivative samples. Research specimens can be added to permanent scientific collections following

different pathways: from intramural federal sources, from one federal agency to another, from non-federal researchers, from private collectors, and from international collaborations and exchange (IWGSC, 2009).

In addition, in order to organize the samples in any given collection, sample identification needs to be standardized. One such approach is to assign a Universally Unique Identifier (UUID) to each sample and its associated metadata. In the geosciences, the System for Earth Sample Registration (SESAR) (SESAR, 2018), hosted at the Lamont-Doherty Earth Observatory of Columbia University, and supported by NSF as part of the Interdisciplinary Earth Data Alliance, operates a registry that distributes the International Geological Sample Number (IGSN). The IGSN consists of an alphanumeric code assigned to specimens and related sampling features to ensure both unique identification and unambiguous referencing of data generated by the study of the samples with UUIDs (USFSC, 2018). SESAR catalogs and preserves sample metadata profiles, and provides access to the sample catalog via the Global Sample Search. Individual researchers can obtain their own accounts, which allows them to register their samples. Using UUIDs such as the IGSN is a concrete step towards making samples FAIR (AGU, 2017b). Multidisciplinary meetings (EOS, 2017) are bringing together researchers from disciplines with different approaches to sampling and informatics specialists to discuss relationships between data and samples, issues of data representation, and the challenges of creating and maintaining links between the physical samples and the data derived from them at different collection scales. The Integrated Digitized Biocollections website is an initiative aimed at making "data and images for millions of biological specimens" available online (iDigBio, 2018). Box 3-8 describes examples of open data in the discipline of the earth sciences.

The Role of Universities

In addition to government and museum repositories, universities have played an important role in the curation and archiving of scientific collections. They maintain scientific collections that are funded from both governmental and nongovernmental sources. While many are members of the Natural Science Collections Alliance (http://nscalliance.org), several large repositories are not. A few examples of such university repositories are the International Ocean Discovery Program, the Oregon State University Marine and Geology repository, the Scripps Institution of Oceanography Collections, and the University of California, Berkeley Museum of Paleontology.

Many university repositories have maintained funding through difficult times, but an alarming number are facing budget cuts that have led to closure and loss of valuable scientific collections. The fate of collections held by individual scientists working in university settings can be particularly complex. As the current generation of senior scientists retires, their collections become the responsibility of institutions that must decide what to keep and what to discard and also to find and manage space for such collections. It is not uncommon that the scientist's university department disposes of the collections once he/she retires, with the loss

of potentially valuable samples and the information associated with them. This can be the case despite the fact these samples may be unique and irreplaceable (AGU, 2017b). Even in the cases where the scientist is proactive and tries to place these collections in museums or other institutions before retirement, success is not guaranteed. One of the main reasons given for the rejection of these collections by the institutions is the high cost associated with their proper curation and storage. Universities or other institutions holding collections sometimes decide, usually because of lack of space, funds, curatorial staff, or because of a change in scientific direction, to divest themselves from those collections. (For a recent case regarding a collection of Antarctic marine sediment cores, see Witze, 2016.) While the Antarctic collection has found a new home (Oregon State University, 2017), many other high-value research collections remain at risk.

BOX 3-8
Earth Sciences

Perhaps the best developed model for open data in the earth sciences is in support of the scientific ocean drilling effort whose current incarnation is the International Ocean Discovery Program (IODP; http://iodp.org/about-iodp/history). Scientific coring of the seafloor started in the 1940s and has evolved into an international collaboration with several platforms that allow drilling and recovery of sub-seafloor materials, enabling scientists to investigate samples of sediment, rock, fluids and biota. IODP is the current implementation of this decades-long endeavor. IODP coordinates the international efforts, maintains core repositories for the physical samples, and supports an open (after an embargo period) database for most of the data generated on the ship (http://web.iodp.tamu.edu/OVERVIEW/) and an open publication portal that archives the initial publications related to the research (http://publications.iodp.org). Data generated after the expedition and in shore-based research, and publication in journals outside of IODP, however, are not part of the IODP structure.

Another example of open data in the earth sciences can be found in the seismological community. The study of earthquakes (seismology) began centuries ago, and now relies on a global network of sensitive instruments that record ground motion, many of which report data in real time. Seismology has many applications, including preparing for and mitigating seismic hazards and distinguishing between explosions and earthquakes, among many others. Interpretation of seismic (or nuclear) events relies on records from around the globe, so seismologists began early on to share data.

The Incorporated Research Institutions for Seismology (IRIS) plays a leading role in archiving and providing access to observed and derived data for the global earth science community, in particular, ground motion, atmospheric, infrasonic, hydrological, and hydroacoustic data (https://www.iris.edu/hq). Earthquake data from around the world are accessible via an “earthquake browser” (http://ds.iris.edu) that displays earthquake locations in near real-time and

(Continued)

BOX 3-8 Continued

allows searching and downloading of data in several formats. There are links to open source code for analyzing the data. IRIS also has an array of materials useful for educators from K-12 through graduate programs with many open access publications and online videos.

The twin fields of paleomagnetism and rock magnetism involve magnetic measurements on geological and archaeological materials. These endeavors contribute key evidence to a number of challenging research problems in Earth sciences, including (1) understanding of past climate changes and their relation to the Earth's magnetic field; (2) the evolution of structure in the Earth's core, its boundary and associated influences on the geomagnetic field; (3) the geodynamics of the Earth's mantle, where magnetic data are crucial in determining the fixity of mantle plumes like Hawaii and the possibility of true polar wander; (4) biogeomagnetism; and (5) magnetism at high pressures and in extraterrestrial bodies including other planets. The Magnetics Information Consortium (MagIC; http://earthref.org/MagIC) provides a data archive that allows the discovery and reuse of such data for the broader earth sciences community.

MagIC began in 2002 as an NSF-funded project to develop a comprehensive database for archiving of paleo- and rock magnetic data, from laboratory measurements to a variety of derived data and metadata such as the positions of the spin axis of the Earth from the point of view of the wander continents and the variations of the strength and direction of the field through time, to changes in environmentally controlled rock magnetic mineralogy. Closely linked to the MagIC project is open source software for the conversion of laboratory data to a common data format that allows interpretation of the data in a consistent and reproducible manner. Once published, the data and interpretations can be uploaded into the MagIC database. All software involved with the MagIC project is freely available on GitHub repositories. MagIC also maintains an open access textbook on rock and paleomagnetism and links the data to the original publications (only a portion of which are currently openly available).

While specimen images and other analytical information can be placed online and used by researchers around the world, this does not mean that the actual specimens can be discarded (Nature, 2017). Technologies not yet developed might yield important discoveries when applied to scientific specimens in the future, and analyses performed with those new techniques can supplement original analyses to test novel questions (McNutt, 2016). One such case is the reconstruction of the 1918 influenza virus through RNA sequencing of highly degraded virus fragments recovered from tissue samples from victims of that pandemic, only possible after the development of PCR techniques in the 1980s. The reconstruction of the 1918 influenza virus allowed the development of novel insights into its biology and pathogenesis, and provided important information about prevention and control of future pandemics (Taubenberger et al., 2012). Box 3-9 describes recent examples of scientific collections in the field of biological sciences.

All researchers in any type of setting need to consider their physical collection and data management plans at the earliest stages of their research. The preservation of physical samples has similar challenges to those presented by digital datasets: accessibility, decisions on what to save and what to discard, how to manage what is being saved, and issues of discoverability and of reuse (Sztein, 2016). Funding considerations frequently determine the preservation of collections, their associated metadata, and the databases that permit the discoverability and reuse of those collections. Funding stability would greatly assist in the preservation of those valuable resources for future generations.

BOX 3-9
Precision Medicine

In 2011, a National Research Council consensus study published a bold new vision for research in health and medicine (NRC, 2011). The significant advances in molecular biology together with the promise afforded by electronic health records made it an opportune time to consider new ways of defining diseases while gaining a deeper understanding of disease mechanisms, pathogenesis, and treatments. In early 2015, as part of his State of the Union address, President Obama announced the Precision Medicine Initiative (PMI) and the funding that would accompany it (The White House, 2015a)

The National Institutes of Health (NIH) announced its plan for implementing the initiative later that year (Collins, 2015; NIH, 2015). The program, named the All of US research program, involves recruiting at least 1 million individuals and collecting biological, health, behavioral, and environmental data about them. Participants in the program must be willing to share their health data, provide a biospecimen, and be recontacted for future research. The PMI data is envisioned as a public resource that will be accessible not only to researchers, but also to interested members of the public, e.g., "citizen scientists." The specifics of data sharing and access are under development and are expected to adhere to a set of privacy and trust principles (The White House, 2015b). These principles include complying with legal and other regulatory requirements, adequately informing participants about how their data will be used, developing multiple tiers of data access based on data type and use, and measures for protecting PMI data from unauthorized use. Notably, and to "enrich the public data resource," the principles require that users of the data publish or publicly post the outcome of their research, including negative outcomes.

The Million Veteran Program, an observational cohort study and "mega-biobank" effort, is a Department of Veterans Affairs (VA) research effort (Gaziano et al., 2016). Veterans are asked to provide a blood sample, respond

(Continued)

BOX 3-9 Continued

to a number of questionnaires, and allow access to their electronic health records housed at the VA. Currently, access to the data is limited to VA-affiliated researchers, but future plans include broadening that access and potential collaboration with the All of US program.

References

Collins, F. S., and H. Varmus. 2015. A new initiative on precision medicine. The New England Journal of Medicine 372(9):793-795.

Gaziano, J. M., J. Concato, M. Brophy, J. Fiore, S. Pyarajan, J. Breeling, S. Whitbourne, J. Deen, C. Shannon, D. Humphries, P. Guarino, M. Aslan, D. Anderson, R. LaFleur, T. Hammond, K. Schaa, J. Moser, G. Huang, S. Muralidhar, R. Przygodzki, and T. J. O'Leary. 2016. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology* 70:214-223.

NIH (National Institutes of Health). 2015. The Precision Medicine Initiative cohort program – Building a research foundation for 21st century medicine. Online. Available at https://acd.od.nih.gov/documents/reports/DRAFT-PMI-WG-Report-9-11-2015-508.pdf. Accessed March 30, 2018.

NRC (National Research Council). 2011. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: National Academies Press.

The White House. 2015a. Remarks by the President on Precision Medicine. Online. Available at https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/remarks-president-precision-medicine. Accessed March 30, 2018.

The White House. 2015b. Precision Medicine Initiative: Privacy and Trust Principles. Online. Available at https://obamawhitehouse.archives.gov/sites/default/files/microsites/finalpmiprivacyandtrustprinciples.pdf. Accessed March 30, 2018.

Open Repositories

A number of organizations provide repositories for archiving datasets. For example, the Registry of Research Data Repositories (Re3Data), formerly DataBib, provides the largest and most comprehensive registry of over 1,500 data repositories, with a wide range of disciplines from around the world. A publication, *Metadata Schema for the Description of Research Data Repositories* (Version 3.0), released in 2015, describes the re3data.org properties (Rücknagel et al., 2015). PLOS has identified a set of trusted repositories that are recognized within their communities (see Table 3-3). For example, the Inter-university Consortium for Political and Social Research (ICPSR) is a large archive of digital social science data (MIT Libraries, 2018). For biomedical and environmental science repositories and field standards, PLOS suggests that researchers utilize FAIRsharing (FAIRsharing, 2017) and Re3Data that provide criteria to identify appropriate

data repositories, including licensing, certificates and standards, policy, and other criteria. Additionally, Scientific Data (http://www.nature.com/sdata/policies/repositories) provides a list of repositories that have been evaluated to ensure that they meet their requirements for data access, preservation, and stability. Box 3-10 illustrates open data practices for economics research.

BOX 3-10
Economics

Unlike other social and behavioral sciences such as sociology and psychology where researchers generate their own data, economics has typically relied upon government-collected data and statistics. As a result, every researcher has had access to these data collections. Economics organizations have worked to make research data more accessible. Since the 1970s the National Bureau of Economic Research (NBER) has maintained a public use data archive (http://www.nber.org/data) that started with lending out 9-track tapes of federal data collections such as the Current Population Survey to NBER researchers. When Internet access became available in the 1990s, NBER added data to its website. Data were shared and made available as a way of treating economics as a science where reproducibility is part of the process. These data are widely used by social science researchers. For example, the NBER working paper associated with the NBER patent database has over 3,000 citations in Google Scholar.

The Federal Reserve Bank of St. Louis started the Federal Reserve Economic Data site (FRED) in the 1990s as a way of compiling economic time series data in one location. FRED started as a dial-in electronic bulletin board that moved onto the web in 1995 (FRED, 2018). FRED currently hosts 504,000 US and international time series data from 87 sources (https://fred.stlouisfed.org/) and features online tools, an API, and tools for smart phones. More recently, the Center for the Advancement of Data and Research in Economics (CADRE) at the Federal Reserve Bank of Kansas City began working to document data inputs and methods for various fields in economics with an emphasis on widely used microeconomic datasets such as the Current Population Survey and the Survey of Income and Program Participation (Federal Reserve Bank of Kansas City, 2018).

As behavioral and experimental economics grew as a field, the American Economic Association developed the Randomized Controlled Trial (RCT) Registry in 2013. By 2017, it had registered over 1,000 RCTs in over 100 countries (AEA, 2017). Investigators can voluntarily register their RCTs and related projects. The economics profession has also responded to issues associated with conflict of interest among researchers. The movie *Inside Job* showed that some economists had been paid to generate research that supported the sponsor's point of view, and that the disclosure of sponsor relationships was sometimes lacking. In 2012, the NBER adopted, and shortly afterward the American Economic Association followed suit, a conflict-of-interest

(Continued)

BOX 3-10 Continued

policy in which researchers are required to disclose financial conflicts of interest associated with sponsored research and publications (NBER, 2012; AEA, 2018). The culture of openness has resulted in the economics profession being a relatively FAIR discipline, which may have extended the intellectual reach of economics research (Angrist et al., 2017).

References

AEA (American Economic Association). 2017. A milestone in research transparency: the AEA's RCT Registry now contains 1,000+ studies from over 100+ countries! Online. Available at https://www.aeaweb.org/news/rct-registry-over-1000. Accessed March 30, 2018.

AEA. 2018. Disclosure Policy. Online. Available at https://www.aeaweb.org/journals/policies/disclosure-policy. Accessed March 30, 2018.

Angrist, J., P. Azoulay, G. Ellison, R. Hill, and S. F. Lu. 2017. Inside Job or Deep Impact? Using Extramural Citations to Assess Economic Scholarship. The National Bureau of Economic Research Working Paper 23698.

FRED (Federal Reserve Bank of St. Louis). 2018. What is FRED? Online. Available at https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred. Accessed March 30, 2018.

Federal Reserve Bank of Kansas City. 2018. Data Services. Online. Available at https://www.kansascityfed.org/research/cadre/dataservices. Accessed March 30, 2018.

NBER (National Bureau of Economic Research). 2012. Research Financial Conflict of Interest Policy. Online. Available at http://admin.nber.org/COI/NBER_ResearchFCOI_Policy.pdf.

A growing number of universities are starting to build research data repositories to help researchers manage data, preserve data for the long term, and allow permanent access to datasets in a reliable environment. MIT offers DSpace, a repository established to capture, distribute, and preserve the digital products of MIT faculty and researchers. The Harvard Dataverse Network (DVN), supported by the Harvard-MIT Data Center and Institute for Quantitative Social Science (IQSS), is a repository infrastructure that includes a large collection of research data in the social sciences (Harvard Dataverse, 2018; MIT Library, 2018). The University of Minnesota Libraries also list popular data repositories categorized by subject, including agricultural sciences; archaeology; astronomy; biological and life sciences; chemistry; computer science and source code; earth, environmental, and geosciences; GIS and geography; health and medical sciences; physics; and social sciences (University of Minnesota Libraries, 2018). Data availability facilitates reproducibility of research; allows validation, replication, reanalysis, new analysis, reinterpretation or inclusion into meta-analyses; and makes citation of data and research articles easier by ensuring recognition for authors (PLOS One, 2018).

TABLE 3-3 Open Data Repositories

Disciplines	Repositories	Links
Cross-disciplinary	Dryad Digital Repository	http://datadryad.org
	Figshare	http://figshare.com
	Harvard Dataverse Network	http://thedata.harvard.edu/dvn
	Open Science Framework	http://osf.io
	Zenodo	http://zenodo.org
Biochemistry	caNanoLab	http://cananolab.nci.nih.gov/caNanoLab
	Kinetic Models of Biological Systems (KiMoSys)	http://www.kimosys.org
	Mass spectrometry Interactive Virtual Environment (MassIVE)	http://massive.ucsd.edu
	PubChem	http://pubchem.ncbi.nlm.nih.gov
	Standards for Reporting Enzymology Data (STRENDA DB)	https://www.beilstein-strenda-db.org/strenda/index.xhtml
Biomedical Sciences	The Cancer Imaging Archive (TCIA)	http://www.cancerimagingarchive.net
	Influenza Research Database	http://www.fludb.org
	National Addiction & HIV Data Archive Program (NAHDAP)	http://www.icpsr.umich.edu/icpsrweb/NAHDAP/index.jsp
	National Database for Autism Research (NDAR)	http://ndar.nih.gov
	PhysioNet	http://physionet.org
	SICAS Medical Image Repository	https://www.smir.ch
Marine Sciences	SEA scieNtific Open data Edition (SEANOE)	http://www.seanoe.org
Model Organisms	The Arabidopsis Information Resource (TAIR)	http://www.arabidopsis.org
	Eukaryotic Pathogen Database Resources (EuPathDB)	http://eupathdb.org/eupathdb
	FlyBase	http://flybase.org
	Mouse Genome Informatics (MGI)	http://www.informatics.jax.org
	Rat Genome Database (RGD)	http://rgd.mcw.edu
	SmedGD	http://smedgd.neuro.utah.edu
	VectorBase	http://www.vectorbase.org/index.php
	WormBase	http://www.wormbase.org/#01-23-6
	Xenbase	http://www.xenbase.org/common
	Zebrafish Model Organism Database (ZFIN)	http://zfin.org

(Continued)

TABLE 3-3 Continued

Disciplines	Repositories	Links
Neuroscience	Functional Connectomes Project International Neuroimaging Data-Sharing Initiative (FCP/INDI)	http://fcon_1000.projects.nitrc.org
	NeuroMorpho.org	http://neuromorpho.org/neuroMorpho/index.jsp
	OpenfMRI	http://neuromorpho.org
		http://openfmri.org
Omics	ArrayExpress	http://www.ebi.ac.uk/arrayexpress
	Biological General Repository for Interaction Datasets (BioGRID)	http://thebiogrid.org
	Database of Interacting Proteins (DIP)	
	dbGAP	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
	The European Genome-phenome Archive (EGA)	http://www.ncbi.nlm.nih.gov/gap
	Gene Expression Omnibus (GEO)	http://www.ebi.ac.uk/ega
	GenomeRNAi	
	GPM DB	http://www.ncbi.nlm.nih.gov/geo
	IntAct Molecular Interaction Database	http://www.genomernai.org
	MetaboLights	http://gpmdb.thegpm.org/index.html
	NURSA	http://www.ebi.ac.uk/intact
	PeptideAtlas	http://www.ebi.ac.uk/metabolights
	ProteomeXchange	https://www.nursa.org/nursa/index.jsf
	Proteomics Identifications (PRIDE)	http://www.peptideatlas.org
		http://www.proteomexchange.org
		http://www.ebi.ac.uk/pride/archive
Physical Sciences	Australian Antarctic Data Centre (AADC)	http://www1.data.antarctica.gov.au
	Cold and Arid Regions Science Data Center (CARD)	http://card.westgis.ac.cn
	Environmental Data Initiative Repository	
	National Climatic Data Center (NCDC)	https://portal.edirepository.org/nis/home.jsp
	National Environmental Research Council Data Centres (NERC)	http://www.ncdc.noaa.gov
	Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)	http://www.nerc.ac.uk/research/sites/data
	PANGAEA	http://daac.ornl.gov
	Reaction Database Standard Search Interface	
	SIMBAD Astronomical Database	http://www.pangaea.de
	UK Solar System Data Centre	http://durpdg.dur.ac.uk/HEPDATA/REAC
	World Data Center for Climate at DKRZ (WDCC)	
		http://simbad.u-strasbg.fr/simbad

		http://www.ukssdc.ac.uk http://www.wdc-climate.de
Sequencing	Database of Genomic Variants Archive (DGVa) dbSNP dbVar DNA DataBank of Japan (DDBJ) EBI Metagenomics EMBL Nucleotide Sequence Database (ENA) European Variation Archive (EVA) GenBank miRBase NCBI Sequence Read Archive (SRA) NCBI Trace Archive Uniprot	http://www.ebi.ac.uk/dgva http://www.ncbi.nlm.nih.gov/snp http://www.ncbi.nlm.nih.gov/dbvar http://www.ddbj.nig.ac.jp http://www.ebi.ac.uk/metagenomics http://www.ebi.ac.uk/ena http://www.ebi.ac.uk/eva/?Home http://www.ncbi.nlm.nih.gov/genbank http://www.mirbase.org http://www.ncbi.nlm.nih.gov/sra http://www.ncbi.nlm.nih.gov/Traces/home http://www.ebi.ac.uk/uniprot
Social Sciences	Data Archiving and Networking Services (DANS) Inter-university Consortium for Political and Social Research (ICPSR) Qualitative Data Repository	https://easy.dans.knaw.nl/ui/home https://www.icpsr.umich.edu/icpsrweb/landing.jsp https://qdr.syr.edu
Structural Databases	Biological Magnetic Resonance Data Bank (BMRB) Cambridge Crystallographic Data Centre (CCDC) Coherent X-ray Imaging Data Bank (CXIDB) Crystallography Open Database (COD) Electron Microscopy Data Bank (EMDB) FlowRepository Protein Circular Dichroism Data Bank (PCDDB) Worldwide Protein Data Bank (wwPDB)	http://www.bmrb.wisc.edu https://www.ccdc.cam.ac.uk http://www.cxidb.org http://www.crystallography.net http://www.emdatabank.org https://flowrepository.org http://pcddb.cryst.bbk.ac.uk http://wwpdb.org

TABLE 3-3 Continued

Disciplines	Repositories	Links
Taxonomic & Species Diversity	Global Biodiversity Information Facility (GBIF)	http://www.gbif.org
	Integrated Taxonomic Information System (ITIS)	http://www.itis.gov
	Knowledge Network for Biocomplexity (KNB)	https://knb.ecoinformatics.org
	NCBI Taxonomy	http://www.ncbi.nlm.nih.gov/taxonomy
Unstructured and/or Large Data	BioStudies	https://www.ebi.ac.uk/biostudies
	CSIRO Data Access Portal	https://data.csiro.au
	GigaDB	http://gigadb.org
	SimTK	https://simtk.org
	Swedish National Data Service	https://snd.gu.se/en

SOURCE: http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories.

Sharing and Preserving Research Software

Sharing and preserving research software code has become an increasingly important issue in recent years Journals have been introducing policies and new capabilities, including new editorial staff and technical tools, to ensure that the analytical code associated with an article meets certain quality standards and is made available (Baker, 2016). Concerns about reproducibility have provided a major impetus for this trend. Reanalyzing or verifying data requires use of the original code.

Regarding long-term preservation of code, the relevant practices, barriers, and considerations are largely the same as those related to data. The challenges of ensuring that data are properly cited are covered in Chapter 4—citation practices for software are even less developed. Some institutional repositories have developed guidelines and best practices in software preservation that they are using in their communities (Rios, 2016).

Reproducing a study or using older data requires the code and/software utilized during the experiment or the research undertaken. Oftentimes, researchers do not believe that they have the right to preserve software due to licensing terms and conditions. Aufderheide et al. (2018) addressed the issue of software preservation and found that "individuals and institutions need clear guidance on the legality of archiving legacy software to ensure continued access to digital files of all kinds and to illuminate the history of technology" (Association of Research Libraries, 2018). Additional information relating to code and reproducibility is further described in Chapter 4.

Considering the importance of code in the vision of open science, there is a need to address non-computational methodologies. These methods include preregistration of studies, most common in clinical research and psychology, which could be expanded in other areas of science. Kimmelman et al. (2014) stressed the importance of separating exploratory from confirmatory research, and in this context, registration of confirmatory experiments in preclinical research has been suggested (Kimmelman et al., 2014; Mogil and Macleod, 2017). Additionally, publication of laboratory protocols via electronic research notebooks or open access repository for science methods such as protocols.io, which allows forking and amendments to existing protocols, is a helpful feature to accelerate methodological development toward an open science enterprise (PLOS, 2017b; Goodman, 2018).

International Approaches

Open science approaches are being broadly assessed and adopted throughout the world. The British Royal Society has prepared an extensive report on issues related to open science (The Royal Society, 2012). In 2015, the European Council and the Group of Seven (G7) adopted open science and the reusability of research data as a priority. The FAIR principles were adopted by Science Europe and endorsed by the G20 in the 2016 Hangzhou summit (Mons et al., 2017). At its September 2017 meeting in Turin, Italy, the G7 committed to giving incentives

for open science activities and to providing global research infrastructures on the basis of FAIR data (G7, 2017). While the FAIR principles are increasingly recognized by governments, the private sector, and the scientific community globally, infrastructure needs have been addressed most intensively in Europe, Australia, and Africa. This section describes key community-driven initiatives toward an open science enterprise at a global level.

Research Data Alliance

The Research Data Alliance (RDA) is a global community-driven organization, launched in 2013 with support from the EC, U.S. National Science Foundation (NSF), U.S. National Institute of Standards and Technology (NIST), and Australia's Department of Industry, Innovation and Science, to accelerate data sharing and data-driven innovation. As of October 2017, the RDA comprises more than 6,000 individuals from over 130 countries, including researchers, policy makers, and open science enablers and promoters (RDA, 2017b). Through its Working and Interest Groups, RDA creates infrastructure (tools, models, preliminary standards, code, curriculum, policy, etc.) that is developed and deployed to support specific challenges in data sharing and data-driven research. For example, the RDA Data Publishing Services Working Group developed a model for "an open, universal literature-data cross-linking service that improves visibility, discoverability, reuse, and reproducibility by bringing existing article/data links together, normalizes them using a common schema, and exposes the full set as an open service" (RDA, 2017b). Other RDA outputs include models for machine readable data type registries, approaches to data citation for data collections that change over time, curriculum for data science instruction, a common metadata vocabulary for agricultural data, and other infrastructure needed to enable data-driven research.

The RDA meets twice a year at Plenaries around the world to accommodate its global community. Its meetings are working meetings where many of its Interest and Working groups get together to advance the conceptualization, development, deployment, and adoption of its infrastructure outputs, and meet with a broad spectrum of stakeholders and communities. Both the U.S. and European regions of the RDA support the engagement of early career professionals with RDA Working and Interest Groups. RDA plenaries, programs, and operations are supported through its regions by funders from around the world including the National Science Foundation, the National Institute of Standards and Technology, the EC, the Australian Government Department of Education and Training, the United Kingdom's nonprofit company JISC (formerly the Joint Information Systems Committee), the Japan Science and Technology Agency, Research Data Canada, University of Montreal, the Alfred P. Sloan Foundation, the John D. and Catherine T. MacArthur Foundation, and others.

International Council for Science

The Committee on Data for Science and Technology (CODATA) was created by the International Council for Science (ICSU) in 1966 with the mission "to improve the quality, reliability, management, accessibility and use of data of importance to all fields of science and technology" (CODATA, 2016). As the ICSU Committee on Data, CODATA promotes international collaboration to improve the availability, usability, and interoperability of research data. CODATA's 2015 Strategic Plan and 2016 Prospectus of Strategy and Achievement identify its three priority areas (CODATA, 2017):

1. Promoting principles, policies and practices for open data and open science;
2. Advancing the frontiers of data science; and
3. Building capacity for open science by improving data skills and the functions of national science systems needed to support open data.

CODATA achieves these objectives through its standing committees, strategic initiatives, and Task Groups and Working Groups. CODATA supports the *Data Science Journal* and collaborates on major data conferences, such as SciDataCon and IDW. A landmark publication of Science International (composed of ICSU, the InterAcademy Partnership, The World Academy of Sciences, and the International Social Science Council) entitled *Open Data in a Big Data World* highlights critical issues related to open data and open science while laying out a framework for how the vision of Open Data in a Big Data World can be achieved (Science International, 2015). Additionally, CODATA supports educational opportunities for early career researchers, including the International Training Workshop in Open Data for Better Science, through a grant from the Chinese Academy of Sciences.

Another ICSU interdisciplinary body, the World Data System (ICSU-WDS), was created in 2008, building on the over 50-year legacy of the World Data Centers and Federation of Astronomical and Geophysical data analysis Services. ICSU-WDS promotes universal and equitable access to scientific data and data services, products, and information across a range of disciplines including the natural and social sciences and humanities (ICSU-WDS, 2017).

European Activities

Open Science is one of three priority areas for research, science, and innovation policy in Europe (EC, 2017d). To support the transition to more effective open science, the EC launched the European Open Science Cloud (EOSC) in 2016, with a vision for "a federated, globally accessible environment where researchers, innovators, companies and citizens can publish, find and reuse each other's data and tools for research, innovation and educational purposes" (EC,

2016). The EOSC High Level Expert Group, including 10 members from European countries, Japan, and Australia, released its first report, *Realising the European Open Science Cloud*, which provides specific recommendations to the Commission regarding actions needed to implement the EOSC (EC, 2016). In June 2017, the EOSC summit was held in Brussels to further discuss how to make EOSC a reality by 2020.

The EC also established the Open Science Policy Platform (OSPP) in 2016, a high-level expert advisory group that will support the development and implementation of open science policy in Europe. The group is tasked with addressing various dimensions of open science, including the establishment of a reward system, the measurement of quality and impact, the change of business models for publishing, FAIR open data, EOSC activities, research integrity, and open education (EC, 2017e). At the request of the Commission, RAND Europe and other entities, such as Deloitte, Digital Science, Altmetric, and Figshare, developed a monitor that tracks open science trends in Europe while identifying the main drivers, incentives, and constraints on its evolution.

Additionally, a group of European Union (EU) member states is preparing the Global Open (GO) FAIR initiative that focuses on involving all networked initiatives, research disciplines, and interested EU member states to make research data FAIR. The Netherlands has initiated and co-leads the early development of the GO FAIR initiative. Three pillars of GO FAIR include: (1) GO CHANGE, which aims to promote cultural change to make the FAIR principles a working standard; (2) GO TRAIN, which deals with locating, creating, maintaining, and sustaining the required data expertise in Europe through training and education; and (3) GO BUILD, regarding the need for interoperable and federated data infrastructures (Dutch Techcentre for Life Sciences, 2016). GO FAIR encourages close cooperation with activities in other regions, such as the NIH Commons (Bonazzi and Bourne, 2017) to build an Internet of FAIR data and services.

The European Southern Observatory (ESO) has recently endorsed the EOSC Declaration and expressed its support for the EOSC initiative on open access to scientific data. The ESO emphasizes that astronomy has been leading well-managed, curated open access to data in scientific research (ESO, 2017).

Other Global Initiatives

Significant efforts are also underway in Australia and Africa to promote the transition towards an open science system. The Australian National Data Service (ANDS), established in 2008, is a joint collaboration between Monash University and the Australian National University and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) that addresses the challenges of managing research data in the country. Another effort is Australia's Academic and Research Network (AARNet), a high speed low latency network infrastructure for research and education across a diverse range of disciplines in the sciences and humanities (AARNet, 2017). In Africa, the East African Community has recently adopted the Dakar Declaration on Open Science in Africa, following the Sci-GaIA

workshops (Barbera et al., 2015) related to the promotion of open science across Africa (CODESRIA, 2017). In South Africa, the African Data Intensive Research Cloud "aims to establish resources to support data intensive radio astronomy research among collaborating partners in South Africa and African Square Kilometer Array telescope partner countries" (Simmonds et al., 2016, p. 1). Additionally, CODATA is supporting the African Open Science Platform to improve the impact of open data across the research community.

4

A Vision for Open Science by Design

SUMMARY POINTS

- In the previous chapters we defined the goals, benefits, and motivations of open science, as well as some barriers and concerns. In this chapter we take a closer look at how open science can be implemented "by design." We define Open Science by Design as a set of principles and practices that fosters openness throughout the entire research lifecycle.
- Scientific progress is largely influenced by factors that promote—or constrain—the dissemination of knowledge. These factors may be social, economic, or both, and they contribute directly not only to the amount of time it takes a community to fully understand and embrace the implications of scientific discoveries, but also, in some cases, to the successful conduct of science itself.
- Over the past two decades, a related number of activities that include open publications, open data, and open source code have gradually been adopted and shared by increasing numbers of researchers in multiple fields. Many researchers are responding to evidence that openly sharing articles, code, and data *in all phases* of the research process is beneficial to the research community, to the broader scientific establishment, to policy makers, and to the public at large.
- The committee's concept of open science by design is by necessity general and idealized. Some discipline-specific nuances cannot be captured in such a broad concept. Also, and importantly, open science by design is intended as a framework to empower the researcher.

PRINCIPLES OF OPEN SCIENCE BY DESIGN

The overarching principle of open science by design is that research conducted openly and transparently leads to *better science*. Claims are more likely to be credible—or found wanting—when they can be reviewed, critiqued, extended, and reproduced by others. All phases of the research process provide opportunities for assessing and improving the reliability and efficacy of scientific research. As an example, even early in the research process, research plans that are made openly available through a preregistration service, such as Registered Reports

(COS, 2018b), allow review and potential revision of the proposed methodology before data are collected and resources are needlessly expended.

A related principle is that integrating open practices at all points in the research process eases the task for the researcher who is committed to open science. Making research results openly available is not an afterthought when the project is over, but, rather, it is an effective way of doing the research itself. That is, in this way of doing science, making research results open is a *by-product* of the research process, and not a task that needs to be done when the researcher has already turned to the next project. Researchers can take advantage of robust infrastructure and tools to conduct their experiments, and they can use open data techniques to analyze, interpret, validate, and disseminate their findings. Indeed, many researchers have come to believe that open science practices help them succeed.

The enormous changes effected by the Internet and the wide range of digital technologies that are now available have had an immense impact in all areas of life, and scientific research is no exception. Digital computing technologies have enhanced scientific knowledge discovery in open networked environments, including information retrieval and extraction, artificial intelligence, data mining, and distributed computing, as a new paradigm in the conduct of research (NRC, 2012a). These technologies have the potential to accelerate the discovery and communication of knowledge, both within the scientific community and in the broader society. A principle that ensues is that open science by design is a *natural consequence* of the fact that "digital science" is rapidly supplanting other ways of doing science.

PRACTICING OPEN SCIENCE BY DESIGN

The researcher is at the center of the practice of open science by design. From the very beginning of the research process, the researcher both contributes to open science and takes advantage of the open science practices of other members of the research community. Practicing open science by design means that researchers develop new habits and customs to promote better science. Figure 4-1 illustrates that open science practices in all phases of the research life cycle are essential to the realization of open science by design.[1]

The research life cycle begins with the *Provocation* phase. During this phase, researchers have immediate access to the most recent publications and have the freedom to search archives of papers, including preprints, research software code, and other open publications, as well as databases of research results, including digital information related to physical specimens, all without charge or other

[1]The committee acknowledges the work groups such as of the Center for Open Science and its Open Science Framework (https://osf.io) and FOSTER (https://www.fosteropenscience.eu/content/what-open-science-introduction) in developing innovative approaches to supporting and conceptualizing openness in the research life cycle. There is a long history of explaining and describing the life cycle of information regarding access to research resources to achieve open science.

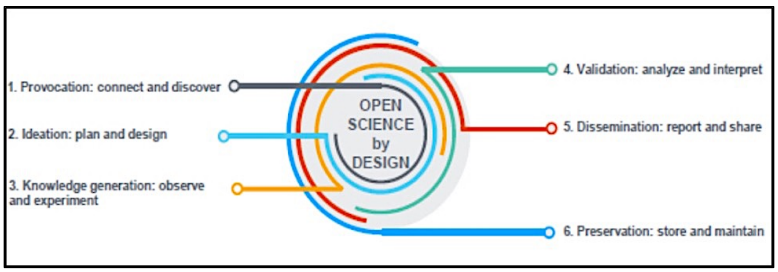


FIGURE 4-1 Phases of Open Science by Design in the research life cycle. SOURCE: Committee generated.

barriers. Researchers use the latest database and text mining tools to explore these resources, to identify new concepts embedded in the research, and to identify where novel contributions can be made. Robust collaborative tools are available to network with colleagues in preparation for the *Ideation* phase of the research.

During the *Ideation* phase, researchers and their collaborators develop and revise their research plans. During this phase they may collect preliminary data from publicly available data repositories and conduct a pilot study to test their new methods on the existing data. When applying for research funding, they develop the required data management plans, stating where data, workflow, and software code will be archived for use by other researchers. In addition, in some cases, they may decide to preregister their research plans and protocols in an open repository, as has, for example, become common practice in clinical research. Publicly preregistering the experimental design and analysis plan in advance of data collection is an effective means of minimizing bias and enhancing credibility in a number of fields. Throughout this phase, they pay close attention to the methods and tools they will use during the *Knowledge Generation* phase, in order to ensure that their final research results will be available in accordance with open principles.

During the *Knowledge Generation* phase, researchers collect data, using tools that guarantee that the dataset will be stored in an interoperable format and includes appropriate documentation and metadata for easy reuse by other interested researchers at some time in the future. Some data are artifacts, physical samples, and specimens, such as rocks, ice core samples, or tissue samples, and researchers develop concrete plans to archive these data according to disciplinary best practices. With the availability of open software, the researcher can document approaches to cleaning and preparing data for analysis in a research notebook. Electronic research notebooks are both human readable and executable documents that can be run to perform data analyses and are useful in the *Validation* phase.

During the *Validation* phase, researchers use open data techniques to analyze, interpret, and validate findings. They present their preliminary findings at conferences and other venues and refine their methods based on relevant comments and critiques. They may deposit their initial working paper in a preprint server of their choice and revise the paper based on the open peer review afforded by the service. They prepare their data in standard formats according to disciplinary norms, and they document and describe both their data and the code that generated their results in optimal ways for reuse and replication. Analysis and interpretation of data are key elements of the scientific process, and the algorithms and workflows for data analysis and interpretation are important research objects in their own right. As they prepare for the *Dissemination* phase, they review and amend, if necessary, their data management plans to ensure that they have met all of the criteria for making their data and code available for broad and open sharing in an appropriate FAIR repository. As discussed in Chapter 3 and Chapter 5, efforts are ongoing to develop new models of scientific communication that rely on open community review, and where the validation stage follows publication. Implementing such models at a large scale will be an important step forward.

During the *Dissemination* phase, researchers select the best venue for open publication of their work, including articles, data, code, and other research products. They revise and, in some cases, substantially improve their work based on the comments of the peer reviewers. Journal articles are currently the primary method for summarizing and sharing scientific results, and the journal's impact factor plays a large role in the assessment of academic achievement. In the digital age, compiling articles in journals for distribution is no longer a requirement for broad distribution. New models are appearing, in which authors publish their work, which then goes through open quality review and certification. The article might then be included in a mega-journal, which are essentially online collections of published articles. Proposals for non-article formats for scholarly communication are also appearing. For example, several neuroscientists have proposed the single figure paper as a form of "nano-publication" that would communicate key findings in a manner optimized for machine-readability (Do and Mobley, 2015). Upon acceptance and before final submission of their work, they select a public copyright license, such as the GNU General Public License for software or a Creative Commons license for other works, including scholarly articles. In preparation for the *Preservation* phase, they make final adjustments to the metadata that describe their research data and code, making sure that these will be reusable by other interested researchers and specific physical samples are preserved and curated for use by other researchers.

During the *Preservation* phase, researchers deposit the final peer-reviewed articles in an openly accessible university archive, or they deposit the articles in another publicly accessible archive as required by their research funders. They deposit their research data and software in one or more FAIR data archives, with clear and persistent links that interlink the article, data, and software. Publicly accessible data may then be used by others in the *Provocation* phase to generate new ideas, marking the beginning of a new research life cycle. Note that data are

often most effectively stewarded and preserved if planned for from the outset, and not at publication time.

Also, and importantly, open science by design is intended as a framework to empower the researcher. As expressed in other NASEM work, the principle for openness of data and other information underlying reported results is that they should ordinarily be available no later than the time of publication, or when the researcher is seeking to gain credit for the work (NAS-NAE-IOM, 2009; NRC, 2003). For journal publication, any sharing prior to the point of final publication is up to the researcher, who is in full control of the decision of when to share.

ENABLING TECHNOLOGIES FOR OPEN SCIENCE BY DESIGN

The practice of open science by design means that the researcher plans for openness right from the start of the research project. Researchers can choose from a growing number of tools, technologies, and platforms as they design and conduct their research. These choices include, for example, the most appropriate data mining algorithms for exploring an unfamiliar dataset, the best workflow tools for capturing and sharing computational workflows, and the established standards for preparing their data for optimal use in FAIR archives. A wide variety of organizations are developing these tools, technologies, and platforms, including community-based nonprofits supported by philanthropy or membership dues, nonprofit coalitions that bring together multiple stakeholders, for-profit startups, and large corporations. A first step might be to register for an author identifier through a service such as ORCID, which provides unique, persistent identifiers for researchers (Meadows, 2016; Wilson and Fenner, 2012). Individuals with an ORCID unique identifier can associate their identifier with their research outputs, whether those outputs are articles, datasets, or other scholarly works. ORCID identifiers can also be used to unambiguously identify researchers in manuscript submission systems, grant application systems, and thesis deposit systems. Because the identifier is unique and persistent, it is not affected by changes in an individual's location, name, or affiliation.

ORCID's User Facilities and Publications Working Group brings together "publishers and facilities to better understand research, publication, and reporting workflows" (ORCID, 2018a). ORCID's Reducing Burden and Improving Transparency (ORBIT) project encourages "funders to use persistent identifiers to automate and streamline the flow of research information between systems" (ORCID, 2018b). Since 2015, Crossref has enabled ORCID records to be automatically updated (Hendricks, 2015). Also, FREYA, a 3-year project funded by the European Commission under the Horizon 2020, builds the infrastructure for persistent identifiers as a core component of open science in the EU and globally.

With eventual publication in mind, researchers can choose an openly available system, such as Docear or Zotero (Beel et al., 2011; Docear, 2018; Vanhecke, 2008; Zotero, 2018) to collect, manage, organize, and format their references. These systems serve as traditional reference managers, but with additional features, including searching and downloading from public databases, tagging and

annotation capabilities, and sharing and exchanging data with collaborators or other software applications. These systems also interoperate with BibTeX, another open source reference manager for those working in the TeX public domain document preparation environment (BibTex, 2018; Hefferon and Barry, 2009; Patashnik, 2003).

Researchers can choose among a variety of open tools for exploring and mining existing datasets. Two popular tools are the R programming language and the WEKA machine learning workbench. The R programming language and software environment is designed for statistical analysis and data mining and integrates with the RStudio user interface (Ihaka, 2010; Tippman, 2015; Verzani, 2011). The WEKA workbench is a toolkit implemented in Java, and like the R software suite, it is also designed to support the entire workflow for experimental data mining, including multiple preprocessing tools, machine learning algorithms, and visualization techniques (Holmes et al., 1994; Eibe et al., 2016).

Automated documentation and sharing of workflows is a key aspect of open science by design. Because many, if not most, areas of science now involve computational analysis of often very large datasets, a variety of tools, both general and domain-specific, have been developed to manage computational data processing and workflows. Importantly, these tools allow researchers to publish their methods and algorithms not only in textual form, but also to publish the code itself, enhancing the reproducibility of the results. In the "Science Code Manifesto," Barnes et al. (2018) emphasized, "the code is the only definitive expression of the data-processing methods used: without the code, readers cannot fully consider, criticize, or improve upon the methods." Stodden et al. stated "Access to the computational steps taken to process data and generate findings is as important as access to the data themselves (Stodden et al., 2016; see Box 4-1)**.** Some recently established journals are dedicated to publishing software, including the Journal of Open Source Software and Journal of Open Research Software, allowing authors to receive credit equivalent to "traditional" journal publication for the code they publish (Shamir et al., 2018).

A growing number of scientific journals, including *Nature*, *PNAS*, and *Science*, require authors to make materials, data, code, and associated protocols available to readers (Nature, 2018; PNAS, 2018; Science, 2018). The *American Economic Review*, the flagship journal of the American Economic Association, has hired a data editor to assist authors with the proper approach to archiving data and code associated with published articles. To incentivize open practices, the Transparency and Openness Promotion (TOP) Guidelines provide a set of recommended standards for scholarly journals to increase reproducibility of research (COS, 2015; Nosek et al., 2015). The TOP Guidelines consist of eight modular standards, with each guideline including three levels of increasing transparency. For example, for the analytic methods (code) transparency standard, a journal that only encourages code sharing or says nothing about it would be at level 0; a journal that requires authors to state whether and where code is available would qualify for level 1; a journal that requires code to be posted on a trusted repository would qualify for level 2; and to qualify for the most transparent level, level 3, the

journal would require that code not only be posted to a trusted repository, but also that reported analyses be reproduced independently before publication.

Mitchum noted that "many are looking to the culture of software programming as a potential model for a more open world of science" (Mitchum, 2015). GitHub, a startup launched in 2008 and originally intended for and used heavily by the open source software development community, is, in fact, increasingly used by researchers as a public platform for sharing their scientific data and code openly (GitHub, 2018; Perkel, 2016). GitHub agreed to be acquired by Microsoft in June 2018 (Ford, 2018). Project Jupyter is an open source framework for scientific software, standards, and services (Project Jupyter, 2018). Jupyter Notebooks, the project's flagship resource, is a domain-independent, web-based platform for supporting reproducible scientific workflows, from "interactive exploration to publishing a detailed record of computation" (Kluyver et al., 2016). Jupyter works with code in several different programming languages and enables notebook sharing when integrated with the recently developed Binder service that provides a computational environment for users to inspect and execute code, and to publish it seamlessly on GitHub (Forde et al., 2018).

BOX 4-1
Reproducibility Enhancement Principles

1. To facilitate reproducibility, share the data, software, workflows, and details of the computational environment in open repositories.
2. To enable discoverability, persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.
3. To enable credit for shared digital scholarly objects, citation should be standard practice.
4. To facilitate reuse, adequately document digital scholarly artifacts.
5. Journals should conduct a Reproducibility Check as part of the publication process and enact the TOP Standards at level 2 or 3.
6. Use Open Licensing when publishing digital scholarly objects e.g. Reproducible Research Standard (Stodden, 2009).
7. To better enable reproducibility across the scientific enterprise, funding agencies should instigate new research programs and pilot studies.

SOURCE: Stodden, V. 2017. Enhancing Reproducibility for Computational Methods. Presentation to the National Academies of Sciences, Engineering, and Medicine Committee on Toward an Open Science Enterprise, First Meeting. July 20, 2017.

Reference

Stodden, V. 2009. Enabling Reproducible Research: Open Licensing for Scientific Innovation. *International Journal of Communications Law and Policy*. Available at SSRN: https://ssrn.com/abstract=1362040.

The Center for Open Science's Open Science Framework (OSF) provides a platform for users to design and create projects, engage with collaborators, manage their research using a suite of tools, prepare their research reports, and preserve their research outcomes. (Foster and Deardorff, 2017; see Box 4-2). The OSF can also be integrated with other open tools, including, notably, support for storage of OSF facilitated research outputs in open repositories. Stodden and colleagues have identified many additional research environments, workflow systems, and dissemination platforms that are now available for researchers' use across a broad spectrum of academic disciplines (Stodden, 2017; Stodden et al., 2014). In addition, other groups, such as the Research Data Alliance, and Earth Science Information Partners, work with communities to create and disseminate open science and open data tools.

BOX 4-2
The Center for Open Science

The Center for Open Science (COS) was created in 2013 as a nonprofit technology and culture change organization with a mission to "increase openness, integrity, and reproducibility of scholarly research" (COS, 2017b, p. 6). The COS has achieved major accomplishments over the past 4 years—in particular on developing and maintaining the Open Science Framework (OSF), a free, open source software tool. The OSF provides cloud-based open project management support for researchers across the entire research lifecycle defined by the framework: (1) search and discover, (2) develop idea, (3) design study, (4) acquire materials, (5) collect data, (6) store data, (7) analyze data, (8) interpret findings, (9) write report, and (10) publish report (COS, 2017a).

As a centralized hub of information, OSF makes research workflow more efficient by keeping all files, data, and protocols in one location; providing controlled access among researchers for effective version control; creating a preprint and meeting abstract automatically; and providing a dependable repository and archive during the research process. The OSF hosts over 86,000 projects and 9,700 registrations by opening their research to the scientific community (COS, 2017a). Every project and file on the OSF has a persistent unique identifier, and every registration (a permanent, time-stamped version of projects and files) can be assigned a digital object identifier (DOI) and archival resource key (ARK) for public sharing or impact measurement (COS, 2017a).

In 2017, COS released a strategic plan for the next 3 years, with a vision for "a future scholarly community in which the process, content, and outcomes of research are openly accessible by default" (COS, 2017b, p.3). The plan proposes the five interdependent activities to accomplish its vision: (1) **metascience** to acquire evidence to encourage culture change; (2) **infrastructure** to build technology to enable change across the entire research lifecycle; (3) **training** to disseminate knowledge to enact change; (4) **incentives** for all

(Continued)

BOX 4-2 Continued

stakeholders to embrace change; and (5) **community** that fosters partnerships among stakeholders to propagate change. The COS is supported by the Laura and John Arnold Foundation, the Defense Advanced Research Projects Agency, the National Institutes of Health, and the National Science Foundation.

References

COS (Center for Open Science). 2017a. Open Science Framework. Online. Available at https://cos.io/our-products/open-science-framework. Accessed December 27, 2017.

COS. 2017b. Strategic Plan. Online. Available at https://osf.io/x2w9h. Accessed December 22, 2017.

Foster, E., and A. Deardorff. Open Science Framework (OSF). *Journal of the Medical Library Association* 105(2):203-206.

Nosek, B. 2017. Achieving Open Science. Presentation to the National Academies of Sciences, Engineering, and Medicine's Committee on Toward an Open Science Enterprise. July 20, 2017.

Curated data are the cornerstone of interoperable systems, allowing others to access, understand, compare, and reuse the data stored within those systems in optimal ways. Curation applies throughout the research lifecycle, from the point of data collection to its eventual deposit in an open repository. Many academic disciplines have established and made available community-driven metadata specifications. The Digital Curation Centre maintains an extensive list of domain-specific metadata standards and, in many cases, includes pointers to tools to help implement those standards (Digital Curation Centre, 2018). For example, the ISA initiative has developed a framework and an open source software suite for creating metadata for -omics-based experiments (Sansone et al., 2012), and the Data Documentation Initiative has developed an international standard and a set of tools for describing data in the social and behavioral sciences (Vardigan, 2013).

There are a variety of efforts underway to ensure that datasets are reliably cited, such that they can be found and appropriately attributed (Altman and Crosas, 2013). Although these practices are not yet as mature as article citation practices, their importance is beginning to be acknowledged. In 2012, the National Academies of Sciences, Engineering, and Medicine (NASEM) hosted a workshop that addressed various dimensions of this topic, including technical requirements, legal and socio-cultural aspects, and disciplinary considerations (NRC, 2012b). Recent community-driven efforts have resulted in a set of principles for data citation (Data Citation Synthesis Group, 2014), the Data Citation Implementation Pilot (DCIP) project (Cousijn et al., 2017), as well as implementation guidelines focused more specifically on scholarly data repositories (Fenner et al., 2016). In

addition, the Center for Expanded Data Annotation and Retrieval (CEDAR), a standards-based metadata authoring system developed under the NIH Big Data to Knowledge Program, is a good example of a general-purpose tool for creating standards-based metadata in a domain-independent manner and that fits into the data submission pipeline for open repositories. There is strong agreement that datasets should have a persistent, globally unique method for identification that is both human understandable and machine-actionable.

The Digital Object Identifier (DOI) is a system for identification of content on digital networks. DOI identifiers are persistent, unique, resolvable, and interoperable for management of content on digital networks (Paskin, 2010). The system is implemented through a federation of registration agencies under agreed upon policies and common infrastructure and is now overseen by the Swiss DONA foundation (DONA). The DataCite organization was founded in 2009 to support data archiving through data citation (Neumann and Brace, 2014; DataCite, 2018). The organization provides persistent DOIs for research data, providing data citation support and services to researchers, data centers, journal publishers, and funding agencies. Many general open data repositories, including Dryad, Figshare, and Zenodo, assign DOIs to the data they store.

As the calls for FAIR research archives have increased, a number of European efforts, including the OpenAIRE project and the European Open Science Cloud, have initiated large-scale infrastructure projects. OpenAIRE's focus is on developing a technical infrastructure for an interoperable network of research repositories from throughout Europe through the establishment of common guidelines and shared metadata standards (Schirrwagen et al., 2013; OpenAIRE, 2018). The project involves the collaboration of researchers from several scientific disciplines as well as librarians, data and information technology experts, and legal specialists. OpenAIRE is collaborating with a number of other groups internationally, including the US-based SHARE (SHared Access Research Ecosystem) project. The SHARE project was launched in 2013 by the Association of Research Libraries, the Association of American Universities, and the Association of Public and Land-grant Universities to strengthen efforts to identify, discover, and track research outputs (COAR, 2015b; Hudson-Vitale et al., 2017; SHARE, 2018).

The European Open Science Cloud (EOSC) project grounds its work in the EOSC Declaration, which includes recommendations and implementation suggestions in the areas of data culture and FAIR data, research data services and architecture, and governance and funding. (EC, 2017a). EOSC has expanded its scope beyond Europe, and they, together with many others, acknowledge that because scientific knowledge is not confined within national boundaries, a globally interoperable open research infrastructure is needed (NASEM, 2018c; Wittenburg and Strawn, 2018).

STRENGTHENING TRAINING FOR OPEN SCIENCE BY DESIGN

As researchers adopt the habits and practices of open science by design, they may need help in identifying and making use of the most effective tools and approaches to use at various stages of their work.[2]

Several recent studies have noted that training of researchers early in their careers is critical, suggesting that open science training can be integrated into existing graduate curricula (OECD, 2015; McKiernan et al., 2016). Others have addressed the practical guidance and training that is needed to help researchers learn how to open up their research processes and results (Carvalho, 2017; EC, 2017f; FOSTER, 2018). The FOSTER (Facilitate Open Science Training for European Research) project, for example, has developed a suite of open science training materials, including courses on open science policies, practices, and resources, research workflow and design, text and data mining methods, data management, legal issues, and responsible conduct of research (See Box 4-3). As described in Chapter 3, the Committee on Data for Science and Technology (CODATA) also supports educational opportunities for early career researchers.

Librarians and other information professionals who have extensive experience in the preservation, publication, and dissemination of digital scientific materials, may, nonetheless, need additional training to address the challenges of curating large-scale open data resources. Several years ago, NASEM undertook a consensus study that examined the career paths for individuals working in the field of digital curation, which they defined broadly as the "active management and enhancement of digital information assets for current and future use" (NRC, 2015). They concluded that although the number of educational opportunities has grown, the available opportunities are well below what is needed to meet the demands of the current data-rich era. Individuals who have discipline-specific knowledge as well as skills in computer and information science are best suited to meet these demands.

Similarly, a recent report from the EC noted that "there is an alarming shortage of data experts both globally and in the European Union" and that there is a "chasm" that needs to be addressed between those who work on digital infrastructure and scientific domain specialists (EC, 2016). An analogous point is made in the current strategic plan for the National Library of Medicine. In discussing the intersection between biomedical data science and open science, the report suggests that more training is needed that involves "enhancing the computational and statistical skills of researchers with biomedical knowledge, and training computer scientists to apply their work to biomedical problems." (NLM, 2018a). The National Science Foundation's Research Traineeship Program trains graduate students in high priority interdisciplinary research areas, including "Harnessing the Data Revolution" (NSF, 2018b).

[2]Indeed, the OSTP 2013 memo explicitly highlights the importance of supporting "training, education, and workforce development related to scientific data management, analysis, storage, preservation, and stewardship" (OSTP, 2013).

BOX 4-3
FOSTER Open Science Training Portal

The Facilitate Open Science Training for European Research (FOSTER) project was initiated in 2014 to support different stakeholders, especially early career researchers, in adopting open access in the context of the European Research Area (ERA) and in complying with the open access policies and rules for the Horizon 2020, the European Union's (EU's) program for research and innovation (Schmidt et al., 2014). The main achievement during the first phase of this project (2014 to 2016) is the creation of a FOSTER portal, an e-learning platform that contains training resources covering a wide range of open science topics for different users, including early career researchers, data managers, librarians, funders, research administrators, and graduate students. The portal includes over 1,800 reusable training materials, 17 self-learning courses, and eight moderated e-learning courses in six languages (EIFI, 2017; FOSTER, 2017). FOSTER's open science taxonomy, which defines and structures the different components of open science (see Chapter 2 and Figure 2-1), has been used to structure the e-learning portal for users to navigate open science topics (Schmidt et al., 2016). Over 6,000 participants, mostly early-career researchers from diverse disciplinary communities, participated in more than 100 training events hosted by FOSTER in 28 European countries (FOSTER, 2017).

As a second phase of the project (2017-2019), the FOSTER Plus (Fostering the Practical Implementation of Open Science in Horizon 2020 and Beyond) currently focuses on creating discipline-specific guidance and resources, in partnership with expert organizations representing the scientific areas of social sciences, humanities, and life sciences. With 11 partners across six countries (Portugal, Germany, United Kingdom, Netherlands, Denmark, and Spain), FOSTER Plus aims to ensure that open science becomes the norm among European researchers. Key objectives of FOSTER Plus include supporting cultural change, consolidating and sustaining a training support network, and strengthening training capacity by addressing the current skills and content gaps at the community, discipline, and institutional levels on the practical implementation of open science (FOSTER, 2017).

References

EIFI (Electronic Information for Libraries). 2017. FOSTER project: Open science training in Europe. Online. http://www.eifl.net/eifl-in-action/foster-project-open-science-training-europe. Accessed December 22, 2017.

FOSTER (Facilitate Open Science Training for European Research). 2017. Online. Available at https://www.fosteropenscience.eu. Accessed December 22, 2017.

Schmidt, B., A. Orth, G. Franck, I. Kuchma, P. Knoth, and J. Carvalho. 2016. Stepping up Open Science Training for European Research. Multidisciplinary Digital Publishing Institute 4(16):1-10.

There are a growing number of collaborative activities among universities, nonprofit organizations, and the philanthropic community related to open science training. The University of California at Riverside and the Center for Open Science (COS) have initiated an NSF-supported randomized trial to evaluate the impact of receiving training on the use of the Open Science Framework for managing, archiving, and sharing lab research materials and data (McKiernan et al., 2016; Nosek, 2017; COS, 2018c). The COS also works with the University of Virginia on an NIH supplemental grant for the Biotechnology Training Program that develops and presents reproducible and open practice curriculum content (COS, 2018d).

Since 2012, the Berkeley Initiative for Transparency in the Social Sciences at the University of California, Berkeley has developed coursework to promote open science in social sciences research (BITSS, 2018). The Data Curation Network of nine major academic institutions, supported by the Alfred P. Sloan Foundation, provides a cross-institutional staffing model for training data curators to build an innovative community to promote data curation practices (Johnston et al., 2017; Data Curation Network, 2018). The Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation, in partnership with several universities, have recently created the Data Science Environments project to "advance data-intensive scientific discovery." Among their efforts is the development of educational materials for researchers at all levels of their academic careers (MSDSE, 2018).

Several related areas of education and training might provide opportunities to incorporate content related to open science practices. First, the field of data science is rapidly growing and evolving. According to NASEM (2018a, p. 1), "Data science is a hybrid of multiple disciplines and skill sets, draws on diverse fields (including computer science, statistics, and mathematics), encompasses topics in ethics and privacy, and depends on specifics of the domains to which it is applied." Those trained in data science at the undergraduate and graduate levels will go on to a wide range of careers, many outside of research. Nevertheless, the core skills and capabilities of data science are clearly relevant to the practice of open science.

In addition, several federal research agencies mandate that some subset of students or trainees that they support receive responsible conduct of research (RCR) training (NASEM, 2017b). Progress toward open science by design will certainly affect the treatment of some traditional RCR topics such as data handling and responsible authorship. RCR training and education programs might benefit from new approaches that incorporate more open science content as a way of ensuring that students and other researchers possess the knowledge and skills needed to practice open science by design.

OTHER CONSIDERATIONS

The committee's concept of open science by design is by necessity general and idealized. It is important to note that research methods and processes vary by

field, and that some discipline-specific nuances cannot be captured in such a broad concept. As discussed in Chapter 3, some fields rely on large data resources that are shared by a defined community. In other fields, experimental data are generated by individuals or research groups, and there may or may not be incentives or rewards for making data openly available. Many disciplines are focused on the study of nondigital data and materials, and face the challenge of ensuring long-term availability. In fields that are focused on the study of one-time phenomena such as earthquakes, specific practices such as preregistration may not make sense or add value.

Likewise, publication cultures vary widely. Some fields such as physics and economics have a long history of utilizing preprints, while this practice is just beginning to gain popularity in the life sciences. In computer science, conferences are a central mechanism for the dissemination of results, and conference proceedings are more important publication venues than are journals (Vardi, 2010). Different fields can and should adapt the open science by design concept to fit their practices and circumstances.

In addition, open science by design raises some questions about roles and responsibilities that have not been fully resolved. For example, should the burden of deposition in publicly accessible archives be shouldered by the organization (publisher or otherwise) that provides services for Dissemination and Preservation, as opposed to the researcher? In order to secure researchers' rights and assuage concerns about reuse and credit, should the choice of license be determined as soon as the research is disseminated, even if it allows for downstream changes (with a separate license for preservation)? Should the researcher be solely responsible for semantically linking research products or should this responsibility be shared?

Given that this is a U.S.-based project undertaken by a U.S. committee, much of the report's discussion and analysis reflects the U.S. experience. However, international perspectives and examples are utilized frequently. At first, wide availability of the tools and infrastructure needed to practice open science by design may be limited to researchers working in well-resourced institutions, primarily in developed countries. Over time, as open science by design demonstrates its value, tools and resources will become more widely available. For example, the African Academy of Sciences recently launched AAS Open Research as "a platform for rapid publication and open peer review for researchers supported by AAS…" (AAS, 2018).

Finally, in order to take hold as a core concept for the future of research, open science by design needs to serve the needs of early career researchers. Chapter 2 discusses how a lack of incentives and supportive culture around open practices is a particular problem facing early career researchers. Open principles and practices must demonstrate their value by enabling early career researchers to become more effective, productive scientists than they would be in a closed environment.

5

Transitioning to Open Science by Design

SUMMARY POINTS

- Despite the significant progress that has been made around the world, the research enterprise remains some distance from completely open science.
- In order to develop effective strategies for achieving open science by design, it is necessary to take stock of what has worked and not worked around the world. Factors such as costs, researcher incentives, policy and legal frameworks, and publishing strategies need to be taken into account.
- The ultimate goals for an ecosystem that supports open science by design are clear: articles immediately available under gold open access, with data available under FAIR (findable, accessible, interoperable, and reusable) principles, and other research products also available. Additional funding, mandates, and community initiatives can be deployed to push towards open science, but careful planning of stakeholder buy-in will be needed to avoid unintended negative consequences and disruptions.
- Still, there are clear short-term actions that can be taken to achieve further progress, and options for longer-term solutions that can be further explored and pursued.

BARRIERS AND LIMITATIONS

In order to realize the vision of open science by design described in Chapter 4, it will be necessary to develop new tools, technologies, and practices. Researchers will need to see the value in adopting them. Training and reward systems will need to be revamped. Discussion in this chapter covers several approaches that have been proposed or are being tried to overcome the most formidable barriers.

Several barriers to open science are discussed in Chapter 2, including those related to the structure of the scholarly communications market as it has evolved over the years, and particularly developments of the past several decades. The vision of open science by design described in Chapter 4 contemplates that all products of the research process will be available immediately at no charge. This vision conflicts with the traditional subscription-based mode of scientific journal

distribution and related aspects of scholarly communications practices. Many traditional publishers are offering open publication options and new open publishers have emerged, with most using a business model based on article processing charges (APCs) that are paid by the author, the author's institution, or the sponsor.

Fully open publications are immediately accessible to all researchers at no cost and are available to all researchers under a copyright license that permits them to perform text and data mining or other productive reuses of the literature without the need for any negotiations or further permissions. While some subscription publishers have begun to offer researchers some forms of access for text and data mining and other productive reuses, their terms of access usually impose some restrictions on reuse.

Another important aspect of the transition to open science relates to the availability of data, code, and other research products under FAIR principles. In contrast to the market for distributing articles, the markets for distributing digital research products such as data are unevenly developed.

This chapter covers possible options and pathways for realizing open science by design, taking into account the legal and policy frameworks that apply, and the landscape of organizations and initiatives that are working in this area.

LEGAL FRAMEWORK

This section focuses on how the law treats "openness" as it relates to access and use of scientific information. While intellectual property law is the most common legal regulation of open science, the relevant law begins with the free speech guarantees of the First Amendment of the U.S. Constitution, along with international agreements. Free speech includes the right to speak, the right not to speak, and the right to listen. These fundamental liberties are the baseline condition governing open access to scientific information. When applied to scientific research, they guarantee the right to share and have access to research results.

The Constitution also grants Congress the power to depart from this baseline when creating intellectual property laws consistent with the First Amendment. Intellectual property law balances the public's right to know against the private interests of researchers to restrict the use of their works for limited times. In the United States, this guarantee provides the legal basis for open science when intellectual property law does not apply. Potentially applicable branches of intellectual property law are: (1) copyright, (2) special database rights in the EU, South Korea, and parts of Eastern Europe, (3) contracts and licenses, (4) patents, and (5) trade secrets. Each of them, except patents, applies automatically and attaches exclusive rights to the protected information.

Although there are cases where private companies have made proprietary data available for research and analysis by the broader community, issues related to proprietary research at companies are not central to open science. Further, although inventions based on university research are often patented, patenting need not interfere with reporting results and making data available. Therefore, issues related to proprietary research that results in patented inventions and trade secrets

lie largely outside the scope of this report. Patents and trade secrets are not covered in what follows. Also, legal issues related to the research use of data generated in other contexts (e.g., social media data) and issues related to the utilization of research results by policy makers are not considered here.

Copyright

Copyright law is the most salient form of intellectual property for this report because it applies automatically to most informational outputs of scientific research, including journal articles, less formal research reports, the organization of datasets, and software. In the United States, federal copyright protection has been granted automatically since 1978, and the requirement that publications carry a copyright notice to maintain protection stopped in 1989. Copyright law is founded on certain science-friendly concepts and imposes no restrictions on sharing the basic building blocks of knowledge—facts and ideas. Researchers rely on this freedom to copy in their daily practice. While, for example, patents apply to specific applications of the CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR associated protein 9) process for gene editing (NLM, 2018b), the ideas and facts underlying the process can be freely built upon. Similarly, raw observational and experimental data are considered "facts" for copyright purposes, free to be shared and reused (NRC, 2009).

Copyright applies to original works of authorship. With respect to journal publications, the "author(s)" who own the copyright sometimes differ from listed authors. Scholarly norms about who receives authorship credit vary by discipline and usually are based on some measure of contribution. The 2016 article that reported the first observation of gravitational waves listed around 1,000 authors, and articles reporting on large clinical trials may have hundreds of authors (Abbott et al., 2016). For copyright purposes, authors are those individuals who wrote the text of an article, created figures or charts, or who otherwise contributed original expression in the work.

Copyright grants the author(s) exclusive right to publicly reproduce the work, distribute copies, display, perform, or otherwise communicate it and make adaptations. When a copyrighted work is created within the scope of employment, the employer is treated as the author and owns the copyright. There is some uncertainty about how this so-called "work for hire" rule applies to outputs of research by full-time faculty, but many institutions have adopted IP policies that address this uncertainty, often recognizing researchers as the authors (and therefore copyright owners) of journal articles they write, datasets they produce or assemble, and software they create.

Most countries also provide authors with some level of "moral right" to their works. These rights are personal to the author and cannot be transferred. Outside the United States, authors have rights of attribution, as well as the right to end attribution if they no longer wish to be associated with the work. A strong version of such rights even gives the author the right to retract a work from publication and to enjoin any further publication or duplication.

Under U.S. law, authors can transfer some or all of their copyright if they sign an agreement to this effect. Subscription-based journals usually require authors to transfer all or part of their copyright(s) to the journal, designating a "corresponding author" who signs on the others' behalf. This allows publishers to restrict access to paying customers and use the threat of a copyright infringement lawsuit to deter republishing or reusing content without a license. Alternatively, the grant of copyright permission (a nonexclusive license) can be executed less formally. In a case where authors never sign a publication agreement, the publisher holds a nonexclusive license and the authors retain copyright.

The rights of copyright holders are constrained by statutory limitations and by exceptions to the owner's exclusive rights to certain reuses. These limitations and exceptions vary by country, so that the right to use the copyrighted layer of a dataset—for example, by copying the whole set—without permission depends on where the copying occurs. All countries have a list of uses permitted by law, but these lists vary widely, and the identified uses are often specific and narrow. Countries also create their own exceptions to determine whether a use is permitted, such as the fair use doctrine in the United States and Israel or fair dealing in many Commonwealth countries.

Sui Generis Database Rights—Europe and South Korea

In the EU, certain candidate countries in Eastern Europe, and South Korea, research data may also be subject to a special database right. As frustrating as this may be to a globalized research community, this right could apply to a substantial amount of computerized data downloaded in Europe or South Korea, but not elsewhere.

When sui generis database rights were introduced in 1996, some experts warned that expanded copyright protections, new technologies restricting access to digital content, and database protections could enable proprietary claims to factual matter that previously entered the public domain as soon as it was disclosed (Reichman and Uhlir, 2003). Others asserted that this legal right would be a significant barrier to sharing research data were it not subject to a limitation for noncommercial research. Since then, courts have interpreted this database right in a manner that limits its potential impact on researchers. However, the European Commission launched a review of its Database Directive in 2017, and a 2018 report supporting this evaluation found that European database rights added complexity to data-intensive research and created barriers to making databases open (EC, 2018d).

Contracts and Licenses

When one or more intellectual property rights apply to research outputs, the owner of such rights can grant permission for reuse through a license. In legal terms, a grant of permission is a nonexclusive license. An exclusive license is one

in which the rights holder agrees to give up any rights to use the intellectual property, usually in return for some form of compensation. From a legal perspective, terms of use or other "licenses" fall into one of two groups. In the first group, there is an underlying intellectual property right associated with data that would be violated by the user in the absence of the permission granted by the terms. That is an intellectual property license. Violation of such a license could lead to a court order requiring the user to cease any further use. Damages and attorneys' fees may also be assessed against the breaching user.

In the second group, there is a collection of data that has no underlying intellectual property right associated with it, such as a large collection of sensor data that is organized in an unoriginal manner—say, chronologically. If one were to download these data from a site with "terms of use" associated, those terms are still enforceable as a contractual agreement, but there would be no intellectual property right to infringe. Enforcing any use restrictions in this second group of agreements is much more difficult because the author of the terms has to prove that the use has caused measureable economic damages.

Although there are policy arguments against enforcing the terms of use in this second group—because they impose use restrictions on data that intellectual property law treats as in the public domain—courts in the United States and elsewhere generally have found these terms of use to be enforceable as long as the basic requirements for voluntary agreements have been met. For example, a Maryland district court upheld a terms-of-use agreement even when a third-party user obtained database access merely by clicking a box to accept, but failed to review, the terms of use. Since the practice is legal and enforceable, it should be a topic for community discussion whether it is ethical or appropriate to condition access to data on agreement to a contract that imposes use restrictions on data that is otherwise free of any intellectual property rights.

Types of Licenses

Rights of use can be shared or granted by several types of licenses. The broadest is the Creative Commons Attribution (CC BY) license, which requires only that the user provide attribution as directed by the licensor. This license is used by open access publishers, including PLOS; creators of open educational resources, such as OpenStax College and Rice Connexions; and a range of other creators.

The owner of IP rights can also grant free permission for use through a nonexclusive license, which applies most directly to data without monetary value. If the data are valuable, the owner may grant an exclusive license, in which the rights holder gives up any rights in return for some form of compensation.

In cases where permission has strings attached, mapping how intellectual property law does—and does not—apply to research data may be of use. For those seeking to understand which reuses of another's data are permitted by law, regrettably, the answers to the above questions are more context dependent than many

would like. This is so for two reasons. First, the source of all intellectual property rights is national law, so that users' rights vary by country. Second, as is discussed above, some countries have an additional right that applies to certain factual databases.

Copyright protection can also be permanently removed in most parts of the world if the owner of the rights publicly states the intention to relinquish the rights. Creative Commons provides a tool called CC0 (CC Zero) to remove the rights, and even in countries that deny owners this right, CC0 functions as a license for the user (Box 5-1).

RESEARCH FUNDER POLICIES

This section covers the context for realizing open science by design shaped by the policies and requirements of research funders. The ability to obtain grants from funders to support scientific studies and publish in credible peer-reviewed scientific journals is important for scientists to advance their research careers and receive recognition nationally and internationally for their work. Funders award research projects that align with their values and mission, providing resources to scientists for collecting the data in order to offer solutions to important topics that the funders want to have an impact on. Therefore, the advances that funding agencies and publishers have made are essential in understanding how a transition to open science can be undertaken.

Traditionally, research sponsors do not publish the articles or host the data generated by the work that they support. They may seek to impose certain conditions on awards, which the grantee can accept, reject, or try to modify. In recent years, research funders have taken a more active role in ensuring that work that they support is publicly available, with some going further in their support for open science. Although much of the discussion that follows focuses on U.S. funder policies, international policies are also included because they help shape the global environment for open science.

U.S. Federal Government Policies

Over the past several decades, the federal government has adopted a number of policy changes relevant to open publication or open data. Many of these have affected access to information resources of the federal government itself, or to data that the government produces or uses. Examples include legislative changes such as the Data Access Act of 1999, the Data Quality Act of 2001, and executive branch policies such as the Obama administration's memoranda on Transparency and Open Government (2009) and Open Data Policy (2013) (NRC, 2009; The White House, 2009; Burwell et al., 2013).

BOX 5-1
Creative Commons

Creative Commons is a global organization that promotes the sharing and reuse of creative, educational, and scientific works by supplying standardized public licenses that anyone can use to permit reuse of works they created or to which they own the rights. The primary tools are six copyright licenses, a copyright waiver, and a label that indicates that a work is free from copyright and in the public domain. The six licenses and the Creative Commons Zero (CC0) waiver are designed to respond to creators who have different appetites for reuse of their works. CC0 is a way to dedicate a work to the public domain by waiving all rights under copyright and any sui generis database rights that may apply. This tool is used by those who create public domain clipart, for example, and in connection with sharing data for which copyright is only an incidental consideration. Unlike CC0, the six licenses, including Creative Commons Attribution (CC BY), Creative Commons Attribution-ShareAlike (CC BY-SA), Creative Commons Attribution-NonCommercial (CC BY-NC), Creative Commons Attribution-NoDerivs (CC BY-ND), Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA), and Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND), impose some conditions on reuse.

As the broadest license, the CC BY license "lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation" (Creative Commons, 2018). As recommended for maximum dissemination and use of licensed materials, the CC BY license is used by open access publishers and creators of open educational resources. The remaining five licenses keep the attribution requirement and add other conditions. For example, the Share Alike requirement provides that anyone who adapts the licensed work must license the adaptation under the same license as the source work. This requirement is often compared to "copyleft" licenses used for software, such as the GNU General Public License (GNU's Not Unix!). Wikipedia uses the CC BY-SA license, and only materials licensed under CC BY or CC BY-SA can be uploaded to Wikimedia Commons.

The CC BY-NC license limits licensed uses to noncommercial uses. One may permit only copy-paste reuse and not license the creation of derivative works by using the CC BY-ND license. The final two licenses, including CC BY-NC-SA and CC BY-NC-ND (most restrictive) combine the noncommercial condition with either the Share Alike or the No Derivatives condition (Creative Commons, 2018). This may seem like more complexity than it is worth; however, the uses of these licenses on Flickr demonstrates that creators appear to want this full choice set to share their works (Flickr, 2018).

References

Carroll, M. W. 2015. Sharing Research Data and Intellectual Property Law: A Primer. *PLOS Biology* 13(8):e1002235.

Creative Commons. 2018. About The Licenses. Online. Available at https://creativecommons.org/licenses. Accessed April 11, 2018.

Flickr. 2018. Explore/Creative Commons. Online. Available at https://www.flickr.com/creativecommons. Accessed April 11, 2018.

The National Institutes of Health (NIH) was a pioneer in supporting openness in relation to outputs from research that it supports. In 2005, NIH adopted a voluntary public access policy for peer-reviewed literature that resulted from its funding. In 2008, under the Consolidated Appropriation Act, NIH began to require all grantees to submit an electronic version of their final peer-reviewed manuscripts upon acceptance for publication to the National Library of Medicine's PubMed Central. Articles were to be publicly available no later than 12 months after the official publication (NIH, 2008).

The America COMPETES Reauthorization Act of 2010 called on the National Science and Technology Council (NSTC) to set up a working group that would coordinate federal science agency research and policies related to the dissemination and long-term stewardship of the results of unclassified research. The material covered includes digital data and peer-reviewed scholarly publications, supported wholly or in part by funding from U.S. federal science agencies.

The next significant step toward openness was the release of the memorandum Increasing Access to the Results of Federally Funded Scientific Research by the Office of Science and Technology Policy (OSTP, 2013). The "Holdren memo" directs federal agencies with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the federal government. This access includes any results published in peer-reviewed scholarly publications that are based on research that directly arises from federal funds, as defined in relevant Office of Management and Budget (OMB) circulars (e.g., A-21 and A-11).

Several months after the Holdren memo was issued, the National Research Council organized two planning meetings for the federal government to receive public comments (NASEM, 2013c). Over the next several years following the release of the memo, the National Institutes of Health, the National Science Foundation, and other relevant agencies developed their own policies to implement the Holdren memo (NIH, 2015; NSF, 2015). The policies set out requirements for data management plans and public access to scholarly publications to be included in grant applications, though data deposit requirements and publication date requirements varied by agencies (CENDI, 2017). In January 2017, the OSTP published a report to Congress on the progress of these agencies on implementation of their public-access polices (Holdren, 2017). A 2017 analysis of how well the agency plans addressed the themes set out in the Holdren memo related to the availability of research data found unevenness among agencies, with some themes such as digitization/legacy data and digital object identifiers (DOIs) not mentioned or addressed in a significant percentage of plans (Kriesberg et al., 2017).

The Fair Access to Science and Technology Research Act of 2017 is the latest version of legislation that would essentially provide a statutory basis for the policies instituted in the Holdren memo. Bipartisan groups of sponsors have introduced versions of this legislation in both houses of the last several Congresses, but it has not yet passed (S.1701). If adopted, it would provide a stable legal basis for federal support of open science.

Other Open Science Mandates of Funders, Publishers, and Universities

A number of public funding entities around the world have instituted open science policies (ROARMAP, 2018). As might be expected, these policies vary in terms of their coverage, whether compliance is encouraged or mandated, whether article processing charges (APCs) or other costs associated with open publications or open data are covered or not, and so forth. For example, the Research Councils UK policy was adopted in 2012, and was designed to be implemented over a period of five years, with regular assessment (RCUK, 2012). The RCUK policy focuses on open publications, allows for compliance through both immediate open publishing and the use of repositories, and provides for coverage of APCs. The European Open Science Cloud, which was announced by the EC in 2017 with a goal of implementation by 2020, focuses on enabling FAIR data and principles that will underlie data accessibility and stewardship on a Europe-wide basis (EC, 2017a).

Several private foundations have implemented even stronger mandates on open access. For example, as of January 1, 2017, the Bill & Melinda Gates Foundation requires that publications supported by its grants be: (1) deposited in a specified repository(s) with proper tagging of metadata, (2) published under the Creative Commons Attribution 4.0 Generic License (CC-BY 4.0) or an equivalent, and (3) available immediately upon their publication with no embargo period (Hansen, 2017). Further, data underlying the published research should be immediately accessible and open with the foundation paying reasonable fees in order to publish on the terms of OA policy (Hansen, 2017). The Wellcome Trust also demands that results of research that it funds be made openly available within 6 months of publication and provides financial support for those researchers to publish under a CC-BY license (Wellcome Trust, 2016). Other international organizations such as CERN and the United Nations Educational, Scientific and Cultural Organization (UNESCO) have also published open science policies and cover expenses for data sharing. Furthermore, funders are moving from resource provider to knowledge institutions, with increased special funding for programs to understand open science practices and tools (Table 5-1). Several websites have been established so that researchers can check the publication and data sharing policies of funders and publishers (SHERPA/Juliet, 2016; FAIRsharing, 2017; ROARMAP, 2018).

While publishers expect authors to make research data available to the journal or readers upon request for validation and as a supplement to publication, many do not mandate that researchers or institutions provide for FAIR access or long-term curation of the data. Some prestigious subscription journals including *Science*, *Nature*, and *PNAS*, have adopted policies allowing preprint sharing from authors. The SHERPA/RoMEO (SHERPA/RoMEO, 2016) website indexed over 2000 publishers, with 46 percent explicitly allowing preprint posting and 72 percent allowing authors to archive postprints. For example, *Science* allows authors to immediately post the accepted version of their manuscript on their website and to post to larger repositories such as PubMed Central 6 months after publication.

And the journal *Nature* allows archiving of accepted articles in open repositories 6 months after publication. For journals that do not formally support self-archiving, authors can submit an author addendum (a template is provided by SPARC, 2016) to allow them to retain rights to post a copy of their article in an open repository.

Universities such as Harvard and MIT have adopted rights-retention open access policies in which faculty members agree to grant their universities nonexclusive reuse rights for future published works. With this policy in place prior to publication, faculty work is archived freely without the need to negotiate with publishers. Many subscription publications offer an open access option that requires authors to pay an APC. There is a significant range in the prices charged for APCs, with many open access (OA) journals charging nothing (Crawford, 2016; West et al., 2014). The majority of OA publishers charging moderate or high APCs offer fee waivers upon request for authors with financial difficulty such as those from low-income or low-middle-income countries. Some publishers (e.g., BioMed Central, F1000, PeerJ) have membership programs through which institutions pay part of all of the APCs for affiliated authors, some institutions provide discretionary funds for author APCs, and some funders also cover fees for publishing in OA journals.

TABLE 5-1 Special Funding Opportunities for Open Research, Training, and Advocacy

Funding	Description	URL
Shuttleworth Foundation Fellowship Program	Funding for researchers working openly on diverse problems	shuttleworthfoundation.org/fellows
Mozilla Fellowship for Science	Funding for researchers interested in open data and open source	www.mozillascience.org/fellows
Leamer-Rosenthal Prizes for Open Science (UC Berkeley and John Templeton Foundation)	Rewards social scientists for open research and education practices	www.bitss.org/prizes/leamer-rosenthal-prizes
OpenCon Travel Scholarship (Right to Research Coalition and SPARC)	Funding for students and early-career researchers to attend OpenCon, and receive training in open practices and advocacy	www.opencon2016.org
Preregistration Challenge (Center for Open Science)	Prizes for researchers who publish the results of a preregistered study	www.cos.io.prereg
Open Science Prize (Wellcome Trust, NIH, and HHMI)	Funding to develop services, tools, and platforms that will increase openness in biomedical research	www.opensciencеprize.org

SOURCE: McKiernan, 2016.

STRATEGIES FOR ACHIEVING OPEN SCIENCE BY DESIGN

Given the current legal and policy context, how should research enterprise stakeholders work together to facilitate and accelerate the transition to open science by design? This section discusses various strategies and options for achieving open science by design, given the motivations and barriers discussed in Chapter 2, the current approaches discussed in Chapter 3, and the vision of open science by design described in Chapter 4. Transitions to open science should enable the research enterprise and those who utilize research results to reap the benefits—increased reliability of knowledge, more rapid advances, broader participation in science—while minimizing any disruptions. The vision or new status quo should be sustainable in the sense that it needs to succeed over time, create value for stakeholders, and be adaptable to changes in the research and scholarly communications environment.

The committee was not tasked with developing a specific, detailed funding plan and timeline for implementing open science. The committee recognizes the significant cost barriers that remain to widespread implementation of open publication, open data, and open code. The discussion below explores the trends that are likely to affect the adoption of open science, and discusses analysis, policy changes, and options that have been proposed by a variety of groups.

In addition to considering which options might best facilitate a transition to open science, it is important to consider which approaches might be less effective or might have undesirable side effects, such as disadvantaging early career researchers or researchers based in developing countries. Avoiding such missteps will likely be just as important as choosing effective actions.

The committee started with the assumption that all the relevant stakeholders will understand and agree that open science by design is the most desirable future state for the global research enterprise. These stakeholders include public and private research sponsors, universities and other research institutions, and scholarly communicators. The committee also believes that a critical mass of stakeholder organizations will be willing to coordinate policies and funding mechanisms to support both open data and open publications. The committee's discussion has focused on steps that U.S.-based stakeholders might take, while keeping in mind that no single institution or national body can singlehandedly change the system. Working and thinking globally will help to smooth and accelerate progress toward open science.

Paying for Open Science

Open publication is a complicated mélange of traditional subscription journals, green and gold open access journals, hybrid models, archiving services, and others. In discussing open publishing, we must consider not only relevant costs, sources of funds, policies, and appropriate business models for open publishing, but also how to transition from the current mixed closed-open environment to a model that fully supports open publications. The committee prefers a system that

supports author choice for where to publish. Policy and incentive should drive the system toward open science. Table 5-2 provides a basic outline of dissemination systems based on subscriptions, green open publications, and gold open publications.

Researcher incentives are important. Researcher incentives are very important to take into account when considering transitioning to open publications supported by APCs or by other mechanisms from current subscription-based publishing. As discussed in Chapter 3, bibliometric indicators, most notably the Journal Impact Factor (JIF), play an important role in current research evaluation practices, which affects research funding decisions and reward systems for researchers. The importance of publishing in highly prestigious journals varies widely by discipline and has grown significantly over the past few decades. For example, in biomedical research, Ginther et al. (2018) showed that the most significant predictor of NIH funding is the weighted sum of impact factors of journals where principal investigators publish.

The adoption of open science principles and practices holds the promise of changing incentive and reward systems so that this focus on journal prestige may be reduced. Still, some disparities in the prestige of various dissemination venues might be expected to continue, at least for the foreseeable future, with implications for researcher incentives. As long as universities and funders rely heavily on the signals provided by journals with the highest JIFs, which overwhelmingly tend to be subscription-based, those journals will continue to dominate high-quality submissions, and their publishers will continue to have considerable leverage in negotiating access agreements with research libraries.

TABLE 5-2 Costs to the Research Community in Subscription-Based and Open Access Scholarly Communication Systems

Basis of the system	Cost Types
Subscriptions-based	• Subscriptions to journals • Subscriptions to regularly published conference proceedings • Library handling costs, e.g., managing subscriptions, negotiating purchasing packages, etc. • Author charges, e.g., page charges, color plate charges, etc.
"Green" Open Access (provided via repositories)	• Dissemination costs: the costs of building and running repositories • Storage and archiving costs: the costs of running repositories, storing content and associated content migration and other technical procedures involved in long-term archiving
"Gold" Open Access (provided via journals)	• Cost of article-processing charges (APCs) where levied by journals • Cost of systems within research institutions for processing and recording APC payments

SOURCE: Swan, 2016.

The financial incentives of researchers also need to be taken into account. Currently the average cost of publishing in subscription journals is negligible for many researchers. Thus, the willingness of researchers to devote additional resources from their grant funding or other sources is limited (Tenopir et al., 2017).

Taking the above into account, it is clear that the transition to open science will require a concerted effort on the part of all stakeholders to change researcher incentive and reward systems in ways that place higher value on open practices. In particular, reducing and eliminating the power of bibliometrics in evaluation practices is an urgent task. Funding agencies and research institutions could work together to develop broader measures of scientific contribution such as credit for peer review, data and code creation, replicability, and open publication. On the financial side, co-pays for APC fees would encourage cost competition amongst journals.

Maintaining and strengthening quality review. Quality review and certification of research will continue to be important, and needs to be maintained and strengthened in the transition to a world characterized by open publications. Traditional prepublication peer review, typically performed confidentially by volunteers and organized by publishers, is an important component of the current research dissemination system, in that it provides an expert judgment on the quality and importance of an article and serves as a mechanism to select the articles to fill what has traditionally been a limited number of slots. While the limitations of prepublication peer review in performing these functions are significant and well known, it will likely continue to play an important role, at least in the near future (NASEM, 2017b). It is unclear how review systems of the future will adjust to an open science world where dissemination opportunities are not artificially limited, and where other forms of review made possible by technological and cultural changes come to be more valued.

Strategies to achieve open science might include initiatives to develop and deploy new mechanisms for quality review. For example, authors might publish a preprint that undergoes open peer review and certification before the article is included in an appropriate open journal or online collection. The current system of prepublication peer review has challenges, such as lack of transparency, bias, and exclusivity, that open peer review actually has the potential to improve. Although the scientific community has a long tradition of peer review of journal articles, there is no culture for peer review of other digital research objects, such as metadata for experimental datasets. The success of open science will require new mechanisms to extend peer review to all products of scientific research. There are working examples of postpublication and open peer review, such as F1000 and PeerJ that can be learned from.

Resources and shifts in the distribution of costs are important. Dissemination of research under open science principles requires additional resources. For example, curating and storing data and samples are not without cost. Dissemination of research results requires substantial effort and resources in addition to

the organization of quality review. There is considerable debate over the magnitude of these costs, and the level of income that will be necessary for disseminators to provide necessary services and maintain quality (Van Noorden, 2013).

In open science, revenue can come in the form of grant funding, service fees, dues, membership fees, and donations. There is considerable funding for dissemination already in the system, but it tends to support subscriptions with paywall barriers to readers. Research institutions, the federal government, and other research funders already provide significant financial and in-kind support to the existing system of disseminating research results through journals and other publications. For example, one analysis estimated that Carnegie Research 1 university libraries paid an average of $6 million annually in subscription fees to journals in 2009 (Bergstrom et al., 2014). Some portion of subscription fees are covered by the indirect assessments charged by institutions to research contracts and grants. In addition, institutions and research funders sometimes provide direct or in-kind support (e.g., office space, IT infrastructure, staffing) to scientific societies that publish journals. Likewise, the researchers who serve as the volunteer editors of journals and peer reviewers of articles are typically employed by universities and other research institutions, so that their service represents a significant in-kind contribution by these institutions as well as by the volunteers themselves. Finally, some institutions and funders are covering costs associated with publishing in journals that do not rely on subscription fees. New services and capabilities enabled by open science will require additional resources. The costs of other services associated with open science dissemination, particularly those related to data and the associated software code, may not be covered under current business models. These services are likely to grow in importance in the future and include data analysis, processing, visualization, and mashups with other data.

It is important that the transition to open publication results in a system that serves individual researchers and the community at large at least as well as the existing system. Some will question whether the current system might not be the best available, given that most researchers in the developed world have access to most of the articles that they need and that the system delivers millions of peer-reviewed articles per year. Open publication will need to continue to demonstrate its value. Although the desire on the part of institutions to restrain increases in subscription costs often arises in considering how to transition to open publication, they are distinct issues. Commercial and nonprofit publishers operating on a subscription model may well continue to exist as a key component of the research enterprise. The overarching goal of open science by design is an effective and efficient publication system that ensures openness.

Managing transition is important. Transitioning to open science will involve some uncomfortable changes on the part of stakeholders. It is important that transitions are planned and managed effectively. For example, asking a scientific society to shift its publishing business model as if this could be done quickly and easily is unrealistic. As discussed in Chapter 2, many scientific societies generate surpluses through their publishing activities that support society activities. There

is considerable variety among societies and disciplines in the size of their publishing operations and the extent to which their professional ecosystems depends upon publishing income. Some societies would be severely challenged by the imposition of some types of open publication mandates, especially if they did not include transition provisions. At the same time, research institutions are currently experiencing difficulty in absorbing the steady increases in subscription rates of recent years. There is a need to ensure that institutions can continue to function and perform their functions in the system during any transition.

Reducing or eliminating embargo periods. Central to the rationale for open science are the principles of accelerating discovery and making dissemination of results as effective as possible. Embargo periods work against these principles. Even short embargo periods mean that results are available only to paying subscribers, not to the public, nor to researchers outside a specialty field, nor to search engines, nor to companies, artificially inhibiting progress in an era when scientific progress is accelerating and can in principle be made available immediately. The committee assumes that the ultimate goal of open science is that published results are available immediately upon publication without any embargo period. As discussed above, transitions should take account of the sustainability of stakeholders and their activities. Some disciplines such as physics and economics, where sharing preprints is a long-established practice, are essentially operating on a green open access model with no embargo period today. Other disciplines would face challenges in making such a transition.

Mandates

One possible approach to transitioning to open science would be for all funders to simply mandate gold open publications, perhaps adopting a policy similar to the current Gates Foundation policy, with APCs to be covered by funders as a fixed amount or percentage of the grant, and/or through institutional funds. Some percentage of institutional funding used to support journal subscriptions could be reallocated to support open science projects and infrastructure.

However, mandates, applied naively, can have potentially damaging side effects. It would be difficult for larger agencies, especially those that cover the entire spectrum of research (e.g., NSF), to adopt such an approach abruptly. Some journals that have the reputational value required for advancement in a given field may not currently have open science options. Meanwhile, a number of high-prestige journals important to a field, especially those published by nonprofit scientific societies that operate subscription journals on the bare edge of making ends meet, could suffer disruptions.

There are also reasons to believe that such an approach would produce "winners" and "losers," and would likely involve several unintended negative consequences. For example, researchers in their role as consumers of the scientific literature would win by having better and cheaper access to publications. Open journals themselves and subscription-based journals with open access options

would benefit from additional revenue. Many authors would gain additional readers. Losers would include the "low-demand" authors who do not have research funding or do not have access to funds from their institutions to pay for APCs (McCabe et al., 2013).

Pushing these results to their logical conclusion, a completely open publication model (without any form of subsidies for APCs) could harm early career, less well-funded scientists and those at less prestigious institutions and institutions in low- and middle-income countries. In addition to early career researchers and those at less research-intensive institutions being "priced out" of publication activity, this might result in less research being undertaken and fewer publications. The Pay It Forward Project (see Chapter 3) estimates that moving to an entirely OA model will be costlier than the current system for research-intensive universities. This contrasts with the current situation where subscription fees paid by less research-intensive universities subsidize the publications of the more research-intensive universities: "Considering both the scholar-as-author and scholar-as-reader roles simultaneously, assessing the net value of OA for scholars appears complicated" (McCabe et al., 2013).

In addition, some are questioning whether a long-term scholarly communications model based on APCs is the best or only answer for science. As discussed above, research funders increasingly recognize that communication of the result is integral to the research process. Without communication, funder investment in research is of no value. As a result, funders are open to covering reasonable communication costs as part of their funding responsibility. This coverage does not have to be in the form of covering APCs. It includes posting papers to preprint servers (some with discipline-organized peer review) and disseminating annotated datasets. Major funders (Welcome, Gates, European Commission) are creating their own platforms, suggesting that the journal's days of exclusive primacy may be numbered.

Stakeholders might examine how current mandates are operating, including the current NIH policy, the 2013 OSTP Memorandum, and the EC Horizon 2020. Future changes might be aimed at producing measurable advances in open science that feature transparent costs and ensures that key infrastructure facilitates openness and community control.

Community-Based Initiatives

There are excellent examples of well-defined communities publishing in a specific set of journals and negotiating a business model tailored to that community to achieve open science. In this model, an agent representing a community negotiates with the major journals that publish community-specific papers to provide a price considered acceptable to cover the costs of publications in return for making them open upon publication.

An example is the Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP3), which, after many years of discussions and negotiations in the high-energy physics (HEP) community, arrived at an arrangement for

open publishing (SCOAP3, 2018). In this case, much of the science community working in HEP is connected to CERN, which represents SCOAP3. Agreements were negotiated over time with numerous publishers in this field; as of January 1, 2018, for example, all HEP articles published by the American Physical Society in their relevant journals have been paid for centrally by SCOAP3. Those articles are now available upon publication. In return, APS reduced subscription rates to libraries. This arrangement will be re-examined 2 years after its initiation.

This arrangement has several notable features: (1) a hybrid model is in effect where certain articles are made available upon publication; (2) a central fund, formed out of a complex arrangement of international agency funds, is used to pay for the cost of publications; and (3) participants used a transition period that moved the system toward an open science environment. Assuming it is successful, it may be renewed and extended to additional parts of the physics community. This arrangement does not include data services, but data are available to the broader community of HEP scientists through data sharing agreements of collaborating scientists that are funded by cooperating science agencies.

It is important to note that in this arrangement, authors are not required by their funding agencies to publish in open access journals, nor are any journals required to switch to open access. Authors can still choose where to publish, and journals can decide if they wish to join the agreement and offer similar services. In principle, a HEP author might still choose to publish in a subscription-based journal that is not open, but given these incentives and community norms it is highly unlikely that one would choose to do so.

Issues Raised by Data and Related Services

Open science means more than open publishing. Researchers need to archive data and code and scientific collections associated with preprints and publications. The committee started with the assumption that the desired end state is for all data underlying reported results to be openly available under FAIR principles, with disciplinary standards in place to determine which data should be preserved over what time period and clarity over how the costs will be covered. The economic issues around data are very different from those around publishing, largely because there is already a mature publishing industry with established funding sources and business models in play that evolved over centuries, whereas data services are not well developed nor funded in many fields.

Most data in repositories today are not available under FAIR principles, and the complexities of realizing this will entail significant costs. Making data FAIR is a difficult task for investigators, and substantial public investment is going to be required to change the current situation. Making data "findable" is going to require better standards for metadata; new ontologies for the vast majority of scientific disciplines, which currently lack standardized, granular terms that can be used by data search engines; and new tools to enable investigators and curators to author scientific metadata that are sufficiently comprehensive and standardized so that search engines can locate appropriate datasets with adequate precision and recall. Making

data "interoperable" and "reusable" can only be achieved if the data are annotated with comprehensive, standardized, high-quality metadata. Again, the absence of necessary metadata standards, appropriate ontologies, and easy-to-use annotation tools is a significant barrier. There is a misconception in the scientific community that simply putting experimental datasets in the cloud will make them FAIR. Scientific publications become findable and useable to others only when they are well indexed; scientific data require nothing less.

The practices of data curation and dissemination lag behind article publishing in most fields. For data curation and services to be provided routinely, significantly more agency funding will be required. This is absolutely necessary to support the vision of open science. With over 2.5 million peer-reviewed journal articles published each year with a projected annual growth rate of +3.5 percent (EC, 2012), completely new funding sources, business models, and even businesses will be needed to support not only storage, discovery, access, and delivery of data, but also new solutions that could entail curation, replication tools and services, and the like.

The premise of open science by design is that scholarly articles and associated datasets should be open and available for others without paywall barriers, and that the agencies and universities that support and perform this research should consider the cost and curation of these results to be part of the cost of performing research. However, the services of analysis, manipulation of data, visualization, and so on do not all have to be borne by the funding sources. Such value-added services might be provided by for-profit businesses or by nonprofit organizations. Service providers might compete for subscribers in an open market.

When creating a change, some steps might be eliminated, new steps might be added, and the existing relationship between value, revenue, and cost will change. For example, in the recorded music industry, disruption was caused by changes in recording technology—phonograph to LP to 8-track to CD to streaming and packaging—and purchasing a curated form (an album) gave way to single-track selection and adoption. The "unbundling" and "cord cutting" phenomena seen in cable television represents a similar transition. In the current publishing model, data might be thought of as part of the new "listening" stream of open science, and sharing of open data is perhaps analogous to the "sampling" of music.

Some of the conditions needed to ensure that the necessary technologies and infrastructure exist to realize open and FAIR data and code on a universal or near-universal basis are described in Chapter 3 and Chapter 4. Important elements include: (1) clarification and standardization of data management plan requirements with enforcement mechanisms, (2) training/assistance in data/code archival best practices, (3) development of disciplinary guidelines for the resources that need to be preserved, and (4) resources to support these activities.

It is important to understand the barriers to realizing open and FAIR data across the sciences. At the same time, it is important to recognize and learn from successful efforts to create community data resources described in Chapter 3, such as the National Center for Biotechnology Information (funded publicly by NIH)

and the Sloan Digital Sky Survey (funded privately by the Alfred P. Sloan Foundation). These efforts have accelerated progress in their respective fields.

Changes in the Business Environment

An additional issue to take into consideration in developing strategies to achieve open science is the shift in the business environment around scholarly communication. For example, commercial publishers have undertaken significant horizontal and vertical integration in recent years. As a result of mergers and acquisitions, just five firms now publish over half of the world's peer-reviewed literature (Lipton, 2006). In addition, commercial publishers are acquiring important pieces of the scholarly communications infrastructure, such as preprint servers and institutional repositories, and expanding data archiving and analytics services associated with their journals (Posada and Chen, 2017; Schonfeld, 2017). For example, Elsevier is working to capitalize on movement toward open access by integrating services and tools for researchers and institutions that address needs from "the research design and grant application stages through laboratory research and to and even beyond publishing" (Schonfeld, 2017).

At the same time, a number of libraries and library consortia around the world have taken a harder line in negotiating with commercial publishers (Schiermeir, 2017). The emergence of national consortia with greater bargaining power than individual universities means that commercial publishers may have less room to raise prices (Normile, 2018). The "big deal" pricing strategies of journal publishers have affected the market for research journals, as described in Chapter 2. The growing trend of "big deal" journal cancellations and the rise of availability of free, legally vetted copies of manuscripts (through use of oaDOI and other tools, such as UnPaywall, the Open Access Button, etc.) signals growing support for open science.

Short-Term Steps

Given these trends, what steps might the research enterprise and its stakeholders take to move closer to open science? Possible steps might be framed in terms of short-term and long-term actions that uphold the principles of open science.

Realize Universal Green Open Publication

While an immediate, universal gold open publication mandate might have negative unintended consequences, a universal green open publication mandate would have a number of beneficial effects. This would not be the end solution, since green open publishers might not have final versions of articles and would include material that ends up not being quality-reviewed at all, but such a mandate would be an important step toward a fully open system. There is some useful experience and precedent. For example, the advent and growth of arXiv and other

preprint servers over the last quarter-century has expanded access without disrupting scholarly communication in physics and astronomy. An appropriate embargo policy would need to be determined, such as the current *Science* policy that allows immediate self-archiving and deposit in an open access repository after 6 months.

In order to support such a mandate, funders will need to articulate a desired rights retention or licensing designation in their funding terms and conditions. Likewise, institutions would need to include rights-retention provisions in their campus policies. Researchers would be encouraged to use available tools, such as author addenda, to retain rights to research outputs. The University of California's University Committee on Library and Scholarly Communication issued a Declaration of Rights and Principles to Transform Scholarly Communication, which is one example of a supportive institutional approach (UCOLASC, 2018).

Green open publication and preprints are not synonymous. Other models for green open publishers that incorporate more elaborate quality review exist and can be expanded. In addition, green open publishers often require a support model that includes institutional membership funding and grants from foundations.

The committee believes that a comprehensive plan to bring the market to a world of fully open science could start with an approach along these lines, but it would need to be accompanied by additional steps as discussed below.

Devote More Resources to Data Management and Other Open Infrastructure

Additional resources that will be required for open science could come from several sources. Although a rapid repurposing of the fees that libraries devote to journal subscriptions might be difficult to manage, the idea of devoting some of these funds to support open infrastructure (the 2.5 percent commitment) has gained some currency within the research library community (Lewis et al., 2018). The Open Research Funders Group and its members are another source of funding for open infrastructure (ORFG, 2018).

Some voices have stressed the importance of community control of open infrastructure: "Everything we have gained by opening content and data will be under threat if we allow the enclosure of scholarly infrastructures" (Bilder et al., 2015).

Adopt Approaches to Evaluation and Reward Systems That Avoid Misuse of Bibliometric Indicators and Value Open Science

As discussed above and in Chapter 2, approaches to evaluating research and researchers that misuse bibliometric indicators constitute a significant barrier to more rapid adoption of open science practices. Quite a few funders and journals have signed the San Francisco Declaration on Research Assessment (DORA, 2013). Some funders are taking specific steps such as limiting the number of citations that can be included in a proposal biosketch, and allowing preprints and

other interim research products to be included in applications and reports (NIH, 2017a; NSF, 2016b).

In addition to addressing deficiencies in evaluation practices, research institutions and sponsors could make efforts to visibly reward open practices as a direct effort to improve progress towards their core missions. The idea is that open science is not an "add-on" for these stakeholders, but an important enabling strategy to become more effective.

Strengthen Consortia of Libraries and Other Research Consumers

An important element of controlling costs in a transition to fully open science will be to ensure that stakeholders are not simply adding additional resources to what they are already paying for access to research results. The NorthEast Research Libraries Consortium coordinates and negotiates terms for 30 core academic research libraries.

Greater transparency in the prices that research libraries are paying for subscriptions would also be of wide value to the community (Ploeger, 2017). Although some commercial publishers routinely use nondisclosure agreements, public institutions are often subject to state freedom of information laws, and disclosure of terms can be pursued through this mechanism. The UCOLASC's Declaration of Rights and Principles to Transform Scholarly Communication, cited above, includes 18 principles regarding rights and principles when negotiating with publishers during journal license renewals (UCOLASC, 2018). If adopted, it will influence future journal license negotiations with publishers.

Possible Long-Term Actions

Depending upon the degree of progress made in taking some of the short-term actions outlined above, a more significant set of steps might be explored by stakeholders. Such an approach might combine more far-reaching mandates and other coordinated policy changes, a defined transition period, and a "burst" of funding to cover costs associated with the transition, including those associated with both transitioning to open publishing (e.g., with a temporary hybrid period) and provisioning additional data services (e.g., minimally, new services for data associated with published articles).

The transition period is likely to require some years for the market to adjust, analogous to the period when the music industry experienced a dramatic realignment of revenue and cost reduction stimulated by new delivery models. It will also require policies coordinated across agencies and countries that will differ for science communities that currently operate in different ways. Likewise, the transition is also likely to need a temporary infusion of funds to initiate and cover the transition. Sources of this "burst" funding could come from philanthropic investment, federal agencies, and universities. Since this funding would be designed to incentivize different behavior, clear benchmarks and requirements would need to

be developed. Government participation is likely to require extensive time due to the processes of approval and budgeting.

Communities that have already begun this process may experience little difficulty, but others may face extreme disruption as the market sorts itself out.

Define and Frame a Commitment to Open Science

A more aggressive pathway toward open science might involve one or more commitments on the part of stakeholders to achieve openness by a given date in the future. Given the time that has passed since the 2013 OSTP memorandum, it might be worthwhile to revisit and update this key federal policy. Also, Congressional passage of the Fair Access to Science and Technology Research Act would constitute a stronger commitment to open science than the current policy based on an executive branch memo. The OA2020 statement, established at the 12th Berlin Open Access conference in 2015, is an international initiative aimed at moving the community toward open science on a global basis (Samberg et al., 2018). Its support is not very broad, however. As of April 2018, 107 scholarly organizations worldwide have officially signed the Expression of Interest, including the University of California Los Angeles and UC Riverside (Open Access 2020, 2018).

Expand Voluntary Market Initiatives

The SCOAP3 initiative described above is an example of a voluntary agreement, negotiated by a central organization, that makes certain articles in numerous journals available without an embargo period. This is a big step toward an open science model during a transition period, but it is limited to a specific sub-field of physics and to specific parts of existing journals; the other parts are still behind subscription paywalls. These ideas illustrate how one might go further and coordinate policy across different countries, different funders, and different fields to incentivize the creation of conditions for an open science environment. The goal is to transition to open science without resorting to simple mandates or complex negotiations around a central organization in every subfield (which may nonetheless still be helpful in transitioning to a fully open science model).

Generalizing from the SCOAP3 example, journals might be willing to switch voluntarily to a business model of gold open access if enough funders that cover a given domain (say, physics, computer science, or civil engineering) all agree to an open science policy and provide funds to support it. However, several features are required for such an approach to work: (1) critical mass—support would likely need to span multiple countries; (2) regional flexibility—different countries have different models for funding research and publications (e.g., direct vs. indirect funding of subscriptions); and (3) discretionary waiving of APC costs—there is a need to account for authors without support whose papers merit publication in important journals with an APC-based open science business models.

It is important to note that the complexities of the SCOAP3 arrangement are not all presented here, and that it benefits from the concentration of the HEP community around one international, centrally administered institution (CERN). Adapting the model to other communities may not be straightforward.

Develop Concepts for Transitional Funding

As discussed above, there is currently significant debate over the desired future state of scholarly communication, with some assuming that the future will be dominated by open access and hybrid journals that charge APCs, and others resistant to a future where APCs are the primary mechanism for financing scholarly communication. Whichever approach is ultimately pursued, a burst of transitional funding might be provided to support researchers and institutions that lack the resources necessary to cover APCs or other costs, to support societies that rely on subscription income from their journals to support general society functions, and to launch a new scholarly communications infrastructure that does not rely on APCs. In addition, availability of funding for publication might be made contingent on publishers meeting certain conditions.

During the transition period, additional funds would likely be needed to initiate the movement of the market toward open science so that costs for new approaches could be paid while subscriptions are still also being paid. One issue encountered in other countries that have mandated open publications is that the average APCs of fully open journals are about half as much as those "hybrid" subscription journals that offer single articles in open form (Swan, 2016). To the extent that APCs are supported, a key to this transition would be to ensure that publishers are only eligible for hybrid APC charges if they lower subscription fees as an increasing fraction of their articles are open. A "hybrid-to-open process" would need to be carefully monitored. There is a need to prevent "double dipping"—accruing APC and subscription revenue from the same articles—on the part of publishers (Swan, 2016).

Such an approach would require negotiation and monitoring, but lessons learned by SCOAP3 may be useful. For example, agencies might agree to pilot a switch to such a voluntary market in some specific disciplines, with the goal of opening the entire market after a period of, for example, 5 years.

Explore New Standards and Governance Approaches

An ambitious transition might require new approaches for managing open science within the community at large. This can be complex or very simple. Federal agencies would need to work with research institutions and other stakeholders, perhaps through a new coordinating mechanism.

The research community might also need to create and accept standards that can lower operating costs, provide the structure and guidance that allow for smoother operations, and speed the process of getting data into an open science

format. Standards developed by the community will speed the growth of markets and numbers of users which will justify and allow for investment by service and support organizations in new systems, services, and products. It is also possible for researchers to support services themselves at a low barrier of access as they do for commercial services. Transition to this model would need to overcome considerable resistance from many in the community. As described in Chapter 3, openness varies significantly by discipline, with the highest levels in biomedical research and mathematics, and lower levels in chemistry and engineering (Piwowar et al., 2018). Common issues in different disciplines are availability of infrastructures, policies and standards, and culture. There is a need for raising awareness within different disciplines of the importance of setting standards to move towards open science.

With the creation of standards, a development road map can be created to allow existing services to be extended and new solutions to be developed. A community-based model could ensure that FAIR processes are followed, requirements from the community are worked on at the appropriate time, and resources are allocated to the needs deemed most urgent. A development road map done clearly, transparently, and with appropriate context will benefit the funders, recipients, and consumers of open science.

Provide for Training

Open science will introduce new steps and new processes. As discussed in Chapter 4, training in open practices has to be an important part of the rollout and of ongoing education of all stakeholders, including researchers, librarians, universities, and funders. Training can also help gain supporters and advocates to improve the system toward an open science enterprise.

Ensure Fair Pricing

The open science workflow will incur costs at various points, some of them at service or vendor prices. Creating an ecosystem in which there is a monopoly or single supplier could cause costs to rise without being checked. Conversely, too much competition will fragment or balkanize the market so no individual or group of stakeholders can achieve a critical mass associated with low-cost behavior. Fair pricing also allows for capital that can be applied to investment in future development (features, enhancements, or even disruptive innovation).

Cover New Service Costs

While traditional funding sources can help support publishing in an open science model, additional costs will be needed for new services, particularly those of data storage and analysis. In the special cases of very large data science projects

(e.g., LSST, SDSS), data management and services are fundamental to the projects themselves, and plans to cover costs are usually in place from the outset. But for the majority of research projects, across virtually all fields, the practice of data stewardship is not supported, and significant funds to cover additional data services, both disciplinary and interdisciplinary, will be needed.

For new areas required to support open science, the three main funding sources—government agencies (e.g., NSF, NIH), private foundations (e.g., Gates), and universities—will need to collaborate on several strategies: covering costs through a mixture of local (university) data stewardship; project-based stewardship, in which the receiver of a grant spends a portion of a grant on the new services; and centrally funded national data services that together support open science.

Strengthen Community Leadership

We can already see that the answer depends on community leadership. It is clear that the technology enablers of open science are causing disruptions in many markets, and that open science is just one of many outcomes. But as science is largely performed and funded by governments, foundations, and universities, this "market" is not able to adjust as quickly and freely as other fields such as the music industry (although its transition might still be traumatic, e.g., publishers may go out of business, early career open science practitioners may have difficulty with promotion and tenure, etc.). Therefore, leadership will be needed to develop, coordinate, and implement policy in order to ensure that the disruptive transition to fully open science is orderly and complete, and the goals of open science can be realized effectively, affordably, and quickly.

In the U.S., a possible example for community leadership might be an organization experienced in setting standards and developing technologies, such as the National Institute of Standards and Technology (NIST). While the committee does not make a formal recommendation with regard to NIST, it suggests that such a combination of skills would be appropriate, as long as the entity is tasked at the level of OSTP with working across agencies and foundations. The task of developing a set of coordinated policies that accelerate and manage the transition to open science publishing is indeed a daunting one, whomever is charged with its execution. In addition, community groups would need to be established to advise and guide the processes, which will vary for communities in very different stages of open science practice. For an undertaking of this scale, international coordination would be required, especially with other OECD nations that have already begun this process.

In addition to these considerations, for a voluntary market to operate effectively there must be mechanisms to keep costs down. If coordinated policies from funders of science result in simply moving funds from covering subscriptions in libraries to APCs—say, from a general fund—competition among journals may not be enough to limit price hikes. Authors would have little incentive to choose

lower cost journals if costs were fully covered by a central fund. One strategy would be to require APCs to be paid partly by a central fund and partly from an author's grant. This would ensure that authors are mindful of costs and incentivized to choose lower-cost journals. Together such policies could enable open publishing to operate voluntarily. After a transition period, during which adjustments are made to the funding mechanisms (e.g., how much is allocated centrally vs. from a grant, what revenues might be generated by value-added data services, etc.), the market should become more efficient and driven to lower cost.

BOX 5-2
The Enabling Environment

In 2011–2012, the National Science Foundation and the Max Planck Society hosted a series of meetings in Washington and Berlin with numerous international research supporting organizations, research performing organizations, publishers (particularly of physics journals), libraries, and universities to examine principles and business models around open science, focusing on models for open access publishing and paths to transition from the current system to one that supports broad open access publishing.

These meetings culminated in the outlines of a voluntary open science model called the enabling environment. The aim was to create a marketplace for publishing articles that are open to the broader community, where incentives rather than mandates motivate research organizations, publishers, universities, and authors to participate voluntarily in a new and flexible funding model that supports costs while removing access barriers to scholarly publications and data. Essentially four key points were established that preserve benefits of the current system of publishing in a move to open science:

- Institutional policies: Funders adopt coordinated policies to provide funds for open access publishing of work supported by them. In return, it is assumed that the payment of these funds eliminates all barriers to electronic access to and reuse of articles in journals that choose to enter the marketplace.
- Publishers' choice: The marketplace has to contain enough authors with access to the guaranteed funds so that publishers can reasonably expect to be able to replace their subscription income with income from APCs. Publishers are not compelled to do this. They make decisions based on competitive advantage. Within the marketplace, publishers are free to set their own article fee levels that allow open access to the content of the articles, and to offer other value-added services, possibly for a fee, to authors and readers.

(Continued)

BOX 5-2 Continued

- Authors' choice: The decision about what journal to publish in is left to the author. In order to encourage competition among journals and provide incentives to authors to keep fees low, funders provide support according to a co-payment scheme: part would come from central funds that authors can access only for the purpose of publishing in open journals, and a smaller part would come from funds that authors can spend on articles or on other research activities.
- Flexibility: The details of how this is implemented will vary by publisher and organization, depending on the specific communities that are involved. For example, a public grant-giving agency such as the NSF would presumably have a different implementation from an institute-based organization, such as the Max Planck Society. At the same time, participating publishers would implement flexible charging schemes that allow high-quality articles to be published even from authors who do not have access to sufficient funds. But all participants' policies need to support the basic principles.

With these assumptions, the market could enable publishers to switch to APC business models, provided the way that the funders support the publication charges of their scholars successfully establishes a functioning market between publishers and authors. Access by scholars to APC support could be regulated by just two requirements:

- Co-payment: The author can claim most but not necessarily all of the cost of the article charge from the funder's central publication fund; the rest—the co-payment—is expected to come from other research monies available to the author that could be spent in other ways.
- Journal eligibility: The fund will support articles only in journals which subject articles to peer-review; which provide, on a website, access to reading and downloading an article without charge; which provide a search interface for machine access to a plain-text version of the article; and which provide a license for unrestricted use of the article, even for commercial purposes, as long as there is proper attribution.

The co-payment rule is designed to encourage authors to publish in open journals by providing enough central funding to authors that paying author fees is not a disincentive to their choosing an open journal. The author's share, however, is critical in establishing competition among publishers in this market. It gives authors an incentive to look for less-expensive journals, and it means that they will pay higher charges only if the journal offers higher quality (such as higher impact). Different funders may adopt different co-payment formulas, and they might change their formulas from time to time as they gain experience with the system.

6

Accelerating Progress to Open Science by Design

The benefits of open science are accruing to researchers themselves, research sponsors, research institutions, disciplines, and scholarly communicators. Yet, despite significant progress toward creating an open science ecosystem, today's science is not completely open. Most scientific articles are only available on a subscription basis. Sharing data, code, and other research products is becoming more common, but is still not routinely done across all disciplines. Barriers to more rapid progress include an academic culture and researcher incentives that can work against open science, insufficient infrastructure and training, issues related to data privacy and national security, and the economic structure of the scholarly communications market.

Open science also needs to overcome less defined sources of skepticism, which it can only do by proving its value to the research enterprise over time. Many important transformations and innovations in the history of science, and in history more broadly, have been opposed at first because of difficulty in quantifying or even imagining the benefits. For example, much of the biomedical research community was strongly opposed to the Human Genome Project when it was first proposed, believing that it diverted resources from more valuable investigator-driven work (Palca, 1992). The project and its impact look much different in hindsight. Today's advances in biomedical research, and many other fields such as archaeology, would not be imaginable without genomic mapping and analysis. Also, researchers who are used to a framework where they are accountable to colleagues, to their disciplines, and to their institutions may be uneasy with open science's implication that they are or should be accountable to the broader public.

The open science movement stands at an important inflection point. A new generation of information technology tools and services holds the potential of further revolutionizing scientific practice. For example, the ability to automate the process of searching and analyzing linked articles and data can reveal patterns that would escape human perception, making the process of generating and testing hypotheses faster and more efficient. These tools and services will have maximum impact when used within an open science ecosystem that spans institutional, national, and disciplinary boundaries. At the same time, a number of organizations around the world are adopting new policies and launching new initiatives aimed at fostering open science.

The vision of open science by design presented in this report seeks to enable the large population of stakeholders to move more rapidly toward open science as the default condition for the research they support. These stakeholders include the researchers themselves, universities, private and nonprofit organizations, publishers and journal editors, scientific societies, the philanthropic community, and federal agencies. Despite the barriers that must still be overcome to implement open science, the momentum of the movement toward open science is generally apparent, and strategies for accelerating access have been outlined by many members of the scientific community. To help accelerate this progress further, the committee has reviewed several recent recommendations, including those of a report by the Association of American Universities (AAU) and Association of Public and Land-grant Universities (APLU) and the European Open Science Cloud (EOSC) Declaration, before developing an action statement for specific stakeholders.

RECENT DEVELOPMENTS

AAU-APLU Public Access Working Group Report

A joint working group on public access convened by the AAU and APLU released a report in November 2017 that provides recommendations and summarizes actions for federal agencies and universities to advance public access to data in a sustainable manner. The report recognizes that a significant culture shift at universities and among their faculty is required, in addition to carefully crafted new federal policies and investment in data infrastructure that support open access (APLU-AAU, 2017). The report also suggests, "by committing to a set of shared principles and minimal levels of standardization across institutions and agencies, we can help minimize costs, enhance interoperability between institutions and disciplines, and maximize the control institutions can exert over how they ensure access to publicly funded scholarship" (AAU-APLU, 2017, p. 1).

EOSC Declaration

Internationally, the European Commission released the EOSC Declaration in October 2017 calling on all scientific stakeholders to endorse and commit to the principles of the declaration by 2020. The declaration, which emerged as a result of the EOSC Summit held in June 2017, recognizes the challenges of data-driven research in pursuing excellent science; grants the vision of European Open Science as widely inclusive of all disciplines and Member States in the long term; and confirms the implementation of the EOSC as a process based on constant learning and mutual alignment (EC, 2017a). Regarding data culture, it notes that "only a considerable cultural change will enable long-term reuse for science and for innovation of data created by research activities: no disciplines, institutions or countries must be left behind" (EC, 2017a, p. 1).

FINDINGS, RECOMMENDATIONS AND IMPLEMENTATION ACTIONS

The Committee on Toward an Open Science Enterprise has developed the following set of findings and recommendations based on its review and synthesis of the information gathered throughout the course of the study. Each recommendation is the focus of a section that includes a discussion of relevant issues drawing on other parts of the report and a set of findings. Each of the five recommendations is followed by implementation actions specifying agencies, universities, or other organizations to guide stakeholder efforts to fostering open science by design.

Building a Supportive Culture

The motivations for and barriers to open science discussed in Chapter 2 present something of a paradox, which is clearly expressed by Nosek et al. (2015):

> Transparency, openness, and reproducibility are readily recognized as vital features of science. When asked, most scientists embrace these features as disciplinary norms and values. Therefore, one might expect that these valued features would be routine in daily practice. Yet, a growing body of evidence suggests that this is not the case.

The actual and anticipated benefits of open science include more reliable knowledge, more rapid and creative generation of results, and broader and more inclusive participation in the research process. Significant barriers to wider and quicker adoption of open practices include the incentives and underlying cultural assumptions that operate in many fields.

The specific ways in which cultural barriers to open science operate vary significantly by field or discipline. Overuse and misuse of bibliographic metrics such as the Journal Impact Factor in the evaluation of research and researchers is one important "bug" in the operation of the research enterprise that has a detrimental effect across disciplines, as explained in Chapter 2. The perception and/or reality that researchers need to publish in certain venues in order to secure funding and career advancement may lock researchers into traditional, closed mechanisms for reporting results and sharing research products. These pressures are particularly strong for early career researchers.

Initiatives such as the San Francisco Declaration on Research Assessment seek to achieve broad buy-in on the part of stakeholders to move toward evaluation systems that use other methodologies. Concrete actions, such as the National Institutes of Health (2017a) decision to encourage investigators to use and cite interim research products such as preprints in seeking funding, can have a beneficial effect.

Continued effort by stakeholders, working internationally and across disciplinary boundaries, is needed to change evaluation practices and introduce other

incentives so that the cultural environment of research better supports and rewards open practices.

Findings

- The culture of academia does not adequately reward and support researchers engaged in open science practices.
- University tenure and promotion committees give credit for journal publications, but rarely give explicit credit to investigators who make their publications and data openly available for use by the broader community and thus do not incentivize such practices.
- There are increasing opportunities for authors to make their research products openly available. Many high-quality open access journals exist. An increasing number of high-quality open access publishers are supported by philanthropy and host institutions and offer fee waivers to authors in case of economic hardship (Shieber, 2009; Lawson, 2015). There are even peer-reviewed open access publishers that charge a nominal article processing charge or none at all. The Directory of Open Access Journals can be searched to find appropriate journals (DOAJ, 2018). Many journal publishers do not prohibit prospective authors from depositing their initial manuscripts in preprint servers. Most journal publishers do not prohibit authors from posting their accepted articles on their personal websites or depositing them in their university's open access repository. Most federal agencies require deposit of federally funded research results in public repositories.
- Journal articles are currently the primary method for summarizing and sharing scientific results, and the journal's impact factor plays a large role in the assessment of academic achievement. In the digital age, while the journal framework may well continue for branding and content integration purposes, compiling articles in journals for distribution is no longer a requirement for broad distribution.

Recommendation One

Research institutions should work to create a culture that actively supports Open Science by Design by better rewarding and supporting researchers engaged in open science practices. Research funders should provide explicit and consistent support for practices and approaches that facilitate this shift in culture and incentives.

Implementation Actions

- Universities and other research institutions should explicitly reward the effort needed to make science open by design.

- Universities and other research institutions should partner with federal agencies in developing innovative approaches to assessing the impact of research in ways that include the impact of open science outputs. This should include, but is not limited to, the development of metrics for assessing the impact of interim research products such as preprints, with a view toward comparing those with existing methods for measuring impact.
- Universities and other research institutions should move toward evaluating published data and other research products in addition to published articles as part of the promotion and tenure process. Archived data should be valued, just as the publications that result from them are valued.
- Researchers should make full use of the many opportunities that are available for making their research products openly available, and they should include that information in their curriculum vitae so that they can be appropriately credited and rewarded.
- In fields where this is not already common practice, research funders should encourage and reward the use of data and other research products that are available in publicly accessible databases.
- Universities and other research institutions should encourage and reward studies that focus on the replication and reproducibility of published research. Such studies should be published and made openly available.

Training for Open Science by Design

The importance of training for open science by design is discussed in several places in the report, particularly Chapter 4. Initiatives such as the European Union's FOSTER project and the Berkeley Initiative for Transparency in the Social Sciences (BITSS) have emphasized training in open science and reproducibility. The emergence of data science as a recognized interdisciplinary field has highlighted the need for new educational content and approaches related to data (NASEM, 2018a).

Several federal agencies require that students or trainees supported by grants receive training in the responsible conduct of research, or RCR (NASEM, 2017b). Training and education that covers issues such as open science and reproducibility would complement the existing focus of RCR education and orient these programs toward supporting both research integrity and quality.

Findings

- Few academic institutions provide formal training and education in the principles and practices of open science.
- The university library community has an important role to play in the promulgation and support of open science principles and practices.

- Federal training programs, while requiring training in the responsible conduct of research, do not explicitly require training in the many aspects of open science principles and practices.

Recommendation Two

Research institutions and professional societies should train students and other researchers to implement open science practices effectively and should support the development of educational programs that foster Open Science by Design.

Implementation Actions

- Universities should provide training in best practices for open science and data stewardship as part of the regular curriculum in graduate and postgraduate education and should expect these practices in all onboarding/orientation processes of universities, including new student orientation, new faculty orientation, library orientations, and lab training as a default. Course curricula should be developed and implemented to complement domain-specific courses that support open science by design.
- Research funders should support the development of training programs in the principles and practices of open science by design. Federal agencies should require this training as part of all federally funded graduate training grants (e.g., NSF research traineeships and NIH training grants) to foster open science by design.
- Library and information science schools, professional societies, and other interested organizations should develop course curricula and offer courses in the principles and practices of open science.
- Research funders and professional societies should create programs or contests that seek the creative and innovative integration and (re)use of open data for new and impactful research.
- The private sector and other interested parties should create innovative educational tools for open science principles and practices.

Ensuring Long-Term Preservation and Stewardship

The issues and challenges related to preservation and stewardship of research products, particularly data, code, and other nonarticle products, are considered in several places in the report. On the one hand, some of the technical and cost barriers to long-term data stewardship are falling, as tools for automated metadata tagging and classification become more widely used and cloud storage becomes cheaper over time. At the same time, the outputs of research continue to grow in volume and complexity, meaning that significant additional resources will still be required. In addition, ensuring preservation and long-term stewardship—

particularly beyond the time period specified by the grant—requires standards and institutional capabilities that need to be developed by stakeholders and updated over time.

Findings

- Ensuring long-term preservation and stewardship of data and other research products requires a commensurate long-term commitment of resources.
- Public access to data and scientific collections created with federal support is required by federal agencies but the infrastructure and funding to store; curate; and preserve data, code, samples, and other research products are not necessarily available.
- Although some of the technical and cost barriers to large-scale data storage are falling, the outputs of research continue to grow in volume and complexity, meaning that significant additional resources will still be required. Significant cultural and institutional barriers also remain.
- The library community, including archivists, curators, and other information scientists, play an important role in effecting long-term preservation and stewardship.
- Scientific disciplines vary to the extent that data and other research products are shared and archived.
- Not all data and other research products should be preserved for the long term, and most research communities do not have well-defined criteria for determining what data and physical collections should be preserved and for what length of time. The rise of interdisciplinary research implies that data preservation criteria should consider possible use outside of the discipline in which the research was originally conducted.
- Most federal agencies require a data management plan as part of grant applications, although there is insufficient guidance for compliance expectations and institutional responsibilities.
- Developing and sustaining the infrastructure required for long-term stewardship of research products will present a continuing challenge. The work of developing necessary standards and policies on the part of stakeholders will enable effective planning of new infrastructure and associated financing.
- Approaches should be flexible enough to adapt and change over time. The size and complexity of data in many fields are changing rapidly, so that the solutions that are effective today might not be effective in a few years. At the same time, we have seen new tools and platforms continue to emerge that allow researchers to address challenges that were previously intractable.

Recommendation Three

Research funders and research institutions should develop the policies and procedures to identify the data, code, specimens, and other research products that should be preserved for long-term public availability, and they should provide the resources necessary for the long-term preservation and stewardship of those research products.

Implementation Actions

- Research institutions, professional societies, and research funders should work together to develop selection guidelines and long-term stewardship best practices for the most valuable community datasets and other research products.
- Federal agencies should, consistent with the 2013 and 2014 Office of Science and Technology Policy (OSTP, 2013, 2014) memoranda for expanding public access to the results of federally funded research, continue to develop and standardize requirements for research products planning, management, reporting, and stewardship.
- Private research funders who have not already done so should adopt approaches compatible with those developed for publicly funded research products planning, management, reporting, and stewardship.
- Researchers should describe the plan for dissemination and stewardship of their research products with some specificity, consistent with the standardized sponsor requirements described above, including where their research products will be made publicly available and for what period of time.
- Research funders and research institutions should work together to resource and provide the infrastructure needed for long-term preservation, stewardship, and community control of research products. This infrastructure could be supported through direct costs or through an ear-marked percentage of each funded grant.

Facilitating Data Discovery, Reuse, and Reproducibility

As progress toward open science by design continues, it is important that the community adhere to the ultimate goal of achieving the availability of research products under FAIR (findable, accessible, interoperable, reusable) principles. Open science under FAIR principles has the potential to deliver benefits to those researchers and disciplines that are participating, which will help make the case for supporting openness. Utilizing advanced machine learning tools in analyzing datasets or literature, for example, will facilitate new insights and discoveries. Ensuring FAIR access should be a key consideration in deciding how to build repositories and other new resources.

As is the case with ensuring long-term stewardship, new standards should be developed by funders in collaboration with research institutions and researchers. Fields and disciplines that do not already have well-developed standards and practices for making research products available under FAIR principles will need time and help to create these. Where meeting new standards imposes costs, funders should make the necessary resources available. Open science will be realized more quickly and effectively by avoiding the imposition of unfunded mandates. Specific actions enabling a transition need to be developed in a transparent manner, and avoid disrupting researchers and their work to the extent possible.

Findings

- It is difficult to determine how much data (open or otherwise) are generated through federally sponsored research projects and where they can be found. It is difficult to plan agency or budgetary data strategies based on this missing information.
- For certain types of data in several disciplines (e.g., computational biology, genomics, proteomics), papers cannot be submitted to major journals unless the relevant data have already been deposited in an open domain repository. This has facilitated the discovery and reuse of data as well as the reproducibility of research. At the same time this has only happened in a small number of fields.
- It is difficult to discover datasets and code through search, making the "findable" part of the FAIR principles challenging.
- There is considerable variation among different disciplines for what constitutes ethical practices in the publication and usage of open data.
- Public access to research data is not sufficient to ensure usability and enable reuse. Uncurated data are often difficult to use. Data curation, management, and stewardship allow for optimal discovery, reuse, and validation of the results of scientific research.
- The value of open data depends heavily on the proper usage of such data, which in turn relies on a proper understanding of how the data were generated and organized. Disciplinary differences are considerable, and some very large and complex datasets require considerable knowledge and expertise to use effectively.
- For most researchers, the amount of the relevant published literature is beyond the human capacity to gather, read, and analyze without the assistance of automated discovery and analytical tools. Such tools are in development, but that development is impeded by the lack of ready access to the entire corpus of published scientific research by tool developers.
- Open access publications are legally available for all, although not all open access publishers make their content readily available for bulk transfer to tool developers or users of text and data mining tools.

- Subscription publishers have varying policies concerning the availability and use of their publications for text and data mining, with the largest publishers making this content available only under the terms of a negotiated license agreement.
- Open access to the data and metadata, along with the code used to generate and/or interpret those data, supports reproducibility, replicability, and the reliability of reported results.

Recommendation Four

Funders that support the development of research archives should work to ensure that these are designed and implemented according to the FAIR data principles. Researchers should seek to ensure that their research products are made available according to the FAIR principles and state with specificity any exceptions based on legal and ethical considerations.

Implementation Actions

- Researchers should preferentially use open repositories that have been designed for interoperability and ease of discovery.
- Research funders should work to ensure that research products are available in repositories that allow for bulk transfer of digital objects to developers or users of automated discovery and analysis tools.
- Researchers and research funders should require that research products designated for long-term preservation and stewardship are assigned persistent unique digital identifiers.
- Professional societies and research funders should support efforts to network and federate existing repositories for improved discoverability.
- Research funders should continue to support the development of methods and tools that improve the interoperability of heterogeneous data. Metadata schemes, commonly accepted workflows for the processing and analysis of data, and other standards should be developed and used for improved data discovery.
- Research funders should commission an independent assessment of the state of university and federal data archives. The assessment should address how the FAIR principles have or have not been adhered to and make recommendations for improving accessibility to distributed or federated archives.

Developing New Approaches to Fostering Open Science by Design

As the report discusses in Chapters 3 and 5, there is a great deal of activity on the part of public and private research funders, research institutions, commercial and nonprofit publishers, community-organized groups, and others aimed at

preparing for and shaping a future research enterprise characterized by open science. Significant progress has been made, but a great deal of work needs to be done before open science by design is a reality. The committee focused on the choices facing U.S. organizations and institutions, realizing that the transition to open science by design is inherently a global process.

Chapter 5 describes a number of issues, a few possible scenarios, and options for action. The recent AAU-APLU report emphasizes the need for federal and other research sponsors to clarify requirements. In addition, revisiting federal policies supporting open science would allow for approaches to be modified and updated. Specific actions enabling a transition need to be developed in a transparent manner, and avoid disrupting researchers and their work to the extent possible.

The research enterprise is at an important point in the transition to open science, where research sponsors, both public and private, have an opportunity to shape the future through their investments.

Findings

- Significant progress in open science practices has been made in recent years, but the majority of research products are not open, and very little research output meets the FAIR guidelines.
- Many, though not all, research funder policies are moving toward open science principles and practices.
- Infrastructure for open science is being designed and deployed, although with variation across fields of study.
- Disciplinary preprint servers, such as arXiv, RePEc and BioRxiv, have successfully provided an open platform to post prepublication versions of manuscripts at no charge. These platforms have had an important positive effect on these disciplines.
- Open publications and open data provide an opportunity for the private sector and others to develop useful products for researchers and other communities.
- The current subscription-based business model for many publishers conflicts with the goal of immediate open access to publications and data.
- Article processing charges are a possible replacement for subscription fees as a business model, but they also have limitations, since the payment of the charges will still be a burden on some part of the ecosystem and will fall unevenly on different stakeholders.
- Certain approaches to implementing open publication have the potential to affect the research ecosystem in significant ways, with differential impacts on different stakeholders. In planning new policies and transitions, it will be necessary to anticipate differential impacts to the extent possible, consider ways of avoiding these, and build in evaluative and corrective mechanisms to address unanticipated consequences.

Recommendation Five

The research community should work together to realize Open Science by Design to advance science and help science better serve the needs of society.

Implementation Actions

- The federal government should revisit and update its open science policy, which is expressed in the 2013 and 2014 OSTP memoranda.
- Funders, institutions, and researchers should align policies and incentives to realize open publication, including rights-retention provisions.
- Research funders should support the establishment of a consortium of research community stakeholders to develop additional concrete methods for implementing open science by design.
- Professional societies—individually and collectively—should work to transition from current business models to new ones that foster open science by design.
- Journal editors should work with publishers to transition from current business models to new ones that foster open science by design.
- Research funders should explore innovative means to support the transition from subscription-based systems to new publication strategies that enable open science by design.
- Librarians should work together with other members of the research community to promote and implement open science by design.
- The research community should develop tools and other applications that depend on the long-term availability of open research products, thereby providing new sources of revenue for the private sector, enhancing the value of research products, and leading to an acceleration of scientific progress.

References

23 Things. 2018. "23 Things: Libraries for Research Data", Research Data Alliance Libraries for Research Data Interest Group. Online. Available at https://rd-alliance.org/system/files/documents/23Things_Libraries_For_Data_Management.pdf. Accessed April 13, 2018.

985 Scientists. 2018. Don't Restrict EPA's Ability to Rely on Science. Letter to EPA Administrator Scott Pruitt. April 23. Online. Available at https://s3.amazonaws.com/ucs-documents/science-and-democracy/secret-science-letter-4-23-2018.pdf. Accessed May 30, 2018.

AAAS (American Association for the Advancement of Science). 2018. Historical Trends in Federal R&D. Online. Available at https://www.aaas.org/page/historical-trends-federal-rd#Char. Accessed April 16, 2018.

AARNet (Australia's Academic and Research Network). 2017. Homepage. Online. Available at https://www.aarnet.edu.au. Accessed October 12, 2017.

AAS (African Academy of Sciences). 2018. AAS Open Research. Online. Available at https://aasopenresearch.org. Accessed May 31, 2018.

AAU (Association of American Universities) and APLU (Association of Public and Land-grant Universities). 2017. AAU-APLU Public Access Working Group: Report and Recommendations. Online. Available at https://www.aau.edu/sites/default/files/AAU-Files/Key-Issues/Intellectual-Property/Public-Open-Access/AAU-APLU-Public-Access-Working-Group-Report.pdf. Accessed January 5, 2018.

Abbott, A., D. Cyranoski, N. Jones, B. Maher, Q. Schiermeier, and R. Van Noorden. 2010. Metrics: Do metrics matter? *Nature* 465(7300):860-862. doi: 10.1038/465860a.

Abbott, B. P., et al. 2016. Observation of gravitational waves from a binary black hole merger. *Physical Review Letters* 116(6):061102. doi: 10.1103/PhysRevLett.116.061102.

Académie des Sciences, Leopoldina, and The Royal Society. 2016. Statement on Scientific Publications by Three National Academies. Online. Available at https://royalsociety.org/~/media/news/2016/Statement%20on%20scientific%20publications.pdf. Accessed February 7, 2018.

Académie des Sciences, Leopoldina, and Royal Society. 2017. Statement by Three National Academies on Good Practice in the Evaluation of Researchers and Research Programmes. Online. Available at https://royalsociety.org/~/media/policy/Publications/2017/08-12-2017-royal-society-leopoldina-and-academie-des-sciences-call-for-more-support-for-research-evaluators.pdf. Accessed February 7, 2018.

Adams, C. 2018. SPARC Innovator: The Bill & Melinda Gates Foundation. Online. Available at https://sparcopen.org/our-work/innovator/gates-foundation. Accessed May 25, 2018.

Aghion, P., M. Dewatripont, J. Kolev, F. Murray, and S. Stern. 2010. The public and private sectors in the process of innovation: Theory and evidence from the mouse genetics revolution. *American Economic Review* 100(2):153-158.

Aghion, P., A. Bergeaud, T. Boppart, P. Klenow, and H. Li. 2016. Missing growth from creative destruction. Mimeo, London School of Economics.

AGU (American Geophysical Union). 2012. AGU partnership with Wiley-Blackwell on journals and book publishing. News release. July 20. Online. Available at https://about.agu.org/president/presidents-message-archive/agu-partnership-with-wiley-blackwell. Accessed June 22, 2018.

AGU. 2013. Update on AGU Publishing: A Focus on Open Access. Online. Available at https://publications.agu.org/announcement/update-on-agu-publishing-a-focus-on-open-access. Accessed January 5, 2018.

AGU. 2016. Promoting Discovery in Earth and Space Science for the Benefit of Humanity. Online. Available at https://sciencepolicy.agu.org/files/2016/12/016_3871_Data-Geosciences-Fact-Sheet.pdf. Accessed March 29, 2018.

AGU. 2017a. Open Access. Online. Available at https://publications.agu.org/open-access. Accessed January 5, 2018.

AGU. 2017b. Where Will Your Samples Go? Online. Available at https://agu.confex.com/agu/fm17/meetingapp.cgi/Paper/208976. Accessed March 29, 2018.

Albert, K. M. 2006. Open access: Implications for scholarly publishing and medical libraries. *Journal of the Medical Library Association* 94(3):253-262.

Alberts, B., M. W. Kirschner, S. Tilghman, and H. Varmus. 2014. Rescuing US biomedical research from its systemic flaws. *Proceedings of the National Academy of Sciences of the United States of America* 111(16):5773-5777.

Altman, M., and M. Crosas. 2013. The evolution of data citation: From principles to implementation. *IASSIST Quarterly*:62-70.

Anderson, I. 2017. Getting to Open: Challenges, Drivers, and Opportunities for Transforming The Global Publication System. Presentation to the National Academies of Sciences, Engineering, and Medicine's Committee on Toward an Open Science Enterprise, Public Symposium. September 18, 2017.

Anderson, K. 2016. Guest Post: Kent Anderson UPDATED—96 Things Publishers Do (2016 Edition). Online. Available at https://scholarlykitchen.sspnet.org/2016/02/01/guest-post-kent-anderson-updated-96-things-publishers-do-2016-edition. Accessed June 6, 2018.

Anderson, K. 2018. The Race to the Bottom—Short-term Bargains versus Long-term Vitality. Online. Available at https://scholarlykitchen.sspnet.org/2018/04/23/the-race-to-the-bottom-short-term-bargains-versus-long-term-vitality. Accessed June 6, 2018.

ANDS (Australian National Data Service). 2017. Online. Available at http://www.ands.org.au. Accessed October 12, 2017.

Arza, V., and M. Fressoli. 2017. Systematizing benefits of open science practices. *Information Services & Use* 37:463–474. DOI:10.3233/ISU-170861.

Association of Research Libraries. 2018. The Copyright Permissions Culture in Software Preservation and Its Implications for the Cultural Record. Online. Available at http://www.arl.org/publications-resources/4468-the-copyright-permissions-culture-in-software-preservation-and-its-implications-for-the-cultural-record#.WwhPYdMzWUn. Accessed May 25, 2018.

Aufderheide, P., B. Butler, K. Cox, and P. Jaszi. 2018. The Copyright Permissions Culture in Software Preservation and Its Implications for the Cultural Record. Online. Available at http://www.arl.org/storage/documents/2018.02.09_CopyrightPermissionsCulture.pdf. Accessed May 25, 2018.

Azoulay, P., W. Ding, and T. Stuart. 2009. The impact of academic patenting on the rate, quality and direction of (public) research output. *The Journal of Industrial Economics* 57(4):637-676.

Baker, M. 2016. Why scientists must share their research code. *Nature* doi:10.1038/nature.2016.20504.

Barbaro, M., and T. Zeller. 2006. A face is exposed for AOL searcher No. 4417749. *The New York Times*, August 9, 2006.

Barbera, R., S. J. E. Taylor, T. M. Banda, B. Becker, A. Cornea, J. Eriksen, L. L. Gustafsson, A. Nungu, O. Oaiya, B. Pehrson, R. Ricceri, and M. Torrisi. 2015. Energising Scientific Endeavour through Science Gateways and e-Infrastructures in Africa: The SciGaIA project. Online. Available at http://oar.sci-gaia.eu/record/108/files/PUBLICATIONSSCIGAIA-2015-015.pdf. Accessed June 28, 2018.

Barnes, N., D. Jones, P. Norvig, C. Neylon, R. Pollock, J. Jackson, V. Stodden, and P. Suber. 2016. Science Code Manifesto. Online. Available at http://sciencecodemanifesto.org. Accessed March 22, 2018.

Bastian, H. 2017. Bias in Open Science Advocacy: The Case of Article Badges for Data Sharing. Online. Available at http://blogs.plos.org/absolutely-maybe/2017/08/29/bias-in-open-science-advocacy-the-case-of-article-badges-for-data-sharing. Accessed March 23, 2018.

Bastian, H. 2018. A Reality Check on Author Access to Open Access Publishing. Online. Available at http://blogs.plos.org/absolutely-maybe/2018/04/02/a-reality-check-on-author-access-to-open-access-publishing. Accessed June 4, 2018.

Beall, J. 2016. Essential information about predatory publishers and journals. *International Higher Education* 86:2-3. Online. Available at https://ejournals.bc.edu/ojs/index.php/ihe/article/viewFile/9358/8368. Accessed December 4, 2017.

Beall, J. 2017. Potential, possible, or probable predatory scholarly open-access publishers. Online. Available at https://beallslist.weebly.com. Accessed December 1, 2017.

Beel, J., B. Gipp, S. Langer, and M. Genzmehr. 2011. Docear: An academic literature suite for searching, organizing and creating academic literature. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. *Joint Conference on Digital Libraries* 11:465-466.

Begley, G., and L. M. Ellis. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483:531-533. doi:10.1038/483531a.

Berg, J. M., N. Bhalla, P. E. Bourne, M. Chalfie, D. G. Drubin, J. S. Fraser, C. W. Greider, M. Hendricks, C. Jones, R. Kiley, S. King, M. W. Kirschner, H. M. Krumholz, R. Lehmann, M. Leptin, B. Pulverer, B. Rosenzweig, J. E. Spiro, M. Stebbins, C. Strasser, S. Swaminathan, P. Turner, R. D. Vale, K. VijayRaghavan, and C. Wolberger. 2016. Preprints for the life sciences. *Science* 352(6288):899-901.

Berghmans, S., H. Cousijn, G. Deakin, I. Meijer, A. Mulligan, A. Plume, S. de Rijcke, A. Rushforth, C. Tatum, T. van Leeuwen, and L. Waltman. 2017. Open Data: The Researcher Perspective. Leiden University's Centre for Science and Technology Studies, Elsevier, and Universiteit Leiden. Online. Available at https://www.elsevier.com/__data/assets/pdf_file/0004/281920/Open-data-report.pdf. Accessed May 25, 2018.

Bergstrom, T. C. 2001. Free labor for costly journals? *The Journal of Economic Perspectives* 15(4):183-198.

Bergstrom, T. C., P. N. Courant, R. P. McAfee, and M. A. Williams. 2014. Evaluating big deal journal bundles. *Proceedings of the National Academy of Sciences of the United States of America* 111(26):9425-9430.

Berman, F., and V. Cerf. 2013. Who will pay for public access to research data? *Science* 341(6146):616-617. DOI: 10.1126/science.1241625.

Bethesda Statement on Open Access Publishing. 2003. Available at http://legacy.earlham.edu/~peters/fos/bethesda.htm. Accessed March 29, 2018.

BibTeX. 2018. Your BibTeX Resource. Online. Available at http://www.bibtex.org. Accessed March 22, 2018.

Bilder, G., J. Lin, and C. Neylon. 2015. Principles for Open Scholarly Infrastructures. Science in the Open. Online. Available at http://cameronneylon.net/blog/principles-for-open-scholarly-infrastructures. Accessed March 30, 2018.

Bill & Melinda Gates Foundation. 2017. Bill & Melinda Gates Foundation Open Access Policy. Online. Available at https://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy. Accessed January 5, 2018.

Bill & Melinda Gates Foundation. 2018. Gates Open Research. Online. Available at https://gatesopenresearch.org. Accessed January 8, 2018.

BITSS (Berkeley Initiative for Transparency in the Social Sciences). 2018. Mission and Objectives. Online. Available at https://www.bitss.org/about. Accessed February 14, 2018.

Björk, B-C. 2017a. Scholarly journal publishing in transition—from restricted to open access. *Electronic Markets* 27(2):101-109.

Björk, B-C. 2017b. Gold, green, and black open access. *Learned Publishing* 30(2):173-175.

BOAI (Budapest Open Access Initiative). 2002. Read the Budapest Open Access Initiative. Online. Available at http://www.budapestopenaccessinitiative.org/read. Accessed March 29, 2018.

BOAI. 2012. Ten years on from the Budapest Open Access Initiative: Setting the default to open. Online. Available at http://www.budapestopenaccessinitiative.org/boai-10-recommendations. Accessed November 20, 2017.

Bonazzi V. R., and P. E. Bourne. 2017. Should biomedical research be like Airbnb? *PLOS Biology* 15(4):e2001818.

Borgman, C. L. 2010. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: The MIT Press.

Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: The MIT Press.

Bosman, J., and B. Kramer. 2018. Open access levels: A quantitative exploration using Web of Science and oaDOI data. PeerJ Preprints. Online. Available at https://peerj.com/preprints/3520. Accessed March 15, 2018.

Bourne, P. E., J. K. Polka, R. D. Vale, and R. Kiley. 2017. Ten simple rules to consider regarding preprint submission. *PLOS Computational Biology* 13(5):e1005473. Online. Available at https://doi.org/10.1371/journal.pcbi.1005473. Accessed November 9, 2017.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. P. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics* 29:365-371. doi:10.1038/ng1201-365.

Buckley, K. 2017. Open Access to Harvard Research. Online. Available at http://library.harvard.edu/02282017-1546/open-access-harvard-research. Accessed November 22, 2017.

Buranyi, S. 2017. The long read: Is the staggeringly profitable business of scientific publishing bad for science? Online. Available at https://www.theguardian.com/science/2017/jun/27/profitable-business-scientific-publishing-bad-for-science. Accessed January 22, 2018.

Burwell, S. M., S. VanRoekel, T. Park, and D. J. Mancini. 2013. Open Data Policy-Managing Information as an Asset. Executive Office of the President. Office of Management and Budget. Online. Available at https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf. Accessed March 30, 2018.

Butler, D. 2017. Gates to launch open-access publishing site. *Nature* 543:599.

Byrne, M. 2017. Making Progress Toward Open Data: Reflections on Data Sharing at PLOS ONE. Online. Available at http://blogs.plos.org/everyone/2017/05/08/making-progress-toward-open-data. Accessed May 25, 2018.

Cafarella, M. J., A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. 2008. WebTables: Exploring the power of tables on the Web. *Proceedings of the Very Large Database Endowment* 1:538-549.

Camerer, C. F., A. Dreber, E. Forsell, T-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351(6280):1433-1436.

Carroll, M. W. 2011. Why full open access matters. *PLOS Biology* 9(11):e1001210.

Carroll, M. W. 2015. Sharing research data and intellectual property law: A primer. *PLOS Biology* 13(8):e1002235.

Carvalho, J. 2017. FOSTER: Training Resources on Open Science. Online. Available at https://helios-eie.ekt.gr/EIE/bitstream/10442/15545/1/FOSTERplus-Training-resources-on-open-access.pdf. Accessed February 14, 2018.

Casadevall, A, and F. C. Fang. 2015. Impacted science: Impact is not importance. *mBio* 6(5):e01593-15.

Cavoukian, A., and M. Chanliau. 2013. Privacy and security by design: A convergence of paradigms. Ontario, Canada: Office of the Privacy Commissioner.

Cavusoglu, H., and S. Raghunathan. 2007. Efficiency of vulnerability disclosure mechanisms to disseminate vulnerability knowledge. *IEEE Transactions on Software Engineering* 33(3). DOI: 10.1109/TSE.2007.26.

CDC (Centers for Disease Control and Prevention). 2003. HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services. *Morbidity and Mortality Weekly Report* 52:1-12.

CEDAR (Center for Expanded Data Annotation and Retrieval). 2018. Better Data for Better Science. Online. Available at https://metadatacenter.org. Accessed June 4, 2018.

CENDI. 2017. Implementation of Public Access Programs in Federal Agencies. Online. Available at https://www.cendi.gov/projects/Public_Access_Plans_US_Fed_Agencies.html. Accessed September 15, 2017.

Chandrasekharan, S., S. Kumar, C.M. Valley, and A. Rai. 2009. Proprietary science, open science and the role of patent disclosure: the case of zinc-finger proteins. *Nature Biotechnology* 27(2):140-144.

Chang, A. C., and P. Li. 2015. Is economics research replicable? Sixty published papers from thirteen journals say "usually not." Finance and Economics Discussion Series 2015-083. Washington, DC: Board of Governors of the Federal Reserve System. Online. Available at http://dx.doi.org/10.17016/FEDS.2015.083. Accessed January 10, 2018.

CNI (Coalition for Networked Information). 2017. Rethinking Institutional Repository Strategies. Online. Available at https://www.cni.org/wp-content/.../CNI-rethinking-irs-exec-rndtbl.report.S17.v1.pdf. Accessed January 29, 2018.

COAR (Confederation of Open Access Repositories). 2014. COAR Roadmap: Future Directions for Repository Interoperability. https://www.coar-repositories.org/files/Roadmap_final_formatted_20150203.pdf. Accessed December 19, 2017.

COAR. 2015a. Promoting Open Knowledge and Open Science Report of the Current State of Repositories. Online. Available at https://www.coar-repositories.org/files/COAR-State-of-Repositories-May-2015-final.pdf. Accessed March 28, 2018.

COAR. 2015b. COAR Roadmap Future Directions for Repository Interoperability. Online. https://www.coar-repositories.org/files/Roadmap_final_formatted_20150203.pdf. Accessed January 29, 2018.

CODATA (Committee on Data for Science and Technology). 2016. CODATA Strategy and Achievement 2015-2016. Online. Available at http://www.codata.org/news/128/62/CODATA-Prospectus-Strategy-and-Achievement-2015-2016. Accessed October 11, 2017.

CODATA. 2017. About CODATA. Online. Available at http://www.codata.org/about-codata. Accessed October 11, 2017.

CODESRIA (Council for the Development of Social Science Research in Africa). 2016. Dakar Declaration on Open Science in Africa and the Global South. Online. Available at http://wiki.lib.sun.ac.za/images/5/50/Dakar-declaration-2016.pdf. Accessed October 12, 2017.

Coffman, L., M. Niederle, and A. J. Wilson. 2017. A proposal to organize and promote replications. *American Economic Review* 107(5):41-45.

Columbia University. 2017. Open Access Policy Frequently Asked Questions. Online. Available at https://scholcomm.columbia.edu/open-access/open-access-policies/frequently-asked-questions. Accessed November 20, 2017.

Congress.gov. 2017. S.1701 - Fair Access to Science and Technology Research Act of 2017. 115th Congress (2017-2018). Online. Available at https://www.congress.gov/bill/115th-congress/senate-bill/1701/text. Accessed March 30, 2018.

Conley, J. P., and M. Wooders. 2009. But what have you done for me lately? Commercial publishing, scholarly communication, and open-access. *Economic Analysis and Policy* 39(1):71-88.

Conniff, R. 2012. When continental drift was considered pseudoscience. *Smithsonian Magazine*. Online. Available at https://www.smithsonianmag.com/science-nature/when-continental-drift-was-considered-pseudoscience-90353214. Accessed June 28, 2018.

COPDESS (Coalition on Publishing Data in the Earth and Space Sciences). 2015. COPDESS Statement of Commitment. Online. Available at http://www.copdess.org/statement-of-commitment. Accessed May 25, 2018.

COPE (Committee on Publication Ethics). 2017. Promoting Integrity in Research and Its Publication. Online. Available at https://publicationethics.org. Accessed December 3, 2017.

Cornell University Library. 2017. arXiv.org. Online. Available at arxiv.org. Accessed November 9, 2017.

COS (Center for Open Science). 2015. Guidelines for Transparency and Openness Promotion (TOP) in Journal Policies and Practices "The TOP Guidelines" Version 1.0.1. Online. Available at https://osf.io/ud578/?show=revision. Accessed March 22, 2018.

COS. 2017. Six New Preprint Services Join a Growing Community across Disciplines to Accelerate Scholarly Communication. Online. Available at https://cos.io/about/news/six-new-preprint-services-join-growing-community-across-disciplines-accelerate-scholarly-communication. Accessed November 10, 2017.

COS. 2018a. Open Science Badges Enhance Openness, a Core Value of Scientific Practice. Online. Available at https://cos.io/our-services/open-science-badges. Accessed March 23, 2018.

COS. 2018b. Registered Reports. Online. Available at https://cos.io/rr. Accessed May 25, 2018.

COS. 2018c. Partner with COS on Grant Funding Proposals. Online. Available at https://cos.io/about/our-partners/partner-cos-grant-funding-proposals. Accessed March 22, 2018.

COS. 2018d. Training Services. Online. Available at https://cos.io/our-services/training-services. Accessed March 22, 2018.

Cousijn, H., A. Kenall, E. Ganley, M. Harrison, D. Kernohan, T. Lemberger, F. Murphy, P. Polischuk, S. Taylor, M. Martone, and T. Clark. 2017. A Data Citation Roadmap for Scientific Publishers. bioRxiv. doi: http://dx.doi.org/10.1101/100784.

Crawford, W. 2016. *Gold Open Access Journals 2011-2015*. Livermore, CA: Cites & Insights Books. Online. Available at https://waltcrawford.name/goaj1115.pdf. Accessed October 20, 2017.

Crawford, W. 2018. *Gold Open Access Journals 2012-2017*. Livermore, CA: Cites & Insights Books. Online. Available at https://waltcrawford.name/cntry1217.pdf. Accessed June 7, 2018.

Cross, J. 2009. Impact factors—the basics. The E-Resources Management Handbook. United Kingdom Serials Group. Online. Available at www.uksg.org/sites/uksg.org/files/19-Cross-H76M463XL884HK78.pdf. Accessed February 23, 2018.

Crotty, D. 2016. The Pay It Forward Project: Confirming What We Already Knew About Open Access. Online. Available at https://scholarlykitchen.sspnet.org/2016/08/09. Accessed December 11, 2017.

DataCite. 2018. Locate, Identify and Cite Research Data with the Leading Global Provider of Dois for Research Data. Online. Available at https://www.datacite.org. Accessed March 22, 2018.

Data Curation Network. 2018. Our Mission. Online. Available at https://sites.google.com/site/datacurationnetwork. Accessed May 25, 2018.

Davis, P. M. 2010. Access, Readership, Citations: A Randomized Controlled Trial of Scientific Journal Publishing. Online. Available at https://ecommons.cornell.edu/bitstream/handle/1813/17788/Davis%2C%20Philip.pdf?sequence=1. Accessed January 11, 2018.

Davis, P. M. 2011. Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *Federation of American Societies for Experimental Biology* 25(7):2129-2134. doi: 10.1096/fj.11-183988.

Davis, P. 2017. Scientific Reports Overtakes PLOS ONE as Largest Megajournal. Online. Available at https://scholarlykitchen.sspnet.org/2017/04/06/scientific-reports-overtakes-plos-one-as-largest-megajournal. Accessed May 25, 2018.

Davis, P. M., B. V. Lewenstein, D. H. Simon, J. G. Booth, and M. J. L. Connolly. 2008. Open access publishing, article downloads and citations: randomized trial. *BMJ* 337:a568. doi: https://doi.org/10.1136/bmj.a568.

dbGaP (Database of Genotypes and Phenotypes). 2018. Online. Available at https://www.ncbi.nlm.nih.gov/gap. Accessed June 6, 2018.

Digital Curation Centre. 2018. Disciplinary metadata standards. Online. Available at http://www.dcc.ac.uk/resources/metadata-standards. Accessed March 22, 2018.

Do, L., and W. Mobley. 2015. Single Figure Publications: Towards a novel alternative format for scholarly communication [version 1; referees: not peer reviewed]. F1000Research 2015, 4:268. doi: 10.12688/f1000research.6742.1.

DOAJ (Directory of Open Access Journals). 2018. Homepage. Online. Available at https://doaj.org. Accessed March 29, 2018.

Docear. 2018. The academic literature suite. Online. Available at https://www.docear.org. Accessed March 22, 2018.

DONA Foundation. 2018. Homepage. Online. Available at https://www.dona.net. Accessed March 22, 2018.

DORA (Declaration on Research Assessment). 2013. San Francisco Declaration on Research Assessment: Putting Science into the Assessment of research. Online. Available at http://www.ascb.org/files/SFDeclarationFINAL.pdf. Accessed February 23, 2018.

Dryad. 2018. Open data best practices. Online. Available at https://datadryad.org. Accessed March 22, 2018.

Dutch Techcentre for Life Sciences. 2016. Global Open FAIR Implementation Nodes. Online. Available at https://www.dtls.nl/wp-content/uploads/2016/11/3.-GO-FAIR-cover-proposal.for-comments.docx. Accessed October 13, 2017.

Dwork, C. 2008. Differential privacy: A survey of results. Pp. 1-19 in *Theory and Applications of Models of Computation*. TAMC 2008. Lecture Notes in Computer Science 4978, M. Agrawal, D. Du, Z. Duan, and A. Li, eds. Berlin, Heidelberg: Springer.

EC (European Commission). 2012. Commission Staff Working Document: Impact Assessment. Accompanying the Document: Commission Recommendation on Access to and Preservation of Scientific Information in the Digital Age. Online. Available at http://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/swd/2012/0222/COM_SWD(2012)0222_EN.pdf. Accessed April 16, 2018.

EC. 2016. *Realising the European Open Science Cloud: First Report and Recommendations of the Commission High Level Expert Group on the European Open Science Cloud*. Brussels, Belgium: EC.

EC. 2017a. EOSC (European Open Science Cloud) Declaration: New Research & Innovation Opportunities. Online. Available at https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf. Accessed January 11, 2018.

EC. 2017b. Evaluation of Research Careers fully acknowledging Open Science Practices: Rewards, incentives and/or recognition for researchers practicing Open Science. Written by the Working Group on Rewards under Open Science. Luxembourg: European Union.

EC. 2017c. Information Note: Towards a Horizon 2020 Platform for Open Access. Online. Available at https://ec.europa.eu/research/openscience/pdf/information_note_platform_public.pdf#view=fit&pagemode=none. Accessed April 13, 2018.

EC. 2017d. Europe's Future: Open Innovation, Open Science, and Open to the World. Reflections of the Research, Innovation and Science Policy Experts (RISE) High Level Group. Brussels, Belgium: EC.

EC. 2017e. Report on the Governance and Financial Schemes for the European Open Science Cloud: Adopted by the Open Science Policy Platform- May 2017. Online. Available at https://ec.europa.eu/research/openscience/pdf/ospp_euro_open_science_cloud_report-.pdf. Accessed October 13, 2017.

EC. 2017f. Providing Researchers with the Skills and Competencies They Need to Practice Open Science: Open Science Skills Working Group Report. Online. Available at https://ec.europa.eu/research/openscience/pdf/os_skills_wgreport_final.pdf. Accessed March 22, 2018.

EC. 2018a. Open Access to Publications. Online. Available at https://ec.europa.eu/research/openscience/index.cfm?pg=access&section=monitor. Accessed April 23, 2018.

EC. 2018b. Commission Staff Working Document: Implementation Roadmap for the European Open Science Cloud. Online. Available at https://ec.europa.eu/research/op

enscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf#view=fit&pagemode=none. Accessed April 23, 2018.

EC. 2018c. Open Science Monitor. Online. Available at https://ec.europa.eu/research/openscience/index.cfm?pg=drivers&section=monitor. Accessed February 23, 2018.

EC. 2018d. Study in support of the evaluation of Directive 96/9/EC on the legal protection of databases. Online. Available at http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51599. Accessed June 8, 2018.

The Economist. 2013. Unreliable research: Trouble at the lab. Online. Available at https://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble. Accessed February 23, 2018.

Edwards, C. 2017. Federal R&D Funding. Online. Available at https://www.downsizinggovernment.org/federal-rd-funding. Accessed April 16, 2018.

Eibe, F., M. Hall, and I. H. Witten. 2016. The WEKA workbench. Online Appendix for *Data Mining: Practical Machine Learning Tools and Techniques*, Fourth Edition. Burlington, MA: Morgan Kaufmann.

e-IRG (e-Infrastructure Reflection Group). 2016. Long Tail of Data: e-IRG Task Force Report. The Hague, The Netherlands: e-IRG Secretariat. Online. Available at http://e-irg.eu/documents/10920/238968/LongTailOfData2016.pdf. Accessed December 11, 2017.

EOS (Earth and Space Science News). 2017. Connecting Scientific Data and Real-World Samples. Online. Available at https://eos.org/meeting-reports/connecting-scientific-data-and-real-world-samples. Accessed March 29, 2018.

EPA (U.S. Environmental Protection Agency). 2018. Transparency in Regulatory Decisionmaking. 40 CFR 30. Available at https://www.gpo.gov/fdsys/pkg/FR-2018-04-30/pdf/2018-09078.pdf. Accessed June 7, 2018.

Erway, R., and A. Rinehart. 2016. *If You Build It, Will They Fund? Making Research Data Management Sustainable*. Dublin, Ohio: OCLC Research. Online. Available at https://www.oclc.org/content/dam/research/publications/2016/oclcresearch-making-research-data-management-sustainable-2016.pdf. Accessed March 29, 2018.

ESO (European Southern Observatory). 2017. ESO Endorses the European Open Science Cloud Declaration. Online. Available at http://www.eso.org/public/australia/announcements/ann17068/?lang. Accessed October 9, 2017.

EUA (European University Association). 2015. EUA's Open Access Checklist for Universities: A Practical Guide on Implementation. Online. Available at http://www.eua.be/Libraries/publications-homepage-list/Open_access_report_v3.pdf?sfvrsn=4. Accessed November 22, 2017.

Evans, J. A., and J. Reimer. 2009. Open Access and Global Participation in Science. Science 323(5917):1025.

Eveleth, R. 2014. Free Access to Science Research Doesn't Benefit Everyone. *The Atlantic*. Online. Available at https://www.theatlantic.com/technology/archive/2014/12/free-access-to-science-research-doesnt-benefit-everyone/383875. Accessed February 23, 2018.

Eysenbach, G. 2006. Citation advantage of open access articles. *PLOS Biology* 4(5):e157. doi:10.1371/journal.pbio.0040157.

F1000 Research. 2018. Article processing charges. Available at https://f1000research.com/for-authors/article-processing-charges. Accessed June 7, 2018.

Fabrizio, K., and A. DiMinin. 2008. Commercializing the laboratory: Faculty patenting and the open science environment. *Research Policy* 30(30). Online. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1129542. Accessed January 10, 2018.

FAIRsharing. 2017. FAIRsharing Policies: A Catalogue of Data Preservation, Management and Sharing Policies from International Funding Agencies, Regulators and Journals. Online. Available at https://fairsharing.org/policies. Accessed October 20, 2017.

Fanelli, D. 2018. Is science really facing a reproducibility crisis? *Proceedings of the National Academy of Sciences* Mar 2018, 201708272; DOI:10.1073/pnas.1708272114.

Fang, F. C., and A. Casadevall. 2015. Competitive science: Is competition ruining science? *Infection and Immunity* 83:1229-1233. doi:10.1128/IAI.02939-14.

FDA (Food and Drug Administration). 2007. FDAAA 801 and the Final Rule. Online. Available at https://clinicaltrials.gov/ct2/manage-recs/fdaaa. Accessed May 31, 2018.

FDAAA (Food and Drug Administration Amendments Act) Trials Tracker. 2018. Who's Sharing Their Clinical Trial Results? Online. Available at https://fdaaa.trialstracker.net. Accessed March 23, 2018.

Fecher, B. and S. Friesike. 2014. Open Science: One Term, Five Schools of Thought. Proceedings of the 1st International Conference on Internet, Brussels.

Fehder, D. C., F. Murray, and S. Stern. 2014. Intellectual property rights and the evolution of scientific journals as knowledge platforms. *International Journal of Industrial Organization* 36:83-94.

Fenner, M., M. Crosas, and J. S. Grethe. 2016. A Data Citation Roadmap for Scholarly Data Repositories. Available at http://dx.doi.org/10.1101/097196. Accessed March 22, 2018.

Figshare. 2017. The State of Open Data 2017: A selection of analyses and articles about open data, curated by Figshare. Digital Science. DOI: https://doi.org/10.6084/m9.figshare.5481187.

Figshare. 2018. Homepage. Online. Available at https://figshare.com. Accessed March 22, 2018.

FORCE11. 2014. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Online. Available at https://doi.org/10.25490/a97f-egyk. Accessed March 22, 2018.

Ford, P. 2018. GitHub is Microsoft's $7.5 Billion Undo Button. Bloomberg. June 6. Online. Available at https://www.bloomberg.com/news/articles/2018-06-06/github-is-microsoft-s-7-5-billion-undo-button. Accessed June 6, 2018.

Forde, J., C. Holdgraf, and Y. Panda. 2018. Post-training evaluation with Binder. Conference on Fairness, Accountability, and Transparency. Online. Available at https://fatconference.org/static/tutorials/forde_binder18.pdf. Accessed March 22, 2018.

FOSTER (Facilitate Open Science Training for European Research). 2018. What Is Open Science? Introduction. Online. Available at https://www.fosteropenscience.eu. Accessed March 29, 2018.

Foster, E. D., and A. Deardorff. 2017. Open science framework. *Journal of the Medical Library Association* 105 (2):203-206.

FREYA. 2018. The FREYA project. Online. Available at https://www.project-freya.eu/en. Accessed May 25, 2018.

Fyfe, A., J. McDougall-Waters, and N. Moxham. 2015. 350 years of scientific periodicals. Notes and Records 69:227-239. doi:10.1098/rsnr.2015.0036.

G7 (Group of Seven). 2017. G7 Science Ministers' Communiqué. Turin, Italy, 27-28 September, 2017. Online. Available at http://www.g7italy.it/sites/default/files/documents/G7%20Science%20Communiqu%C3%A9.pdf. Accessed October 9, 2017.

Gaboardi, M., H. W. Lim, R. M. Rogers, and S. P. Vadhan. 2016. Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing. Proceedings of the 33rd International Conference on Machine Learning, New York, NY. JMLR: Workshops and Proceedings 48.

Galiani, S., P. Gertler, and M. Romero. 2017. Incentives for Replication in Economics. Online. Available at http://www.paulgertler.com/uploads/4/7/5/1/47512443/galiani-gertler-romano_replication_in_economics_nber.pdf. Accessed January 10, 2018.

Gaule, P., and N. Maystre. 2011. Getting cited: Does open access help? *Research Policy* 40(10):1332-1338. doi:10.1016/j.respol.2011.05.025.

Gelman, A. 2018. Can you criticize science (or do science) without looking like an obsessive? Maybe not. *Slate*. March 26. Online. Available at https://slate.com/technology/2018/03/its-hard-to-criticize-science-without-looking-like-an-obsessive.html. Accessed May 31, 2018.

Ginther, D. K., J. Basner, U. Jensen, J. Schnell, R. Kington, and W. T. Schaffer. Publications as Predictors of Racial and Ethnic Differences in NIH Research Awards. Mimeo, University of Kansas.

GitHub. 2018. Built for Developers. Online. Available at https://github.com. Accessed March 22, 2018.

GO FAIR (Global Open FAIR). 2018. GO FAIR: a bottom-up international approach. Online. Available at https://www.go-fair.org. Accessed April 23, 2018.

Goodman, L. 2018. GigaScience Wins 2018 PROSE Award for Innovation in Publishing. Online. Available at http://gigasciencejournal.com/blog/gigascience-prose-award-for-innovation. Accessed June 6, 2018.

Gordon, G. 2016. SSRN—the leading social science and humanities repository and online community—joins Elsevier. Online. Available at https://www.elsevier.com/connect/ssrn-the-leading-social-science-and-humanities-repository-and-online-community-joins-elsevier. Accessed April 13, 2018.

GSA (The Geological Society of America). 2018. Geoscience Data Preservation. Online. Available at https://www.geosociety.org/documents/gsa/positions/pos9_dataPres.pdf. Accessed March 29, 2018.

The Guardian. 2018. Performance-driven culture is ruining scientific research. Online. Available at https://www.theguardian.com/higher-education-network/2018/feb/16/performance-driven-culture-is-ruining-scientific-research. Accessed February 23, 2018.

Hahn, R. 2018. Many mocked this Scott Pruitt proposal. They should have read it first. *Washington Post*. May 10.

Hamermesh, D. S. 2017. Replication in labor economics: Evidence from data, and what it suggests. *American Economic Review* 107(5):37-40.

Hansen, J. 2017. Providing Support and Solutions for Open Science to Achieve Impact. Presentation to the National Academies of Sciences, Engineering, and Medicine's Committee on Toward an Open Science Enterprise, Public Symposium. September 18, 2017.

Harvard Dataverse. 2018. Share, Archive, and Get Credit for Your Data. Find and Cite Data Across All Research Fields. Online. Available at https://dataverse.harvard.edu. Accessed January 29, 2018.

Harvard Library Office for Scholarly Communication. 2017. Open Access Policies. Online. Available at https://osc.hul.harvard.edu/policies. Accessed November 28, 2017.

Healy, K. 2011. Choosing your workflow applications. *The Political Methodologist* 18(2):9-18.

Heber, J. 2017. Advocating Open Science at PLOS. Presentation to the National Academies of Sciences, Engineering, and Medicine's Committee on Toward an Open Science Enterprise, Public Symposium. September 18, 2017.

Hefferon, J. and K. Berry. 2009. The TeX family in 2009. *Notices of the American Mathematical Society* 56(3):348-354.

Heidorn, P. B. 2008. Shedding light on the dark data in the long tail of science. *Library Trends* 57(2):280-299.

Hendricks, G. 2015. Crossref to Auto-Update ORCID Records. Online. Available at https://www.crossref.org/blog/crossref-to-auto-update-orcid-records. Accessed May 25, 2018.

Hess, C., and E. Ostrom. 2003. Ideas, artifacts, and facilities: Information as a common-pool resource. *Law and Contemporary Problems* 66(1&2): 111-146.

Hicks, D., P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols. 2015. Bibliometrics: The Leiden Manifesto for research metrics. *Nature*. Online. Available at https://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351. Accessed February 23, 2018.

Hitchcock, S. 2018. The effect of open access and downloads ('hits') on citation impact: a bibliography of studies. Online. Available at http://opcit.eprints.org/oacitation-biblio.htm. Accessed March 23, 2018.

Holdren, J. 2017. Public Access- Report to Congress- January 2017. Office of Science and Technology Policy. Online. Available at https://obamawhitehouse.archives.gov/sites/default/files/microsites/public_access-report_to_congress-jan2017-final.pdf. Accessed September 15, 2017.

Holmes, G., A. Donkin, and I. H. Witten. 1994. Weka: A machine learning workbench. Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems.

Howard, J. 2013. Rise of "altmetrics" revives questions about how to measure impact of research. *Chronicle of Higher Education*. Online. Available at https://www.chronicle.com/article/Rise-of-Altmetrics-Revives/139557. Accessed March 29, 2018.

Hrynaszkiewicz, I., and M. J. Cockerill. 2012. Open by default: A proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. *BMC Research Notes* 7(5):494.

Huang, Y., Y. Liu, C. Zheng, and C. Shen. 2017. Investigation of cross-contamination and misidentification of 278 widely used tumor cell lines. *PLOS One*. Online. Available at https://doi.org/10.1371/journal.pone.0170384. Accessed February 23, 2018.

Hudson-Vitale, C. R., R. P. Johnson, J. Ruttenberg, and J. R. Spies. 2017. SHARE: Community-focused infrastructure and a public goods, scholarly database to advance access to research. *D-Lib Magazine* 23(5/6).

Hutchins, B. I., X. Yuan, J. M. Anderson, and G. M. Santangelo. 2016. Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *PLOS Biology*. Online. Available at https://doi.org/10.1371/journal.pbio.1002541. Accessed February 23, 2018.

IAC-IAP (InterAcademy Council and InterAcademy Partnership). 2012. The Global Network of Science Academies. Responsible Conduct in the Global Research Enterprise. Amsterdam, Netherlands: IAC.

ICSU (International Council for Science)-WDS (World Data System). 2017. Trusted Data Services for Global Science. Online. Available at https://www.icsu-wds.org. Accessed October 10, 2017.

iDigBio (Integrated Digitized Biocollections). 2018. iDigBio website. Online. Available at https://www.idigbio.org. Accessed June 22, 2018.

Ihaka, R. 2010. R: Lessons learned, directions for the future. JSM (Joint Statistical Meetings) Proceedings 2010: Statistical Computing Section.

Inglis, J. 2017. The bioRxiv Preprint Server: An Open Science Initiative for the Life Sciences. Presentation to the National Academies of Sciences, Engineering, and Medicine's Committee on Toward an Open Science Enterprise, Public Symposium. September 18, 2017.

INLEXIO. 2017. Preprint Servers: Challenges and Consequences. Online. Available at https://www.inlexio.com/preprint-servers-challenges-consequences. Accessed November 10, 2017.

Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLOS Medicine* 2(8):e124. Online. Available at https://doi.org/10.1371/journal.pmed.0020124. Accessed January 10, 2018.

IOM (Institute of Medicine). 2015. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington, DC: The National Academies Press.

IWGSC (Interagency Working Group on Scientific Collections). 2009. Scientific Collections: Mission Critical Infrastructure of Federal Science Agencies. A report of the Interagency Working Group on Scientific Collections. National Science and Technology Council, Committee on Science, Office of Science and Technology Policy. Online. Available at https://usfsc.nal.usda.gov/sites/usfsc.nal.usda.gov/files/IWGSC_GreenReport_FINAL_2009.pdf. Accessed March 29, 2018.

IWGSC. 2016. About IWGSC. Online. Available at https://usfsc.nal.usda.gov/about-iwgsc. Accessed March 29, 2018.

IWGSC. 2018. Agency Documents. Online. Available at https://usfsc.nal.usda.gov/agency-documents-view. Accessed March 29, 2018.

Jacobs, N. 2018. Open Research in 2018, Real or Fake News? Online. Available at https://wonkhe.com/blogs/open-research-in-2018-real-or-fake-news. Accessed April 13, 2018.

Johnston, L. R., J. R. Carlson, P. Hswe, C. Hudson-Vitale, H. Imker, W. Kozlowski, R. K. Olendorf, and C. Stewart. 2017. Data curation network: How do we compare? A snapshot of six academic library institutions' data repository and curation services. *Journal of eScience Librarianship* 6(1):e1102. https://doi.org/10.7191/jeslib.2017.1102.

Jong, S., and K. Slavova. 2014. When publications lead to products: The open science conundrum in new product development. *Research Policy* 43(4):645-654.

Kennison, R., and L. Norberg. 2015. A Network Approach to Scholarly Communication Infrastructure. *EDUCAUSE Review*. Online. Available at https://er.educause.edu/articles/2015/4/a-network-approach-to-scholarly-communication-infrastructure. Accessed March 23, 2018.

Kidwell, M. C., L. B. Lazarević, E. Baranski, T. E. Hardwicke, S. Piechowski, L-S. Falkenberg,

C. Kennett, A. Slowik, C. Sonnleitner, C. Hess-Holden, T. M. Errington, S. Fiedler, and B. A. Nosek. 2016. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. Online. Available at http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002456. Accessed March 23, 2018.

Kimmelman, J., J. S. Mogil, and U. Dirnagl. 2014. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLOS Biology* 12(5): e1001863.

Kittrie, E., A. A. Atienza, R. Kiley, D. Carr, A. MacFarlane, V. Pai, J. Couch, J. Bajkowski, J. F. Bonner, D. Mietchen, and P. E. Bourne. 2017. Developing international open science collaborations: Funder reflections on the Open Science Prize. *PLOS Biology* 15(8):e2002617. https://doi.org/10.1371/journal.pbio.2002617

Kluyver T., B. Ragan-Kelley, and F. Pérez. 2016. *Jupyter Notebooks – A Publishing Format for Reproducible Computational Workflows. Positioning and Power in Academic Publishing: Players, Agents, and Agendas*. Amsterdam, Netherlands: IOS Press.

Kriegeskorte, N., A. Walther, and D. Deca. 2012. An emerging consensus for open evaluation: 18 visions for the future of scientific publishing. *Frontiers in Computational Neuroscience* 6(94):1-5.

Kriesberg, A., K. Huller, R. Punzalan, and C. Parr. 2017. An Analysis of Federal Policy on Public Access to Scientific Research Data. *Data Science Journal* 16:27. Online. Available at http://doi.org/10.5334/dsj-2017-027. Accessed March 30, 2018.

Kunzmann, M., and F. Reckling. 2017. Austrian Science Fund (FWF) Open Access Compliance Monitoring 2016. Online. Available at https://zenodo.org/record/811924#.WrU2ZtMzWmR. Accessed March 23, 2018.

Larivière, V., S. Haustein, and P. Mongeon. 2015. The Oligopoly of Academic Publishers in the Digital Era. PLoS ONE 10(6):e0127502. doi:10.1371/journal.pone.0127502.

Lawson, S. 2015. Fee waivers for open access journals. *Publications* 3:155-167. doi:10.3390/publications3030155.

Lerner, J., and J. Triole. 2000. The Simple Economics of Open Source. NBER Working Paper No. 7600. March.

Lewis, D. W., L. Goetsch, D. Graves, and M. Roy. 2018. Funding Community Controlled Open Infrastructure for Scholarly Communication: The 2.5% Commitment Initiative. Online. Available at https://theidealis.org/funding-community-controlled-open-infrastructure-for-scholarly-communication-the-2-5-commitment-initiative. Accessed March 30, 2018.

LIBER (The European Library Federation). 2012. Ten Recommendations for Libraries to Get Started with Research Data Management. Online. Available at http://liber europe.eu/wp-content/uploads/The%20research%20data%20group%202012%20v7%20final.pdf. Accessed December 19, 2017.

Library Research News. 2018. Web of Science new features. Online. Available at http://blogs.sun.ac.za/libraryresearchnews/2018/01/04/web-of-science-new-features. Accessed March 19, 2018.

Lipton, M. 2006. Merger Waves in the 19th, 20th and 21st Centuries. Online. Available at http://cornerstone-business.com/MergerWavesTorontoLipton.pdf. Accessed April 16, 2018.

Luther, J. 2017. The Stars Are Aligning for Preprints. The Scholarly Kitchen. Online. Available at https://scholarlykitchen.sspnet.org/2017/04/18/stars-aligning-preprints. Accessed November 10, 2017.

Lynch, C. 2003. Institutional repositories: essential infrastructure for scholarship in the digital age. *Association of Research Libraries Bimonthly Report* 226:1-7.

Machanavajjhala, A., J. Gehrke, D. Kifer, and M. Venkitasubramaniam. 2006. L-diversity: Privacy beyond k-anonymity. Proceedings of the 22nd International Conference on Data Engineering. DOI: 10.1109/ICDE.2006.1.

Mailman, M. D., M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, Y. Shao, Z. Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev,

D. Preuss, E. Yaschenko, A. Graeff, J. Ostell, and S. T. Sherry. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics* 39(10):1181-1186.

Malin, B., and L. Sweeney. 2001. Re-identification of DNA through an automated linkage process. *Proceedings of the AMIA Symposium*:423-427.

McCabe, M. J. 2013. Online Access and the Scientific Journal Market: An Economist's Perspective. Draft Report for the National Academy of Sciences. Online. Available at https://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_063400.pdf. Accessed January 12, 2018.

McCabe, M. J., and C. M. Snyder. 2014. Identifying the effect of open access on citations using a panel of science journals. *Economic Inquiry* 52(4):1284-1300.

McCabe, M. J., and C. M. Snyder. 2015. Does online availability increase citations? Theory and evidence from a panel of economics and business journals. *The Review of Economics and Statistics* 97(1):144-165.

McCabe, M. J., C. M. Snyder, and A. Fagin. 2013. Open access versus traditional journal pricing: Using a simple "platform market" model to understand which will win (and which should). *Journal of Academic Librarianship* 39:11-19.

McKiernan, E., P. E. Bourne, C. T. Brown, S. Buck, A. Kenall, J. Lin, D. McDougall, B. A. Nosek, K. Ram, C. K. Soderberg, J. R. Spies, K. Thaney, A. Updegrove, K. H. Woo, and T. Yarkoni. 2016. How open science helps researchers succeed. *eLife* 5:e16800. http://doi.org/10.7554/eLife.16800.

McNutt, M., K. Lehnert, B. Hanson, B. A. Nosek, A. M. Ellison, and J. L. King. 2016. Liberating field science samples and data: Promote reproducibility by moving beyond "available upon request." *Science* 351(6277):1024-1026.

Meadows, A. 2016. Everything you ever wanted to know about ORCID. *College and Research Libraries* 77(1).

MedOANet (Mediterranean Open Access Network). 2013. MedOANet Guidelines for Implementing Open Access Policies: For Research Performing and Research Funding Organizations. Online. Available at http://medoanet.eu/sites/www.medoanet.eu/files/documents/MED2013_GUIDELine_dp_EN_ws.pdf. Accessed November 22, 2017.

Merton, R. K. 1942. The Normative Structure of Science. Pp. 267-278 in The Sociology of Science: Theoretical and Empirical Investigations. N. W. Storer, ed. Chicago, IL: The University of Chicago Press. Online. Available at https://www.collier.sts.vt.edu/5424/pdfs/merton_1973.pdf. Accessed June 28, 2018.

Miguel, E., L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. Promoting an open research culture. *Science* 348(6242):1422-1425.

MIT (Massachusetts Institute of Technology) Libraries. 2009. MIT Faculty Open Access Policy: Policy adopted by unanimous vote of the faculty on 3/18/2009. Online. Available at https://libraries.mit.edu/scholarly/mit-open-access/open-access-policy. Accessed June 27, 2018.

MIT Libraries. 2018. Data Management. Online. Available at https://libraries.mit.edu/data-management/share/find-repository. Accessed January 29, 2018.

Mogil, J. S., and M. R. Macleod. 2017. No publication without confirmation. *Nature* 542(7642):409-411.

Molloy, J.C. 2011. The Open Knowledge Foundation: Open data means better science. *PLOS Biology* 9(12):e1001195.

Mons, B., C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson. 2017. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services and Use* 37(1):49-56.

Moore, G. E. 1965. Cramming more components onto integrated circuits. *Electronics* 38(8). Online. Available at https://drive.google.com/file/d/0By83v5TWkGjvQkpBcXJKT1I1TTA/view. Accessed March 23, 2018.

Morey, R. D., C. D. Chambers, P. J. Etchells, C. R. Harris, R. Hoekstra, D. Lakens, S. Lewandowsky, C. C. Morey, D. P. Newman, F. D. Schönbrodt, W. Vanpaemel, E-J. Wagenmakers, and R. A. Zwaan. 2016. The Peer Reviewers' Openness Initiative: Incentivizing Open Research Practices through Peer Review. Royal Society Open Science. DOI:10.1098/rsos.150547.

MPDL (Max Planck Digital Library). 2015. Disrupting the Subscription Journals' Business Model. Online. Available at http://pubman.mpdl.mpg.de/pubman/item/escidoc:2148961:7/component/escidoc:2149096/MPDL_OA-Transition_White_Paper.pdf. Accessed December 11, 2017.

MSDSE (Moore-Sloan Data Science Environments). 2018. Moore-Sloan Data Science Environments. Online. Available at http://msdse.org. Accessed March 22, 2018.

Mueller-Langer, F., and R. Watt. 2014. The Hybrid Open Access Citation Advantage: How Many More Cites is a $3,000 Fee Buying You? Max Planck Institute for Innovation & Competition Research Paper 14-02. Online. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2391692. Accessed June 27, 2018.

Mukherjee, A., and S. Stern. 2009. Disclosure or secrecy? The dynamics of open science. *International Journal of Industrial Organization* 27(3):449-462.

Munafò, M. R., B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour* 1(0021). doi:10.1038/s41562-016-0021.

Murray-Rust, P., C. Neylon, R. Pollock, and J. Wilbanks. 2010. Panton Principles, Principles for Open Data in Science. Online. Available at https://pantonprinciples.org. Accessed March 29, 2018.

Murray, F., and S. Stern. 2007. Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior and Organization* 63(4):648-685.

Narayanan, A., and V. Shmatikov. 2008. Robust de-anonymization of large sparse datasets. IEEE Symposium on Security and Privacy. DOI: 10.1109/SP.2008.33.

NAS-NAE-IOM (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine). 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington, DC: The National Academies Press.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2016. *Science Literacy: Concepts, Contexts, and Consequences*. Washington, DC: The National Academies Press.

NASEM. 2017a. *Communicating Science Effectively*. Washington, DC: The National Academies Press.

NASEM. 2017b. *Fostering Integrity in Research*. Washington, DC: The National Academies Press.

NASEM. 2018a. *Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report*. Washington, DC: The National Academies Press.

NASEM. 2018b. Project Information: Reproducibility and Replicability in Science. Online. Available https://www8.nationalacademies.org/cp/projectview.aspx?key=49906. Accessed April 26, 2018.

NASEM. 2018c. International Coordination for Science Data Infrastructure: Proceedings of a Workshop–in Brief. Washington, DC: The National Academies Press.

Nature. 2017. Natural History Collections Face Fight for Survival. *Nature* 544 (7649):137-138.

Nature. 2018. Availability of data, material and methods. Online. Available at http://www.nature.com/authors/policies/availability.html#code. Accessed March 22, 2018.

NERL (NorthEast Research Libraries consortium). 2018. Homepage. Online. Available at http://www.nerl.org. Accessed March 30, 2018.

Neumann J., and J. Brase. 2014. DataCite and DOI names for research data. *Journal of Computer-Aided Molecular Design* 28(10):1035-1041.

Neylon C. 2017. Openness in Scholarship: A Return to Core Values? Proceedings of the 21st International Conference on Electronic Publishing, IOS Press

Nielsen, M. 2011. *Reinventing Discovery: The New Era of Networked Science*. Princeton, NJ: Princeton University Press.

NIH (National Institutes of Health). 2008. Analysis of Comments and Implementations of the NIH Public Access Policy. Online. Available at https://publicaccess.nih.gov/analysis_of_comments_nih_public_access_policy.pdf. Accessed September 15, 2017.

NIH. 2015. National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research. Online. Available at https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf. Accessed March 30, 2018.

NIH. 2017a. Reporting Preprints and Other Interim Research Products. Online. Available at https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-050.html. Accessed April 23, 2018.

NIH. 2017b. NIH Supports International Effort to Create a Central Service for Preprints. Online. Available at https://datascience.nih.gov/preprints/preprints_central_service. Accessed November 9, 2017.

NISO (National Information Standards Organization). 2014. Alternative Metrics Initiative Phase 1 White Paper. Online. Available at https://groups.niso.org/apps/group_public/download.php/13809/Altmetrics_project_phase1_white_paper.pdf. Accessed February 23, 2018.

NLM (National Library of Medicine). 2018a. A platform for biomedical discovery and data-powered health: National Library of Medicine Strategic Plan 2017-2027.

NLM. 2018b. What are genome editing and CRISPR-Cas9? Online. Available at https://ghr.nlm.nih.gov/primer/genomicresearch/genomeediting. Accessed April 16, 2018.

Normile, D. 2018. South Korean Universities Reach Agreement with Elsevier after Long Standoff. Online. Available at http://www.sciencemag.org/news/2018/01/south-korean-universities-reach-agreement-elsevier-after-long-standoff. Accessed March 30, 2018.

Nosek, B., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. Promoting an open research culture. *Science* 348(6242):1422-1425.

Nosek, B. A. 2017. Opening Science. Pp. 89-99 in *Open: The Philosophy and Practices that are Revolutionizing Education and Science*, R. Biswas-Diener and R. Jhangiani, eds. London, United Kingdom: Ubiquity Press.

NRC (National Research Council). 1985. *Sharing Research Data*. Washington, DC: The National Academies Press.

NRC. 1997. *Bits of Power: Issues in Global Access to Scientific Data*. Washington, DC: The National Academies Press.

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

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

NRC. 2012a. *The Future of Scientific Knowledge Discovery in Open Networked Environments: Summary of a Workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18258.

NRC. 2012b. *For Attribution: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: The National Academies Press.

NRC. 2013a. *Copyright in the Digital Era: Building Evidence for Policy*. Washington, DC: The National Academies Press.

NRC. 2013b. *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press.

NRC. 2013c. Public Access to Federally-Supported Research and Development Data and Publications: Two Planning Meetings. Division of Behavioral and Social Sciences and Education. Online. Available at http://sites.nationalacademies.org/dbasse/currentprojects/dbasse_082378. Accessed March 30, 2018.

NRC. 2015. *Preparing the Workforce for Digital Curation*. Washington, DC: National Academies Press.

NSB (National Science Board). 2018. Science & Engineering Indicators 2018. Online. Available at https://www.nsf.gov/statistics/2018/nsb20181. Accessed April 26, 2018.

NSF (National Science Foundation). 2015. NSF's Public Access Plan: Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation. Online. Available at https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf. Accessed March 30, 2018.

NSF. 2016a. Public Access to Results of NSF Funded Research. Online. Available at https://www.nsf.gov/news/special_reports/public_access. Accessed January 10, 2018.

NSF. 2016b. Grant Proposal Guide, Chapter II – Proposal Preparation Guidelines. January 25. Available at https://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg_2.jsp. Accessed June 6, 2018.

NSF. 2018a. Dissemination and Sharing of Research Results. Online. Available at https://www.nsf.gov/bfa/dias/policy/dmp.jsp. Accessed March 21, 2018.

NSF. 2018b. National Science Foundation Research Traineeship (NRT) Program. Online. Available at https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505015. Accessed March 22, 2018.

OASPA (Open Access Scholarly Publishers Association). 2017. The International Community of Open Access Publishers. Online. Available at https://oaspa.org. Accessed December 1, 2017

Odell, J. D., H. L. Coates, and K. L. Palmer. 2016. Rewarding open access scholarship in promotion and tenure: Driving institutional change. *College & Research Libraries News*. Online. Available at https://scholarworks.iupui.edu/bitstream/handle/1805/10343/322.full.pdf?sequence=1&isAllowed=y. Accessed November 24, 2017.

Odell, J., K. Palmer, and E. Dill. 2017. Faculty attitudes toward open access and scholarly communications: disciplinary differences on an urban and health science campus. *Journal of Librarianship and Scholarly Communication* 5(1):eP2169. DOI: http://doi.org/10.7710/2162-3309.2169.

OECD (Organisation for Economic Co-Operation and Development). 2015. Making Open Science a Reality. Online. Available at http://wiki.lib.sun.ac.za/images/0/02/Open-science-oecd.pdf. Accessed March 22, 2018.

Offord, C. 2018. Scientists continue to use outdated methods. *The Scientist*. Online. Available at https://www.the-scientist.com/?articles.view/articleNo/51260/title/Scientists-Continue-to-Use-Outdated-Methods. Accessed February 23, 2018.

O'Neill, L., F. Dexter, and N. Zhang. 2016. The risks to patient privacy from publishing data from clinical anesthesia studies. *Anesthesia & Analgesia* 122(6):2017-2027.

Open Access 2020. 2018. Expression of Interest in the Large-Scale Implementation of Open Access to Scholarly Journals. Online. Available at https://oa2020.org/mission/#eois. Accessed May 25, 2018.

Open Access Directory. 2017. Guides for OA Journal Publishers. Online. Available at http://oad.simmons.edu/oadwiki/Main_Page. Accessed December 1, 2017.

Open Access Max-Planck-Gesellschaft. 2003. Berlin Declaration on Open Access to Knowledge in the Science and Humanities. Online. Available at https://openaccess.mpg.de/Berlin-Declaration. Accessed June 27, 2018.

Open Access Oxford. 2018. Wellcome Trust and Charity Open Access Fund (COAF). Online. Available at http://openaccess.ox.ac.uk/wellcome-and-coaf. Accessed March 16, 2018.

OpenAIRE. 2018. Homepage. Online. Available at https://www.openaire.eu. Accessed March 22, 2018.

OpenAIRE and ICSU World Data System. 2017. Immediate Release: OpenAIRE and ICSU World Data System Announce Cooperation to Advance Open Science. Online. Available at https://www.icsu-wds.org/news/press-releases/openaire-and-icsu-world-data-system-announce-cooperation-to-advance-open-science. Accessed October 12, 2017.

Open Data Handbook. 2018. What Is Open? Online. Available at http://opendatahandbook.org/guide/en/what-is-open-data. Accessed March 29, 2018.

Open Definition. 2018. Open Definition 2.1. Online. Available at http://opendefinition.org/od/2.1/gl. Accessed March 29, 2018.

Open Research Central. 2017. Open Research Central: The Central Portal for Open Research Publishing. Online. Available at https://openresearchcentral.org. Accessed January 8, 2018.

Open Science Training Handbook. 2018. The Open Science Training Handbook. Online. Available at https://legacy.gitbook.com/book/open-science-training-handbook/book/details. Accessed April 23, 2018.

Open Source Initiative. 2018. Homepage. Online. Available at https://opensource.org. Accessed March 29, 2018.

ORCID. 2018a. User Facilities and Publications Working Group. Online. Available at https://orcid.org/content/user-facilities-and-publications-working-group. Accessed May 25, 2018.

ORCID. 2018b. The ORBIT Project. Online. Available at https://orcid.org/organizations/funders/orbit. Accessed May 25, 2018.

ORFG (Open Research Funders Group). 2017. Policy Development Guide. Online. Available at http://www.orfg.org/resources. Accessed December 18, 2017.

ORFG. 2018. Homepage. Online. Available at http://www.orfg.org. Accessed January 5, 2018.

Oregon State University. 2017. OSU to expand sediment core collection to one of largest in the world. Online. http://oregonstate.edu/ua/ncs/archives/2017/mar/osu-expand-sediment-core-collection-one-largest-world. Accessed March 21, 2018.

OSC (Open Science Collaboration). 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716. DOI: 10.1126/science.aac4716.

OSF (Open Science Framework). 2018. Open Science Framework. Online. Available at https://osf.io. Accessed March 22, 2018.

OSTP (Office of Science and Technology Policy). 2013. Increasing Access to the Results of Federally Funded Scientific Research. Memorandum for the Heads of Executive Departments and Agencies from John P. Holdren. Washington, DC: OSTP.

OSTP. 2014. Improving the Management of and Access to Scientific Collections. Online. Available at https://usfsc.nal.usda.gov/sites/usfsc.nal.usda.gov/files/OSTP_MEMO_Scientific_Collxn_FINAL_2014_03(1).pdf. Accessed March 21, 2018.

Palca, J. 1992. The Genome Project: Life after Watson. *Science* 256(5059):956-958.

Panitch, J. M., and S. Michalak. 2015. The Serials Crisis: A White Paper for the UNC-Chapel Hill Scholarly Communications Convocation. Online. Available at http://www.unc.edu/scholcomdig/whitepapers/panitch-michalak.html. Accessed March 23, 2018.

Paskin, N. 2003. Digital Object Identifier (DOI) System. *Encyclopedia of Library and Information Sciences*, Third Edition.

Pasquetto, I. V. et al. 2017. On the reuse of scientific data. *Data Science Journal* 16(8): 1-9. DOI: https://doi.org/10.5334/dsj-2017-008.

Patashnik, O. 1984. BibTeX yesterday, today, and tomorrow. *Proceedings of the TUGboat Annual Meeting* 24(1).

Pearce, N., and A. H. Smith. 2011. Data sharing: Not as simple as it seems. *Environmental Health* 10(107). http://doi.org/10.1186/1476-069X-10-107

Perkel, J. 2016. Democratic databases: Science on GitHub. *Nature* 538(7623):127-128.

Peters, P. 2017. A Radically Open Approach to Developing Infrastructure for Open Science. Online. Available at https://about.hindawi.com/opinion/a-radically-open-approach-to-developing-infrastructure-for-open-science. Accessed March 23, 2018.

Pisanski, K. 2017. Predatory journals recruit fake editor. *Nature* 543(7646).

Piwowar, H., J. Priem, V. Larivière, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J. West, and S. Haustein. 2018. The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *Peer J* 6:e4375. Online. Available at https://doi.org/10.7717/peerj.4375. Accessed June 28, 2018.

Ploeger, L. 2017. Understanding the costs of scholarly publishing – Why we need a public data infrastructure of publishing costs. Online. Available at http://www.openaccessweek.org/profiles/blogs/understanding-the-costs-of-scholarly-publishing-why-we-need-a. Accessed March 30, 2018.

PLOS (Public Library of Science). 2016. Statement on Data Sharing in Public Health Emergencies. Online. Available at http://blogs.plos.org/plos/2016/02/statement-on-data-sharing-in-public-health-emergencies. Accessed June 4, 2018.

PLOS. 2017a. Who We Are. Online. Available at https://www.plos.org/who-we-are. Accessed December 1, 2017.

PLOS. 2017b. Protocols.io Tools for PLOS Authors: Reproducibility and Recognition. Online. Available at http://blogs.plos.org/plos/2017/04/protocols-io-tools-for-reproducibility. Accessed June 6, 2018.

PLOS. 2018. Publication Fees. Online. Available at https://www.plos.org/publication-fees. Accessed May 25, 2018.

PLOS Blogs. 2017. Protocols.io Tools for PLOS Authors: Reproducibility and Recognition. Online. Available at http://blogs.plos.org/plos/2017/04/protocols-io-tools-for-reproducibility. Accessed December 4, 2017.

PLOS One. 2018. Data Availability. Online. Available at http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories. Accessed January 29, 2018.

PNAS (Proceedings of the National Academy of Sciences of the United States of America). 2018. Editorial and Journal Policies. Online. Available at http://www.pnas.org/page/authors/journal-policies. Accessed March 22, 2018.

Pomerantz J., and R. Peek. 2016. Fifty shades of open. *First Monday* 21(5).

Pond, W. 2000. Do security holes demand full disclosure? ZDNet. August 15, 2000.

Posada, A., and G. Chen. 2017. Preliminary Findings: Rent Seeking by Elsevier: Publishers Are Increasingly in Control of Scholarly Infrastructure and Why We Should Care: A Case Study of Elsevier. Online. Available at http://knowledgegap.org/index.php/sub-projects/rent-seeking-and-financialization-of-the-academic-publishing-industry/preliminary-findings. Accessed March 30, 2018.

Poynder, R. 2018. The Open Access Big Deal: Back to the Future. Online. Available at https://poynder.blogspot.ae/2018/03/the-open-access-big-deal-back-to-future.html. Accessed April 18, 2018.

Prinz, F., T. Schlange, and K. Asadullah. 2012. Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10(712). doi:10.1038/nrd3439-c1.

Project Jupyter. 2018. Homepage. Online. Available at http://jupyter.org. Accessed March 22, 2018.

RCUK (Research Councils UK). 2012. RCUK Policy on Open Access and Supporting Guidance. Online. Available at http://www.rcuk.ac.uk/documents/documents/rcukopenaccesspolicy-pdf. Accessed March 30, 2018.

RDA (Research Data Alliance). 2017a. Long tail of research data IG. Online. Available at https://www.rd-alliance.org/groups/long-tail-research-data-ig.html. Accessed December 19, 2017.

RDA. 2017b. Who Is RDA? Online. Available at https://www.rd-alliance.org/node/51727. Accessed October 9, 2017.

Registry of Research Data Repositories. Homepage. Online. Available at https://www.re3data.org. Accessed January 29, 2018.

Reichman, J. H., and P. 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.

Research4Life. 2018. Access to Research in the Developing World. Online. Available at http://www.research4life.org. Accessed January 8, 2018.

Rios, F. 2016. The pathways of research software preservation: An educational and planning resource for service development. *D-Lib Magazine*, July/August. Online. Available at http://www.dlib.org/dlib/july16/rios/07rios.html. Accessed May 31, 2018.

ROARMAP (Registry of Open Access Repository Mandates and Policies). 2018. About the Repository. Online. Available at http://roarmap.eprints.org/information.html. Accessed March 23, 2018.

Rogers, R., A. Roth, A. Smith, and O. Thakkar. 2016. Max-information, differential privacy, and post-selection hypothesis testing. IEEE 57th Annual Symposium on Foundations of Computer Science:487-494. DOI 10.1109/FOCS.2016.59.

Rosenbloom, J. L., D. K. Ginther, T. Juhl, and J. A. Heppert. 2015. The effects of research & development funding on scientific productivity: academic chemistry, 1990-2009. *PLOS One* 10(9):e0138176.

The Royal Society. 2012. Final Report - Science as an Open Enterprise. Online. Available at https://royalsociety.org/topics-policy/projects/science-public-enterprise/Report. Accessed October 12, 2017.

Rücknagel, J., P. Vierkant, R. Ulrich, G. Kloska, E. Schnepf, D. Fichtmüller, E. Reuter, A. Semrau, M. Kindling, H. Pampel, M. Witt, F. Fritze, S. van de Sandt, J. Klump, H-J Goebelbecker, M. Skarupianski, R. Bertelmann, P. Schirmbacher, F. Scholze, C. Kramer, C. Fuchs, S. Spier, and A. Kirchhoff. 2015. Metadata Schema for the Description of Research Data Repositories. Version 3.0. Registry of Research Data Repositories. doi: http://doi.org/10.2312/re3.008.

Samberg, R., R. A. Schneider, A. Taylor, and M. Wolfe. 2018. What's behind OA2020? Accelerating the transition to open access with introspection and repurposing funds. *Scholarly Communication* 79(2). Online. Available at https://crln.acrl.org/index.php/crlnews/article/view/16881/18521. Accessed March 30, 2018.

Sample, I. 2012. Harvard University says it can't afford journal publishers' prices. The Guardian. Online. Available at https://www.theguardian.com/science/2012/apr/24/harvard-university-journal-publishers-prices. Accessed March 23, 2018.

Sansone, S. A., P. Rocca-Serra, and D. Field. 2012. Toward interoperable bioscience data. *Nature Genetics* 44(2):121-126.

Schiermeier, Q. 2017. Hundreds of German universities set to lose access to Elsevier journals: Negotiations to reduce journal prices and promote open access are progressing slowly. Nature 552:17-18.

Schirrwagen, J., P. Manghi, and N. Manola. 2013. Data curation in the OpenAIRE scholarly communication infrastructure. *Information Standards Quarterly* 25(3):13-19.

Schonfeld, R. C. 2017. Elsevier Acquires bepress. The Scholarly Kitchen. August 2. Online. Available at https://scholarlykitchen.sspnet.org/2017/08/02/elsevier-acquires-bepress/. Accessed June 7, 2018.

Shieber, S. M. 2009. Equity for open-access journal publishing. *PLOS Biology* 7(8):e1000165. Online. Available at https://doi.org/10.1371/journal.pbio.1000165. Accessed April 24, 2018.

Science. 2018. Science Journals: Editorial policies. Online. Available at http://www.sciencemag.org/authors/science-journals-editorial-policies. Accessed March 22, 2018.

Science Exchange. 2018. Reproducibility Initiative. Online. Available at https://validation.scienceexchange.com/#/reproducibility-initiative. Accessed May 25, 2018.

Science International. 2015. Open Data in a Big World: An International Accord, Extended Version. In collaboration with the InterAcademy Partnership (IAP), The World Academy of Sciences (TWAS), and the International Social Science Council (ISSC). Online. Available at http://www.science-international.org/sites/default/files/reports/open-data-in-big-data-world_long_en.pdf. Accessed October 11, 2017.

Science-Metrix. 2014. Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996-2013. RTD-B6-PP-2011-2: Study to develop a set of indicators to measure open access. Online. Available at http://science-metrix.com/sites/default/files/science-metrix/publications/d_1.8_sm_eec_dg-rtd_proportion_oa_1996-2013_v11p.pdf. Accessed March 16, 2018.

Science-Metrix. 2018. *Analytical Support for Bibliometrics Indicators: Open Access Availability of Scientific Publications.* Montréal, Canada: Science-Metrix Inc.

Scientific Data. 2018. Recommended Data Repositories. Online. Available at http://www.nature.com/sdata/policies/repositories. Accessed January 29, 2018.

SCOAP3 (Sponsoring Consortium for Open Access Publishing in Particle Physics). 2018. Online. Available at https://scoap3.org. Accessed March 30, 2018.

SESAR (System for Earth Sample Registration). 2018. Online. Available at http://www.geosamples.org. Accessed March 21, 2018.

Shamir, L., B. Berriman, P. Teuben, R. Nemiroff, and A. Allen. 2018. Best Practices for a Future Open Code Policy: Experiences and Vision of the Astrophysics Source Code Library. Online. Available at https://astrocompute.files.wordpress.com/2018/02/shamirlior.pdf. Accessed June 28, 2018.

SHARE (SHared Access Research Ecosystem). 2018. Online. Available at http://www.share-research.org. Accessed March 28, 2018.

SHERPA/Juliet. 2016. Research Funders' Open Access Policies. Online. Available at http://www.sherpa.ac.uk/juliet/index.php. Accessed October 20, 2017.

SHERPA/RoMEO. 2016. Publisher Copyright Policies and Self-Archiving. Online. Available at http://www.sherpa.ac.uk/romeo/index.php. Accessed October 20, 2017.

Shieber, S. 2015. A Model Open Access Policy. https://osc.hul.harvard.edu/assets/files/23.model-policy-annotated_12_2015.pdf. Accessed November 20, 2017.

Shieber, S., and P. Suber, eds. 2015. Good Practices for University Open-Access Policies. Produced by the Harvard Open Access Project and the Berkman Center for Internet and Society at Harvard University. Online. Available at https://cyber.harvard.edu/hoap/sites/hoap/images/Goodpracticesguide-2015.pdf. Accessed November 20, 2017.

Shulenberger, D. 2016. Substituting Article Processing Charges for Subscriptions: The Cure is Worse than the Disease. Association of Research Libraries. Available at http://www.arl.org/storage/documents/substituting-apcs-for-subscriptions-20july2016.pdf. Accessed May 31, 2018.

Simmonds, R., R. Taylor, J. Horrell, B. Fanaroff, H. Sithole, S. J. Van Rensburg, and B. Pretorius. 2016. The African Data Intensive Research Cloud. IST-Africa 2016 Conference Proceedings. Online. Available at http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7530650. Accessed October 13, 2017.

Singal, J. 2016. Inside Psychology's 'Methodological Terrorism' Debate. The Cut. October 12. Available at https://www.thecut.com/2016/10/inside-psychologys-methodological-terrorism-debate.html. Accessed May 31, 2018.

Smith, E., S. Parks, S. Gunashekar, C. A. Lichten, A. Knack, and C. Manville. 2017. *Open Science: The Citizen's Role and Contribution to Research*. Santa Monica, CA: RAND Corporation. Online. Available at https://www.rand.org/pubs/perspectives/PE246.html. Accessed February 23, 2018.

SPARC (Scholarly Publishing and Academic Resources Coalition). 2012. A SPARC Guide for Campus Action. Online. Available at http://sparc.arl.org/news/youve-signed-boycott-now-what. Accessed November 21, 2017.

SPARC. 2016. Author Rights and Author Addendum. Online. Available at http://sparcopen.org/our-work/author-rights. Accessed October 20, 2017.

SPARC, PLOS (Public Library of Science), and OASPA (Open Access Scholarly Publishers Association). 2014. *HowOpenIsIt?* Online. Available at https://www.plos.org/files/HowOpenIsIt_English.pdf. Accessed April 16, 2018.

Stall, S. 2017. Developing Common Standards for Researchers, Repositories, and Publishers to Enable Open and FAIR Data in the Earth and Space Sciences. Presentation to the National Academies of Sciences, Engineering, and Medicine's Committee on Toward an Open Science Enterprise, Public Symposium. September 18, 2017.

State of Open Data. 2018. About. Online. Available at http://www.stateofopendata.od4d.net/about. Accessed March 23, 2018.

Stephan, P. 2012a. *How Economics Shapes Science*. Vol. 1. Cambridge, MA: Harvard University Press.

Stephan, P. 2012b. Research efficiency: Perverse incentives. *Nature* 484:29-31. doi:10.1038/484029a

Stephan, P. E., S. Gurmu, A. J. Sumell, and G. Black. 2007. Who's patenting in the university?

Evidence from the survey of doctorate recipients. *Economics of Innovation and New Technology* 16(2):71-99.

Stodden, V. 2017. Enhancing Reproducibility for Computational Methods. Presentation to the National Academies of Sciences, Engineering, and Medicine Committee on Toward an Open Science Enterprise, First Meeting. July 20, 2017.

Stodden, V., F. Leisch, and R. D. Peng. 2014. *Implementing Reproducible Research*. Boca Raton, FL: Chapman & Hall/CRC. The R series.

Stodden, V., M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M.A. Heroux, J. P. Ioannidis, and M. Taufer. 2016. *Enhancing reproducibility for computational methods*. *Science* 354(6317):1240-1241.

Storer, N. W. 1966. *The Social System of Science*. New York: Holt, Rinehart, and Winston.

Suber, P. 2012. *Open Access*. Cambridge, MA: The MIT Press. Online. Available at https://mitpress.mit.edu/sites/default/files/9780262517638_Open_Access_PDF_Version.pdf. Accessed February 23, 2018.

Suber, P. 2015. Open Access Overview. Online. Available at http://legacy.earlham.edu/~peters/fos/overview.htm. Accessed November 30, 2017

Swan, A. 2012. Policy Guidelines for the Development and Promotion of Open Access. Paris, France: UNESCO (The United Nations Education, Scientific and Cultural Organization).

Swan, A. 2016. The costs and benefits to the research community of Open Access: A briefing paper. Online. Available at http://pasteur4oa.eu/sites/pasteur4oa/files/resource/Costs%20of%20OA%20final_0.pdf. Accessed March 29, 2018.

Sweeney, L. 1996. Replacing personally-identifying information in medical records, the Scrub system. *Proceedings of the AMIA Annual Fall Symposium*:333-337.

Sweeney, L. 1997. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine & Ethics* 25(2-3):98-110.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(05):557-570.

Sweeney. L. 2003. Identifiability of de-identified pharmacy data. Technical report. Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Data Privacy Lab.

Sweeney. L. 2009. Identifiability of de-identified clinical trial data. Technical report. Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Data Privacy Lab.

Szalay, A. S. 2017. From SkyServer to SciServer. *The ANNALS of the American Academy of Political and Social Science* 675(1):202-220.

Sztein, E. 2016. The U.S. Government Role in Preserving Geoscience Sample and Data Collections. American Geophysical Union, Fall General Assembly 2016, abstract #U52A-02. http://adsabs.harvard.edu/abs/2016AGUFM.U52A.02S.

Taichman, D. B., P. Sahni, A. Pinborg, L. Peiperl, C. Laine, A. James, S-T. Hong, A. Haileamlak, L. Gollogly, F. Godlee, F. A. Frizelle, F. Florenzano, J. M. Drazen, H. Bauchner, C. Baethge, and J. Backus. 2017. Data sharing statements for clinical trials: a requirement of the international committee of medical journal editors. *New England Journal of Medicine* 376:2277-2279 DOI: 10.1056/NEJMe1705439.

Taubenberger, J. K., D. Baltimore, P. C. Doherty, H. Markey, D. M. Morens, R. G. Webster, and I. A. Wilson. 2012. Reconstruction of the 1918 influenza virus: Unexpected rewards from the past. *mBio* 3(5):e00201-12.

Tennant, J. P., F. Waldner, D. C. Jacques, P. Masuzzon, L. B. Collister, and C. H. J. Hartgerink. 2016. The academic, economic, and societal impacts of open access: An evidence based review. *F1000 Research* 5:632.

Tenopir, C., E. D. Dalton, L. Christian, M. K. Jones, M. McCabe, M. Smith, and A. Fish. 2017. Imagining a gold open access future: Attitudes, behaviors, and funding scenarios among authors of academic scholarship. *College & Research Libraries* 78(6). Online. Available at https://crl.acrl.org/index.php/crl/article/view/16738. Accessed March 30, 2018.

Think, Check, and Submit. 2017. Choose the right journal for your research. Online. Available at http://thinkchecksubmit.org. Accessed December 1, 2017.

Tippmann, S. 2015. Programming tools: Adventures with R. *Nature* 517(7532):109-110.

UC (University of California), Academic Senate. 2013. Open Access Policy for the Academic Senate of the University of California, Adopted 7/24/2013. Online. Available at https://osc.universityofcalifornia.edu/wp-content/uploads/2013/09/OpenAccess_adopted_072413.pdf. Accessed December 11, 2017.

UC Libraries. 2016. Pay It Forward: Investigating a Sustainable Model of Open Access Article Processing Charges for Large North American Research Institutions. Online. Available at http://icis.ucdavis.edu/wp-content/uploads/2015/07/UC-Pay-It-Forward-Project-Final-Report.pdf. Accessed December 11, 2017.

UC Libraries. 2018. Pathways to Open Access. Online. Available at https://libraries.universityofcalifornia.edu/groups/files/about/docs/UC-Libraries-Pathways%20to%20OA-Report.pdf. Accessed April 13, 2018.

UCOLASC (University Committee on Library and Scholarly Communication). 2018. Declaration of Rights and Principles to Transform Scholarly Communication. Online. Available at https://senate.universityofcalifornia.edu/_files/committees/ucolasc/scholcommprinciples-20180425.pdf. Accessed June 4, 2018.

Universities UK. 2017. Monitoring the Transition to Open Access. Online. Available at http://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2017/monitoring-transition-open-access-2017.pdf. Accessed March 15, 2018.

University of Minnesota Libraries. 2018. Discipline-Based Data Archives: Depositing Your Data. Online. Available at https://www.lib.umn.edu/datamanagement/datacenters. Accessed January 29, 2018.

USFSC (U.S. Federal Scientific Collections). 2018. The Registry of US Federal Scientific Collections. Online. Available at http://usfsc.grscicoll.org. Accessed March 29, 2018.

USGS (U.S. Geological Survey). 2015. Geological Collection Management System – A Master Catalog and Collections Management Plan for U.S. Geological Survey Geologic Samples and Sample Collections. Online. Available at https://pubs.usgs.gov/circ/1410. Accessed March 29, 2018.

USGS. 2018. Data Preservation. Online. Available at https://datapreservation.usgs.gov. Accessed March 29, 2018.

Vale, R. D., and A. A. Hyman. 2016. Point of View: Priority of discovery in the life sciences. *eLife* 5:e16931 doi: 10.7554/eLife.16931.

Vanhecke, T. E. 2008. Citation and research management tool. *Journal of the Medical Library Association* 96(3):275-276.

Van Noorden, R. 2014. Online Collaboration: Scientists and the social network. *Nature* 512:126-129.

Van Noorden, R. 2017. Gates Foundation demands open access: Global-health charity clashes with leading research journals. *Nature* 541:270.

Vardi, M. Y. 2010. Revisiting the publication culture in computing research. *Communications of the ACM* 53(3)5. Online. Available at https://cacm.acm.org/magazines/2010/3/76297-revisiting-the-publication-culture-in-computing-research/fulltext. Accessed May 31, 2018.

Vardigan, M. 2013. The DDI matures: 1997 to the present. *IASSIST Quarterly* 37(1-4):45-50.

Varian, H. R. 1994. Buying, sharing and renting information goods. *The Journal of Industrial Economics* 48(4):473-488.

Varmus, H. 2009. *The Art and Politics of Science*. New York: W.W. Norton.

Verzani, J. 2011. *Getting Started with RStudio*. Newton, MA: O'Reilly Media, Inc.

Vivli. 2018. About Vivli: Overview. Online. Available at http://vivli.org/about/overview. Accessed June 6, 2018.

Voosen, P. 2017. Dueling preprint servers coming for the geosciences. Online. Available at http://www.sciencemag.org/news/2017/09/dueling-preprint-servers-coming-geosciences. Accessed November 10, 2017.

Wagner, B. 2014. Open Access Citation Advantage: An Annotated Bibliography Version 3. Online. Available at https://ubir.buffalo.edu/xmlui/handle/10477/25214. Accessed May 27, 2018.

Wallis, J. C., E. Rolando, and C. L. Borgman. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLOS ONE* 8(7):e67332. https://doi.org/10.1371/journal.pone.0067332.

Wang, X., C. Liu, W. Mao, and Z. Fang. 2015. The open access advantage considering citation, article usage and social media attention. *Scientometrics* 103:555-564. DOI: 10.1007/s11192-015-1547-0

Ware, M., and M. McCabe. 2015. The STM Report: An overview of scientific and scholarly journal publishing. Online. Available at http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf. Accessed February 21, 2018.

Wellcome Trust. 2016. Open Access Policy. Online. Available at https://wellcome.ac.uk/funding/managing-grant/open-access-policy. Accessed October 20, 2017.

Wellcome Trust. 2018. Wellcome Is Going to Review Its Open Access Policy. Online. Available at https://wellcome.ac.uk/news/wellcome-going-review-its-open-access-policy. Accessed May 25, 2018.

West, J. D., T. Bergstrom, and C. T. Bergstorm. 2014. Cost effectiveness of open access publications. *Economic Inquiry* 52(4):1315-1321.

The White House. 2009. Transparency and Open Government: Memorandum for the Heads of Executive Departments and Agencies. Online. Available at https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government. Accessed March 30, 2018.

The White House. 2013. Executive Order-Making Open and Machine Readable the New Default for Government Information. Online. Available at https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government. Accessed March 29, 2018.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018. DOI: 10.1038/sdata.2016.18.

Williams, H. 2013. Intellectual property rights and innovation: evidence from the human genome. *Journal of Political Economy* 121(1):1-27.

Willinsky, J. 2004. Scholarly associations and the economic viability of open access publishing. *Texas Digital Library* 4(2). Online. Available at https://journals.tdl.org/jodi/index.php/jodi/article/view/104/103. Accessed April 16. 2018.

Wilsdon, J., L. Allen, E. Belfiore, P. Campbell, S. Curry, S. Hill, R. Jones, R. Kain, S. Kerridge, M. Thelwall, J. Tinkler, I. Viney, P. Wouters, J. Hill, and B. Johnson. 2015. The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. HEFCE (Higher Education Funding Council for England): United Kingdom. DOI: 10.13140/RG.2.1.4929.1363.

Wilsdon, J., J. Bar-Ilan, R. Frodeman, E. Lex, I. Peters, and P. Wouters. 2017. Next-generation metrics: Responsible metrics and evaluation for open science. Report of the European Commission Expert Group on Altmetrics. Brussels, Belgium: European Commission. Online. Available at https://ec.europa.eu/research/openscience/pdf/report.pdf. Accessed February 23, 2018.

Wilson, B., and M. Fenner. 2012. Open researcher and contributor ID (ORCiD): Solving the name ambiguity problem. *Educase Review* 47(3):54-55.

Winerman, L. 2017. Trends report: Psychologists embrace open science. *American Psychological Association* 48(10):90. Online. Available at http://www.apa.org/monitor/2017/11/trends-open-science.aspx. Accessed February 23, 2018.

Wittenburg, P., and G. Strawn. 2018. Common patterns in revolutionary infrastructures and data. Draft manuscript.

Witze, A. 2016. Iconic Antarctic geology lab gets the boot. *Nature News* 534(7608):448.

Wykstra, S. 2017. Paving the Way to More Reliable Research. Inside Higher Ed. Online. Available at https://www.insidehighered.com/views/2017/07/10/introducing-new-series-reproducibility-scientific-research-essay. Accessed January 10, 2018.

The YODA (Yale University Open Data Access) Project. 2018. Forging a Unified Scientific Community. Online. Available at http://yoda.yale.edu. Accessed June 6, 2018.

Zimmerman, P. 1995. *The Official PGP User's Guide*. Cambridge, MA: MIT Press.

Zenodo. 2018. Homepage. Online. Available at https://zenodo.org. Accessed March 22, 2018.

Zooniverse. Homepage. Online. Available at https://www.zooniverse.org. Accessed February 23, 2018.

Zotero. 2018. Homepage. Online. Available at https://www.zotero.org. Accessed March 22, 2018.

Appendix A

Committee Member Biographies

ALEXA T. MCCRAY (Chair) (NAM) is professor of Medicine at Harvard Medical School and the Department of Medicine, Beth Israel Deaconess Medical Center. She conducts research on knowledge representation and discovery, with a special focus on the significant problems that persist in the curation, dissemination, and exchange of scientific and clinical information in biomedicine and health. McCray is the former director of the Lister Hill National Center for Biomedical Communications, a research division of the National Library of Medicine at the National Institutes of Health (NIH). While at the NIH, she directed the design and development of a number of national information resources, including ClinicalTrials.gov. Before joining the NIH she was on the research staff of IBM's T. J. Watson Research Center. She received a Ph.D. from Georgetown University and for 3 years was on the faculty there. She conducted pre-doctoral research at the Massachusetts Institute of Technology. McCray was elected to the National Academy of Medicine in 2001. She is a fellow of the American Association for the Advancement of Science and a fellow of the American College of Medical Informatics (ACMI). She is past president of ACMI and is a past member of the board of both the American Medical Informatics Association and the International Medical Informatics Association. McCray is past Editor-in-Chief of *Methods of Information in Medicine*, and is a past member of the editorial board of the *Journal of the American Medical Informatics Association*.

FRANCINE BERMAN is the Edward P. Hamilton Distinguished Professor in Computer Science at Rensselaer Polytechnic Institute. Berman was the inaugural recipient of the ACM/IEEE-CS (Association for Computing Machinery/IEEE Computer Society) Ken Kennedy Award for "influential leadership in the design, development, and deployment of national-scale cyberinfrastructure." She is the United States lead of the Research Data Alliance, a community-driven international organization created to accelerate research data sharing worldwide, and has served as director of the San Diego Supercomputer Center and as vice president for Research at Rensselaer Polytechnic Institute. She currently serves as chair of the Anita Borg Institute Board of Trustees, as a member of the National Science Foundation (NSF) Advisory Committee for the Computer and Information Science and Engineering Directorate, as a member of the National Council on the

Humanities, and as a member of the Board of Trustees of the Sloan Foundation. She has previously served on the NSF's Engineering Advisory Committee, the National Institutes of Health's National Institute of General Medical Sciences Advisory Committee, and the U.S. President's Council of Advisors on Science and Technology NITRD Review Working Group. She served as co-chair of the National Academies Board on Research Data and Information, as co-chair of the United States-United Kingdom Blue Ribbon Task Force for Sustainable Digital Preservation and Access, and as chair of the Information, Computing and Communication Section (Section T) of the American Association for the Advancement of Science (AAAS). She is a fellow of the Association of Computing Machinery, the Institute of Electrical and Electronics Engineers, and the AAAS. Berman received her Ph.D. in mathematics from the University of Washington in 1979.

MICHAEL CARROLL is professor of Law and the director of the Program on Information Justice and Intellectual Property (2009- present) at American University Washington College of Law. He teaches and writes about intellectual property law and cyberlaw. Carroll's research focuses on the search for balance in intellectual property law over time in the face of challenges posed by new technologies. He is also recognized as a leading advocate for open access over the Internet to the research that appears in scholarly and scientific journals. In addition, he speaks about and promotes publication of open educational resources and open scientific data. Carroll is a founding member of Creative Commons, Inc., a global organization that provides free, standardized copyright licenses to enable and to encourage legal sharing of creative and other copyrighted works. He also serves on the Board of the Public Library of Science and recently on the National Academies Board on Research Data and Information. He is a member of the Editorial Board of *I/S Journal of Law and Policy* for the Information Society, a non-resident Fellow at the Center for Democracy and Technology, and a member of the Advisory Board of Public Knowledge. Carroll served as a law clerk to Judge Judith W. Rogers, U.S. Court of Appeals for the D.C. Circuit, and Judge Joyce Hens Green, U.S. District Court for the District of Columbia. He practiced law at Wilmer, Cutler & Pickering (now WilmerHale) in Washington, DC. Carroll received his J.D. from the Georgetown University Law Center in 1996.

DONNA GINTHER is professor of Economics and director of the Center for Science Technology & Economic Policy at the Institute for Policy & Social Research at the University of Kansas. She was a research economist and associate policy adviser in the regional group of the Research Department of the Federal Reserve Bank of Atlanta from 2000 to 2002, and taught at Washington University and Southern Methodist University. Her major fields of study are scientific labor markets, gender differences in employment outcomes, wage inequality, scientific entrepreneurship, and children's educational attainments. Ginther has advised the National Academy of Sciences, the National Institutes of Health (NIH), and the Sloan Foundation on the diversity and future of the scientific workforce. She served on the Advisory Committee to the Director of NIH Working Group on the

Biomedical Research Workforce. Ginther was formerly a member of the Board of Trustees of the Southern Economic Association and was formerly on the board of the Committee on the Status of Women in the Economics Profession of the American Economic Association. Ginther received her doctorate in economics from the University of Wisconsin-Madison in 1995.

ROBERT MILLER is the chief executive officer of LYRASIS. He joined LYRASIS in June 2015, bringing more than 25 years of technology industry leadership, global business solutions, and proven executive management experience to the organization. Before joining LYRASIS, Miller served as the general manager of Digital Libraries at Internet Archive (501(c)(3)), a top 200 web company that offers a library of millions of open access books, movies, software, music and more. His tenure boasts many achievements, including successfully building from the ground up the Digital Libraries Division, resulting in more than 2.5 million digitized books globally available with more than 30 million monthly downloads. His work included building key partnerships with over 1,000 state librarians, top libraries, archives, and museums across North America, leading library consortia across Asia, Europe, Africa, and South America. Key relationships were set up with such technology company leaders such as MSN, Adobe, and Canon. Miller is a longtime champion of innovation, entrepreneurship, and global solutions. He has enjoyed a fruitful career as a senior executive in global business as evidenced by his time as founder and co-founder of five start-up companies. As CEO of an Israeli-U.S. information technology company, he led the firm focused on commercializing specialized search technology for health care. Additionally, Miller was co-founder and president of an information technology services company. In this role, he and his team developed and helped patent a consumer behavior referral technology that utilized crowdsourced data with structured metadata, and several consumer product companies that disrupted traditional product models. With a strong commitment to the community, Miller acts as a board member of the Historically Black Colleges and Universities Library Alliance, ArchivesSpace, and CollectionSpace. Miller holds a B.S. in Industrial Engineering from Lehigh University.

PETER SCHIFFER is vice provost for Research and a professor of Applied Physics and Physics at Yale University. As vice provost for Research, he works to support and enhance the research enterprise across all schools and departments in the university. Before joining Yale in 2017, he was the vice chancellor for Research and a professor of Physics at the University of Illinois at Urbana-Champaign, and previously he served in a number of administrative, faculty, and research roles at Pennsylvania State University. Prior to that, Schiffer was on the faculty at the University of Notre Dame, and performed postdoctoral work at AT&T Bell Laboratories. His personal research focuses on artificial spin ice, geometrically frustrated magnets and other magnetic materials. Schiffer has co-authored more than 200 papers, and is the recipient of a Career Award from the National Science Foundation, a Presidential Early Career Award for Scientists

and Engineers from the Army Research Office, an Alfred P. Sloan Research Fellowship recipient, and he received the Faculty Scholar Medal in the Physical Sciences and the Joel and Ruth Spira Award for Teaching Excellence from Penn State. He is also a fellow of the American Physical Society. He has served as the chair of the Topical Group on Magnetism and its Applications and also as the chair of the Division of Materials Physics in the American Physical Society. Schiffer received his B.S. from Yale University in 1988 and his Ph.D. from Stanford University in 1993.

EDWARD SEIDEL is the vice president for Economic Development and Innovation for the University of Illinois System, as well as a founder professor of Physics and professor of Astronomy and Computer Science at the University of Illinois at Urbana-Champaign. He was the director of the National Center for Supercomputing Applications (NCSA) at the University of Illinois from 2014 to 2017. Seidel is a distinguished researcher in high-performance computing and relativity and astrophysics with an outstanding track record as an administrator. His previous leadership roles include serving as the senior vice president for research and innovation at the Skolkovo Institute of Science and Technology in Moscow, directing the Office of Cyberinfrastructure and serving as assistant director for Mathematical and Physical Sciences at the National Science Foundation, and leading the Center for Computation & Technology at Louisiana State University. He also led the numerical relativity group at the Max Planck Institute for Gravitational Physics (Albert Einstein Institute) in Germany. Seidel is a fellow of the American Physical Society and of the American Association for the Advancement of Science, as well as a member of the Institute of Electrical and Electronics Engineers and the Society for Industrial and Applied Mathematics. His research has been recognized with a number of awards. He received his Ph.D. in relativistic astrophysics from Yale University in 1988, earned a master's degree in physics at the University of Pennsylvania, and received a bachelor's degree in mathematics and physics from the College of William and Mary.

ALEXANDER SZALAY is the Bloomberg Distinguished Professor and the Alumni Centennial Professor of Astronomy, and professor of Computer Science at Johns Hopkins University. He is the director of the Institute for Data Intensive Science. Szalay is a cosmologist, working on the statistical measures of the spatial distribution of galaxies and galaxy formation, an interdisciplinary institute to tackle cross-cutting challenges in sciences related to the data deluge. He was heavily involved in the Data Conservancy, a National Science Foundation-funded DataNet project, researching the long-term curation and preservation of scientific data. years he He was part of the U.S. CODATA Council for 4 years and presented at the 2012 Networking and Information Technology Research and Development meeting, commemorating the 20 years of the Internet. He is a fellow of the American Academy of Arts and Sciences. Szalay received his Ph.D. in astrophysics from Eötvös University at Budapest, Hungary in 1975.

LISA TAUXE (NAS) is a distinguished professor of Geophysics in the Geosciences Research Division at Scripps Institution of Oceanography, University of California, San Diego. Her studies concentrate on paleomagnetism, the study of remanent magnetism in geological and archaeological materials. Tauxe has received the George P. Woollard Award of the Geological Society of America (GSA), Outstanding Academic Title in Earth Science from the American Library Association for Essentials of Paleomagnetism, the Antarctic Service Medal, the Benjamin Franklin Medal, and the Arthur L. Day Medal. She has served as president of the Geomagnetism/Paleomagnetism Section and as the General Secretary/Treasurer of the American Geophysical Union (AGU). Tauxe is an elected fellow of the American Association for the Advancement of Science, of the GSA, and of the AGU. She has also been a member of the American Academy of Arts and Sciences since 2016. Tauxe received her Ph.D. in geology from Columbia University and was elected to the National Academy of Sciences in 2015.

HENG XU is associate professor of Information Sciences and Technology at the Pennsylvania State University. She leads the Privacy Assurance Lab, an interdisciplinary research group working on a diverse set of projects related to understanding and assuring information privacy. From 2013 to 2016, Xu served as a program director at the National Science Foundation for Secure and Trustworthy Cyberspace (SaTC) Program in the Directorate for Social, Behavioral, and Economic Sciences. Her research themes emerge from her interests in the fields of information privacy, data analytics, information systems and human-computer interaction. She approaches research issues through a combination of empirical, theoretical, and technical research efforts. Xu was a recipient of an NSF CAREER (Early Faculty Development) Award (2010) and the endowed PNC Technologies Career Development Professorship (2010-2013). Xu received her Ph.D. in Information Systems from the National University of Singapore in 2005.

Appendix B

Glossary

Citation: a well-established measure of research impact; recognition or validation of research by others (Hersh and Plume, 2016).

Delayed open access: articles published in a subscription journal, but are made free to read after an embargo period (Willinsky, 2009; Laakso and Björk, 2013; Piwowar et al., 2017).

Digital object identifier (DOI): a unique alphanumeric string assigned by a registration agency (the International DOI Foundation) to identify content and provide a persistent link to its location on the Internet (American Psychological Association, 2018).

Fully open publication: all articles in the journal freely available to readers immediately upon publication (see Chapter 2).

Gold open access: immediate availability of articles at no cost to the reader beyond that required to access the Internet (see Chapter 2). Articles are published in an open access journal, a journal in which all articles are open directly on the journal website (Archambault et al., 2014; Gargouri et al., 2012; Piwowar et al., 2018).

Green open access: less open approaches to publication in which authors are able to self-archive a version of the article in an open access repository when access to the final published version requires a subscription to the journal (see Chapter 2). Green articles are published in a toll-access journal, but self-archived in an open access archive (Harnad et al., 2008).

Hybrid open access: articles that are published in a subscription journal but are immediately free to read under an open license, in exchange for an article processing charge paid by authors (Piwowar et al., 2018).

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 (NSF, 2007).

Open access: an ambitious goal that aims to ensure the availability and usability of scholarly publications (see Chapter 2). Free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself (Budapest Open Access Initiative, 2002a).

Open access journal: a scientific and scholarly journal that meets high-quality standards by exercising peer review or editorial quality control and use a funding model that does not charge readers or their institutions for access (DOAJ, 2018).

Open code: ensuring the availability and usability of methods, in the case of computational work. The concept of open code is fundamentally linked to open source software and the Open Source Initiative that was founded in 1998 (from Chapter 2).

Open data: data that can be freely used, reused, and redistributed by anyone—subject only, at most, to the requirement to attribute and sharealike (Open Data Handbook, 2018).

Open peer review: peer review where authors' and reviewers' identities are disclosed to one another, as a growing trend in scholarly publishing (Ford, 2015).

Open publication: free and unrestricted access to publications with the only restriction on use being that proper attribution and credit needs to be given to the original creator of the work, as originally advocated by the Budapest Open Access Initiative (see Chapter 2; Budapest Open Access Initiative, 2002b).

Open science: an ambitious goal that aims to ensure the availability and usability of scholarly publications, the data that result from scholarly research, and the methodology, including code or algorithms, that were used to generate those data. Open science typically refers to the entire process of conducting science and harkens back to the original precepts underpinning the conduct and goals of the scientific enterprise (Storer, 1966; Borgman, 2010; Neylon, 2017). (from Chapter 2)

Preprint: a complete written description of a body of scientific work that has yet to be published in a journal (Bourne et al., 2017). Preprint servers can also host other objects such as posters presented at scientific meetings.

Research data: 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)" (GPO, 2012).

Specimen: a portion or quantity of material for use in testing, examination, or study (Merriam-Webster, 2018).

REFERENCES

Archambault, É., D. Amyot, P. Deschamps, A. Nicol, F. Provencher, R. Françoise, L. Rebout, and G. Roberge. 2014. Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels: 1996-2013. Online. Available at http://digitalcommons.unl.edu/scholcom/8. Accessed May 28, 2018.

American Psychological Association. 2018. What is a digital object identifier, or DOI? Online. Available at http://www.apastyle.org/learn/faqs/what-is-doi.aspx. Accessed March 21, 2018.

Borgman, C. 2010. Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, MA: The MIT Press.

Bourne, P. E., J. K. Polka, R. D. Vale, and R. Kiley. 2017. Ten simple rules to consider regarding preprint submission. PLoS Computational Biology 13(5):e1005473. Online. Available at https://doi.org/10.1371/journal.pcbi.1005473. Accessed November 9, 2017.

Budapest Open Access Initiative. 2002a. BOAI15. Online. Available at http://www.budapestopenaccessinitiative.org/boai15-1. Accessed March 21, 2018.

Budapest Open Access Initiative. 2002b. Read the Budapest Open Access Initiative. Online. Available at http://www.budapestopenaccessinitiative.org/read. Accessed March 21, 2018.

DOAJ (Directory of Open Access Journals). 2018. Online. Available at http://doaj.org/about. Accessed March 21, 2018.

Ford, E. 2015. Open peer review at four STEM journals: an observational overview [v2; ref status: indexed, http://f1000r.es/5n1] F1000Research 4:6. doi: 10.12688/f1000research.6005.2.

GPO (U.S. Government Publishing Office). 2012. 2 CFR 215 - Uniform Administrative Requirements for Grants and Agreements with Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations (OMB Circular A-110). Online. Available at https://www.gpo.gov/fdsys/pkg/CFR-2012-title2-vol1/pdf/CFR-2012-title2-vol1-part215.pdf. Accessed February 8, 2018.

Hersh, G., and A. Plume. 2016. Citation metrics and open access: what do we know? Online. Available at https://www.elsevier.com/connect/citation-metrics-and-open-access-what-do-we-know. Accessed March 21, 2018.

Laakso, M., and B. Björk. 2013. Delayed open access: an overlooked high-impact category of openly available scientific literature. Journal of the American Society for Information Science and Technology 64(7):1323-1329 DOI 10.1002/asi.22856.

Merriam-Webster. 2018. Specimen. Online. Available at https://www.merriam-webster.com/dictionary/specimen. Accessed March 21, 2018.

Neylon, C. 2017. Openness in Scholarship: A Return to Core Values? Proceedings of the 21st International Conference on Electronic Publishing. IOS Press Ebooks. Online. Available at http://ebooks.iospress.nl/publication/46638. Accessed March 21, 2018.

NSF (National Science Foundation). 2007. Cyberinfrastructure Vision for 21st Century Discovery. Online. Available at https://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf. Accessed February 12, 2018.

Open Data Handbook. 2018. What is Open? Online. Available at http://opendatahandbook.org/guide/en/what-is-open-data. Accessed March 21, 2018.

Piwowar, H., J. Priem, V. Larivière, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J. West, and S. Haustein. 2018. The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ 6:e4375. DOI 10.7717/peerj.4375.

Storer, N. W. 1966. The Social System of Science. New York, NY: Holt, Rinehart, and Winston.

Willinsky, J. 2009. The access principle: the case for open access to research and scholarship. Cambridge: MIT Press.

Appendix C

Office of Science and Technology Policy 2013 Memorandum: Increasing Access to the Results of Federally Funded Scientific Research[1]

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren, Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the federal government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather data underpins the forecasting industry, and making

[1]The memo is available at https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf. Accessed April 17, 2018.

genome sequences publicly available has spawned many biotechnology innovations. In addition, wider availability of peer-reviewed publications and scientific data in digital formats will create innovative economic markets for services related to curation, preservation, analysis, and visualization. Policies that mobilize these publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.

The Administration also recognizes that publishers provide valuable services, including the coordination of peer review, that are essential for ensuring the high quality and integrity of many scholarly publications. It is critical that these services continue to be made available. It is also important that federal policy not adversely affect opportunities for researchers who are not funded by the federal government to disseminate any analysis or results of their research.

To achieve the Administration's commitment to increase access to federally funded published research and digital scientific data, federal agencies investing in research and development must have clear and coordinated policies for increasing such access.

2. Agency Public Access Plan

The Office of Science and Technology Policy (OSTP) hereby directs each federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the federal government.

This includes any results published in peer-reviewed scholarly publications that are based on research that directly arises from federal funds, as defined in relevant OMB circulars (e.g., A-21 and A-11). It is preferred that agencies work together, where appropriate, to develop these plans.

Each agency plan must be consistent with the objectives set out in this memorandum. These objectives were developed with input from the National Science and Technology Council and public consultation in compliance with the America COMPETES Reauthorization Act of 2010 (P.L. 111-358).

Further, each agency plan for both scientific publications and digital scientific data must contain the following elements:

a) a strategy for leveraging existing archives, where appropriate, and fostering public private partnerships with scientific journals relevant to the agency's research;

b) a strategy for improving the public's ability to locate and access digital data resulting from federally funded scientific research;

c) an approach for optimizing search, archival, and dissemination features that encourages innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research;

d) a plan for notifying awardees and other federally funded scientific researchers of their obligations (e.g., through guidance, conditions of awards, and/or regulatory changes);

e) an agency strategy for measuring and, as necessary, enforcing compliance with its plan;

f) identification of resources within the existing agency budget to implement the plan;

g) a timeline for implementation; and

h) identification of any special circumstances that prevent the agency from meeting any of the objectives set out in this memorandum, in whole or in part.

Each agency shall submit its draft plan to OSTP within six months of publication of this memorandum. OSTP, in coordination with the Office of Management and Budget (OMB), will review the draft agency plans and provide guidance to facilitate the development of final plans that are consistent with the objectives of this memorandum and, where possible, compatible with the plans of other federal agencies subject to this memorandum. In devising its final plan, each agency should use a transparent process for soliciting views from stakeholders, including federally funded researchers, universities, libraries, publishers, users of federally funded research results, and civil society groups, and take such views into account.

3. Objectives for Public Access to Scientific Publications

To the extent feasible and consistent with law; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, the results of unclassified research that are published in peer-reviewed publications directly arising from federal funding should be stored for long-term preservation and publicly accessible to search, retrieve, and analyze in ways that maximize the impact and accountability of the federal research investment.

In developing their public access plans, agencies shall seek to put in place policies that enhance innovation and competitiveness by maximizing the potential to create new business opportunities and are otherwise consistent with the principles articulated in section 1.

Agency plans must also describe, to the extent feasible, procedures the agency will take to help prevent the unauthorized mass redistribution of scholarly publications.

Further, each agency plan shall:

a) Ensure that the public can read, download, and analyze in digital form final peer-reviewed manuscripts or final published documents within a timeframe that is appropriate for each type of research conducted or sponsored by the agency. Specifically, each agency:

   i) shall use a 12-month post-publication embargo period as a guideline for making research papers publicly available; however, an agency may tailor its plan as necessary to address the objectives articulated in this memorandum, as well as the challenges and public interests that are unique to each field and mission combination, and

   ii) shall also provide a mechanism for stakeholders to petition for changing the embargo period for a specific field by presenting evidence demonstrating that the plan would be inconsistent with the objectives articulated in this memorandum;

b) Facilitate easy public search, analysis of, and access to peer-reviewed scholarly publications directly arising from research funded by the federal government;

c) Ensure full public access to publications' metadata without charge upon first publication in a data format that ensures interoperability with current and future search technology. Where possible, the metadata should provide a link to the location where the full text and associated supplemental materials will be made available after the embargo period;

d) Encourage public–private collaboration to:

   i) maximize the potential for interoperability between public and private platforms and creative reuse to enhance value to all stakeholders,

   ii) avoid unnecessary duplication of existing mechanisms,

   iii) maximize the impact of the federal research investment, and

iv) otherwise assist with implementation of the agency plan;

e) Ensure that attribution to authors, journals, and original publishers is maintained; and

f) Ensure that publications and metadata are stored in an archival solution that:

i) provides for long-term preservation and access to the content without charge,

ii) uses standards, widely available and, to the extent possible, nonproprietary archival formats for text and associated content (e.g., images, video, supporting data),

iii) provides access for persons with disabilities consistent with Section 508 of the Rehabilitation Act of 1973,[2] and

iv) enables integration and interoperability with other Federal public access archival solutions and other appropriate archives.

Repositories could be maintained by the Federal agency funding the research, through an arrangement with other Federal agencies, or through other parties working in partnership with the agency including, but not limited to, scholarly and professional associations, publishers and libraries.

4. Objectives for Public Access to Scientific Data in Digital Formats

To the extent feasible and consistent with applicable law and policy;[3] agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze. For purposes of this memorandum, data is defined, consistent with OMB circular A-110, as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports,

[2]Section 508 of The Rehabilitation Act, as amended, available at: https://www.section508.gov/index.cfm?fuseAction=1998Amend.

[3]These policies include, but are not limited to OMB Circular A-130, Management of Federal Information Resources, available at: http://www.whitehouse.gov/omb/circulars_a130_a130trans4.

communications with colleagues, or physical objects, such as laboratory specimens. Each agency's public access plan shall:

a) Maximize access, by the general public and without charge, to digitally formatted scientific data created with Federal funds, while:

   i) protecting confidentiality and personal privacy,

   ii) recognizing proprietary interests, business confidential information, and intellectual property rights and avoiding significant negative impact on intellectual property rights, innovation, and U.S. competitiveness, and

   iii) preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden;

b) Ensure that all extramural researchers receiving federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long-term preservation and access cannot be justified;

c) Allow the inclusion of appropriate costs for data management and access in proposals for Federal funding for scientific research;

d) Ensure appropriate evaluation of the merits of submitted data management plans;

e) Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies;

f) Promote the deposit of data in publicly accessible databases, where appropriate and available;

g) Encourage cooperation with the private sector to improve data access and compatibility, including through the formation of public-private partnerships with foundations and other research funding organizations;

h) Develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan;

i) In coordination with other agencies and the private sector, support training, education, and workforce development related to scientific data management, analysis, storage, preservation, and stewardship; and

j) Provide for the assessment of long-term needs for the preservation of scientific data in fields that the agency supports and outline options for developing and sustaining repositories for scientific data in digital formats, taking into account the efforts of public and private sector entities.

5. Implementation of Public Access Plans

Some Federal agencies already have policies that partially meet the requirements of this memo. Those agencies should adapt those policies, as necessary, to fully meet the requirements. Once finalized, each agency should post its public access plan on its Open Government website.

The agency plan shall not apply to manuscripts submitted for publication prior to the plan's effective date or to digital data generated prior to the plan's effective date. The effective dates can be no sooner than the publication date of the agency's final plan.

OSTP will oversee implementation through regular meetings with agencies. Each agency shall provide updates on implementation to the Directors of OSTP and OMB twice yearly; these updates shall be submitted by January 1 and July 1 of each year for two years after the effective date of the agency's final plan. An agency may amend its public access plan consistent with these objectives, in consultation with OSTP and OMB.

6. General Provisions

Nothing in this memorandum shall be construed to impair or otherwise affect authority granted by law to an executive department, agency, or the head thereof; or functions of the Director of OMB relating to budgetary, administrative, or legislative proposals.

Consistent with the America COMPETES Reauthorization Act of 2010, nothing in this memorandum, or the agency plans developed pursuant to it, shall be construed to authorize or require agencies to undermine any right under the provisions of title 17 or 35, United States Code, or to violate the international obligations of the United States. This memorandum is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity, by any party against the United States; its departments, agencies; or entities, its officers, employees, or agents; or any other person.

Appendix D

Office of Science and Technology Policy 2014 Memorandum: Improving the Management of and Access to Scientific Collections[1]

March 20, 2014

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren, Assistant to the President for Science and Technology, and Director, Office of Science and Technology Policy

SUBJECT: Improving the Management of and Access to Scientific Collections

1. Scientific-Collections Policy Principles

Scientific collections provide an essential base for developing scientific evidence and are an important resource for scientific research, education, and resource management. Scientific collections represent records of our past and investments in our future. They are also tools that can be harnessed to address challenges facing humankind. Federally supported scientific collections are public assets, and their stewardship by federal agencies carries with it trustee responsibilities. Policies and procedures for maintaining, preserving, and developing federal scientific collections while also increasing access to those collections for appropriate use are, therefore, central to their value.

The Administration is committed to ensuring the proper management, preservation, security, and ethical use of federal scientific collections to inform scientific research and maintain the Nation's legacy of exploration and discovery. The federal government has a responsibility to help ensure that scholars and resource managers are able to locate and access federal collections, while also ensuring that

[1]The memo is available at https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_memo_scientific_collections_march_2014.pdf. Accessed April 17, 2018.

collections are appropriately preserved and ethically managed. In some cases, these goals may be served by providing access to digital or other reproductions of elements of the collections.

In response to the policy memorandum I issued on scientific collections in 2010[2] and the requirements of Section 104 of the America COMPETES Reauthorization Act of 2010 (P.L. 111-358),[3] Federal agencies have been working diligently through the Interagency Working Group on Scientific Collections (IWGSC) of the National Science and Technology Council (NSTC) to develop guidelines for the management of scientific collections. Through these efforts, it has become evident that to ensure the faithful stewardship of scientific collections, clear policies for their development, management, and ethical use must be developed by federal agencies.

The policy requirements listed in this memorandum were developed with input from the NSTC IWGSC and in compliance with the America COMPETES Reauthorization Act of 2010 (P.L. 111-358). Each agency's policy on scientific collections shall be consistent with law, agency mission, resource constraints, and U.S. national, homeland, and economic security.

2. Agency Scientific-Collections Policies

Therefore, the Office of Science and Technology Policy (OSTP) hereby directs each federal agency that owns, maintains, or otherwise financially supports permanent scientific collections to develop a draft scientific-collections management and access policy within six months. Agencies should collaborate through the IWGSC while developing these draft policies to reduce redundancy and identify opportunities for common requirements and standards. The end goal will be a systematic improvement of the development, management, accessibility, and preservation of scientific collections owned and/or funded by Federal agencies.

The requirements below are intended to apply to institutional scientific collections owned, maintained, or financially supported by the U.S. government. This policy applies to scientific collections, known in some disciplines as institutional collections, permanent collections, archival collections, museum collections, or voucher collections, which are assets with long-term scientific value. Materials assembled specifically for short-term use, sometimes referred to as "project collections," and not intended for long-term preservation, do not fall under this policy, but such

[2]See http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp-2010-scientific-collections.pdf.

[3]See http://www.gpo.gov/fdsys/pkg/USCODE-2011-title42/html/USCODE-2011-title42-chap79-subchapIIsec6624.htm.

collections should be reviewed periodically and carefully to ensure that they should not be considered institutional collections.

Each agency policy should be consistent with the Executive Order on Making Open and Machine Readable the New Default for Government Information;[4] my earlier memorandum on Increasing Access to the Results of Federally Funded Scientific Research;[5] other relevant Administration initiatives and policies on open data and open government; and the objectives set out in this memorandum. For the purpose of developing agency policies, scientific collections are broadly defined as sets of physical objects, living or inanimate, and their supporting records and documentation, which are used in science and resource management and serve as long-term research assets that are preserved, cataloged, and managed by or supported by federal agencies for research, resource management, education, and other uses. For example, scientific collections can include fossils, tissue specimens, rocks, and many other types of objects essential to scientific research. These policies should apply to scientific collections that are owned, directly managed, or financially supported by federal agencies.

Each agency's policy must include descriptions of the following requirements:

a) the role and importance of collections in advancing the overall mission of the agency, including examples of how specific collections contribute to advancing the agency mission;

b) the legislative and regulatory requirements and authorities related to the agency's scientific collections;

c) the divisions, offices, or other organizational components within the agency that will be responsible for implementing the policy across the agency;

d) the agency officials with responsibility for carrying out policies related to collections, including their specific responsibilities for ensuring compliance;

e) any differences that may exist between department-wide policies and collections-specific policies established by the agency;

f) the methodologies used for the assessment and projection of costs associated with the development, management, and preservation of agency scientific collections;

---

[4]See http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-newdefault-government.-

[5]See http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

g) how the agency budgets for the stewardship of scientific collections, including a description of the overall funding strategy to support scientific collections and ensure online access to information about scientific collections and individual objects;

h) procedures for obtaining or supporting the development of new scientific collections;

i) agency requirements for long-term preservation, maintenance, and accessibility of new and existing collections to maximize public benefit from their use;

j) standards used by the agency for managing collections including ensuring the quality of the collection, its documentation, and for tracking progress on complying with their scientific collections policies and making such progress publicly available;

k) practices for safeguarding individual privacy, confidentiality, intellectual property rights, and national security;

l) a strategy for providing online information about the contents of the agency's scientific collections and, where appropriate, for maximizing access to individual objects in digital form for scientific and educational purposes;

m) how the agency will provide access to the public or other members of the research community, including how collections and information about collections will be disseminated equitably;

n) the process for de-accessioning, transferring, and disposing of scientific collections, including documentation procedures and procedures for moving collections acquired for individual projects to institutional collections; and

o) resources within the existing agency budget to implement the policy.

3. Management Objectives for Scientific Collections:

For the stewardship of scientific collections it supports, each agency shall, where applicable:

a) Develop and clearly describe procedures for making scientific collections more accessible to educators and researchers, including non-federal scientists, to maximize public benefit.

b) Work with the Smithsonian Institution to ensure that information on the contents of and how to access the agency's scientific collections is available

on the Internet in a central federal clearinghouse and to maintain participation in the federal clearinghouse once it is established.

c) Use machine-readable and open formats, data standards, and common-core and extensible metadata for all new information creation and collection to facilitate search and discoverability and provide clear public guidance for accessing collections materials, consistent with the Executive Order on Making Open and Machine Readable the New Default for Government Information.

d) When available and where not limited by law, make freely and easily accessible to the public all digital files in the highest available fidelity and resolution, including, but not limited to, photographs, videos, and digital 3-D models, and associated records and documentation, describing or characterizing objects in government-managed scientific collections.

e) Associate digital files describing or characterizing scientific collections with the agency's collections catalog and the central Federal clearinghouse referenced in Section 3(b) of this memorandum. By default, this information should be in machine-readable and open format.

f) Limit access to collections and information about collections for the purpose of protecting national interests including honoring copyright, international or tribal agreement, confidentiality, privacy and other laws, and regulations, or addressing security concerns. For example, locality information could be withheld or its release limited for the purpose of protecting endangered or otherwise protected species or research sites or complying with the Native American Graves Protection and Repatriation Act, the Archaeological Resources Protection Act, the Paleontological Resources Preservation Act, the National Parks Omnibus Management Act, or the Health Insurance Portability and Accountability Act.

g) In the event that access to objects within or information about a collection must be restricted as described by 3(f), restrictions on access shall be limited to the minimal subset of specific objects and records possible, with all other collection content made public. Where possible, redaction of specific metadata fields should be favored over limiting access to the entire object or subset of objects.

h) Clearly describe how the agency will apply its scientific collections policy as a term and condition, as appropriate, of providing funding for the acquisition and stewardship of scientific collections that are being managed by a third party or that the agency does not own, but supports or for which it has oversight responsibilities.

i) Consistent with each agency's mission and authority, establish standards for de-accession and disposal of scientific collections. When transferring collections, give preference to transferring to other Federal agencies or non-federal institutions that will continue to make the collections and information about the collections accessible for research and education. These standards should include:

   i. review of the research, resource management, and education values of a collection
   ii. consultation with researchers who have used the collection, parties interested in the collection's value for research, resource management, and educational purposes, and other subject matter experts, as needed; and
   iii. procedures to transfer scientific collections that agencies no longer need to researchers at institutions or other entities qualified to manage the collections.

Agencies should work together to share and coordinate policies, where appropriate, through the IWGSC.

OSTP will review draft agency policies to ensure they are consistent with the objectives of this memorandum and other requirements, including the America COMPETES Reauthorization Act of 2010. During the drafting and review process, OSTP will seek opportunities to harmonize policies among federal agencies and will provide feedback to facilitate the development of final agency policies that are consistent with the objectives of this memorandum.

Some federal agencies already have policies that partially or fully meet the requirements of this memorandum. Those agencies should adapt or maintain those policies, as necessary, to fully meet these requirements. Once finalized, each agency should post its scientific collections policy on its Open Government website.

4. General Provisions

Nothing in this memorandum shall be construed to impair or otherwise affect authority granted by law to an executive department, agency, or the head thereof; or functions of the Director of OMB relating to budgetary, administrative, or legislative proposals.

This memorandum is not intended to, and does not create any right or benefit, substantive or procedural, enforceable at law or in equity, by any party against the United States, its departments, agencies, or entities, its officers, employees, or agents, or any other person.

Appendix E

Committee Meeting Agendas

Open Session

FIRST COMMITTEE MEETING

July 20, 2017
Keck Center of the National Academies of Sciences, Engineering, and Medicine
500 Fifth Street NW, Room 101, Washington, DC

1:00 PM **Welcome and Introductions**
Alexa McCray (NAM), Harvard Medical School, Committee Chair

1:15 PM **Sponsor's Briefing on the Statement of Task**
Michael Stebbins, Laura and John Arnold Foundation

1:45 PM **Effective Open Access Policies and Practices**
Heather Joseph, SPARC

2:15 PM **Enhancing Reproducibility for Computational Methods**
Victoria Stodden, University of Illinois at Urbana-Champaign

2:45 PM **A Manifesto for Reproducible Science**
Brian Nosek, Center for Open Science

3:15 PM BREAK

3:30 PM **Remarks from the National Academies of Sciences, Engineering, and Medicine**
Marcia McNutt (NAS), President, National Academy of Sciences

3:45 PM **National Perspective on Toward an Open Science Enterprise**
James Kurose, National Science Foundation

4:15 PM **Q&A and Discussion**

4:40 PM **Open Microphone Session: Brief Comments from Interested Parties**

5:00 PM **Open Session Adjourns**

SECOND COMMITTEE MEETING
September 18, 2017
National Academy of Sciences
2101 Constitution Avenue NW, Room 125, Washington, DC

PUBLIC SYMPOSIUM: TOWARD AN OPEN SCIENCE ENTERPRISE—FOCUS ON STAKEHOLDERS

9:00 AM **Welcome and Introductions**
Alexa McCray (NAM), Harvard Medical School, Committee Chair

9:15 AM **Session I: Perspectives of Publishers and Journal Editors**

Advocating Open Science at PLOS
Joerg Heber, Public Library of Science (PLOS)

Towards an Open Science Enterprise: A Community Organization Perspective
Michael Forster, Institute of Electrical and Electronics Engineers

Approaching Open Science across the Researcher Workflow
Holly Falk-Krzesinski, Elsevier

bioRxiv: A Preprint Server for the Life Sciences
John Inglis, bioRxiv and Cold Spring Harbor Laboratory Press

Consuming Identifiers: A Path to Open Science
Howard Ratner, Clearinghouse for the Open Research of the United States (CHORUS)

10:50 AM BREAK

11:10 AM **Session II: Perspectives of Private Sector and Foundations**

Facilitating the Discovery of Scientific Data
Natasha Noy, Google

Providing Support and Solutions for Open Science to Achieve Impact
Jennifer Hansen, Bill & Melinda Gates Foundation

The Economics of Research Data
Daniel Goroff, Alfred P. Sloan Foundation

12:10 PM **Open Microphone Session: Brief Comments from Interested Parties**

12:30 PM LUNCH

1:15 PM **Session III: Perspectives of Federal Agencies and Academic Libraries**

Opening Science and Scholarship
Michael Huerta, National Library of Medicine, National Institutes of Health

Implementation and Learning Healthcare System Research in the Era of Open Science
Amy Kilbourne, U.S. Department of Veterans Affairs

Challenges to and Progress towards Open Science from both Federal and Community-Driven Perspectives
Lindsay Powers, U.S. Geological Survey

The Open Science "Stack": Infrastructure, Scientific Objects, and Policy
Tyler Walters, Virginia Polytechnic Institute and State University

Getting to Open: Challenges, Drivers, and Opportunities
Ivy Anderson, California Digital Library, University of California

2:55 PM BREAK

3:15 PM **Session IV: Perspectives of Research Community and Scientific Societies**

Fostering Open Science in Meteorological Research, Operations, and Education
Eugene Takle, Iowa State University

Paving the Rocky Road toward Open and FAIR in the Field Sciences

Kerstin Lehnert, Columbia University

Developing Common Standards for Researchers, Repositories, and Publishers to Enable Open and FAIR Data in the Earth and Space Sciences
Shelley Stall, American Geophysical Union

Enabling Open Science without Impeding Open Science
Kenton McHenry, National Data Service

4:35 PM **Open Microphone Session: Brief Comments from Interested Parties**

5:00 PM **Public Symposium Adjourns**