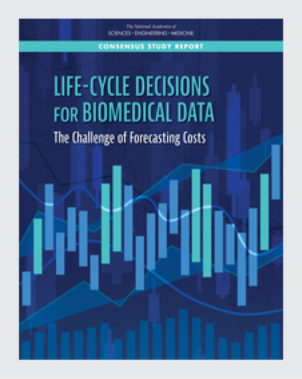

The National Academies of SCIENCES · ENGINEERING · MEDICINE
CONSENSUS STUDY REPORT
LIFE-CYCLE DECISIONS FOR BIOMEDICAL DATA
The Challenge of Forecasting Costs

Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs (2020)

DETAILS

184 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-67003-6 | DOI 10.17226/25639

CONTRIBUTORS

Committee on Forecasting Costs for Preserving and Promoting Access to Biomedical Data; Board on Mathematical Sciences and Analytics; Committee on Applied and Theoretical Statistics; Computer Science and Telecommunications Board; Board on Life Sciences; Board on Research Data and Information; Division on Engineering and Physical Sciences; Division on Earth and Life Studies; Policy and Global Affairs; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2020. *Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25639.

LIFE-CYCLE DECISIONS FOR BIOMEDICAL DATA

The Challenge of Forecasting Costs

Committee on Forecasting Costs for Preserving and Promoting Access to Biomedical Data

Board on Mathematical Sciences and Analytics
Committee on Applied and Theoretical Statistics
Computer Science and Telecommunications Board
Division on Engineering and Physical Sciences

Board on Life Sciences
Division on Earth and Life Studies

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 Contract No. HHSN26300002 with the National Institutes of Health. 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: 978-0-309-67003-6
International Standard Book Number-10: 0-309-67003-9
Digital Object Identifier: https://doi.org/10.17226/25639

Additional copies of this publication are available 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. 2020. *Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Cost*s. Washington, DC: The National Academies Press. https://doi.org/10.17226/25639.

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 FORECASTING COSTS FOR PRESERVING AND PROMOTING ACCESS TO BIOMEDICAL DATA

DAVID S.C. CHU, Institute for Defense Analyses, *Chair*
ILKAY ALTINTAS, University of California, San Diego
G. SAYEED CHOUDHURY, Johns Hopkins University
MARGARET C. LEVENSTEIN, University of Michigan
CLIFFORD A. LYNCH, Coalition for Networked Information
DAVID MAIER, Portland State University
CHARLES F. MANSKI, NAS,[1] Northwestern University
MARYANN MARTONE, University of California, San Diego
ALEXA T. McCRAY, NAM,[2] Harvard Medical School
MICHELLE N. MEYER, Geisinger
WILLIAM W. STEAD, NAM, Vanderbilt University Medical Center
LARS VILHUBER, Cornell University

Staff

SAMMANTHA L. MAGSINO, Senior Program Officer, Board on Earth Sciences and Resources, *Study Director*
SELAM ARAIA, Senior Program Assistant, Board on Mathematical Sciences and Analytics
LINDA CASOLA, Associate Program Officer, Board on Mathematical Sciences and Analytics (until December 2019)
CHRISTOPHER FU, Research Associate, Board on Mathematical Sciences and Analytics (until August 2019)
ADRIANNA HARGROVE, Financial Manager
TYLER KLOEFKORN, Program Officer, Board on Mathematical Sciences and Analytics
MICHELLE SCHWALBE, Director, Board on Mathematical Sciences and Analytics
LINDA WALKER, Program Coordinator, Board on Physics and Astronomy
BEN WATZAK, Research Assistant, Board on Mathematical Sciences and Analytics

[1] Member, National Academy of Sciences.
[2] Member, National Academy of Medicine.

BOARD ON MATHEMATICAL SCIENCES AND ANALYTICS

MARK L. GREEN, University of California, Los Angeles, *Chair*
HÉLÈNE BARCELO, Mathematical Sciences Research Institute
JOHN R. BIRGE, NAE,[1] University of Chicago
RUSSEL E. CAFLISCH, NAS,[2] New York University
W. PETER CHERRY, NAE, Independent Consultant
DAVID S.C. CHU, Institute for Defense Analyses
RONALD R. COIFMAN, NAS, Yale University
JAMES (JIM) CURRY, University of Colorado, Boulder
SHAWNDRA HILL, Microsoft Research
LYDIA KAVRAKI, NAM,[3] Rice University
TAMARA KOLDA, NAE, Sandia National Laboratories
RACHEL KUSKE, Georgia Institute of Technology
JOSEPH A. LANGSAM, University of Maryland, College Park
DAVID MAIER, Portland State University
LOIS CURFMAN McINNES, Argonne National Laboratory
JILL PIPHER, Brown University
ELIZABETH A. THOMPSON, NAS, University of Washington
CLAIRE TOMLIN, NAE, University of California, Berkeley
LANCE WALLER, Emory University
KAREN E. WILLCOX, University of Texas, Austin

Staff

MICHELLE SCHWALBE, Director
SELAM ARAIA, Senior Program Assistant
LINDA CASOLA, Associate Program Officer (until December 2019)
CHRISTOPHER FU, Research Associate (until August 2019)
ADRIANNA HARGROVE, Finance Business Partner
TYLER KLOEFKORN, Program Officer
BEN WATZAK, Research Assistant

[1] Member, National Academy of Engineering.
[2] Member, National Academy of Sciences.
[3] Member, National Academy of Medicine.

COMMITTEE ON APPLIED AND THEORETICAL STATISTICS

ALFRED O. HERO III, University of Michigan, *Chair*
ALICIA CARRIQUIRY, NAM,[1] Iowa State University
RONG CHEN, Rutgers University, The State University of New Jersey
MICHAEL J. DANIELS, University of Florida
KATHERINE BENNETT ENSOR, Rice University
AMY H. HERRING, Duke University
TIM HESTERBERG, Google, Inc.
NICHOLAS J. HORTON, Amherst College
DAVID MADIGAN, Columbia University
XIAO-LI MENG, Harvard University
JOSÉ M.F. MOURA, NAE,[2] Carnegie Mellon University
RAQUEL PRADO, University of California, Santa Cruz
NANCY M. REID, NAS,[3] University of Toronto
CYNTHIA RUDIN, Duke University
AARTI SINGH, Carnegie Mellon University
ALYSON G. WILSON, North Carolina State University

Staff

TYLER KLOEFKORN, Director
SELAM ARAIA, Senior Program Assistant
LINDA CASOLA, Associate Program Officer (until December 2019)
CHRISTOPHER FU, Research Associate (until August 2019)
ADRIANNA HARGROVE, Financial Manager
BEN WATZAK, Research Assistant

[1] Member, National Academy of Medicine.
[2] Member, National Academy of Engineering.
[3] Member, National Academy of Sciences.

COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD

FARNAM JAHANIAN, Carnegie Mellon University, *Chair*
STEVEN M. BELLOVIN, NAE,[1] Columbia University
DAVID CULLER, NAE, University of California, Berkeley
EDWARD FRANK, NAE, Cloud Parity, Inc.
LAURA HAAS, NAE, University of Massachusetts, Amherst
MARK HOROWITZ, NAE, Stanford University
ERIC HORVITZ, NAE, Microsoft Corporation
BETH MYNATT, Georgia Institute of Technology
CRAIG PARTRIDGE, Colorado State University
DANIELA RUS, NAE, Massachusetts Institute of Technology
FRED B. SCHNEIDER, NAE, Cornell University
MARGO SELTZER, NAE, University of British Columbia
MOSHE VARDI, NAS[2]/NAE, Rice University

Staff

JON EISENBERG, Senior Board Director
SHENAE BRADLEY, Administrative Assistant
RENEE HAWKINS, Financial and Administrative Manager
LYNETTE I. MILLETT, Associate Director
KATIRIA ORTIZ, Associate Program Officer
BRENDAN ROACH, Program Officer

[1] Member, National Academy of Engineering.
[2] Member, National Academy of Sciences.

BOARD ON LIFE SCIENCES

JAMES P. COLLINS, Arizona State University, *Chair*
A. ALONSO AGUIRRE, George Mason University
VALERIE H. BONHAM, Ropes and Gray LLP
DOMINIQUE BROSSARD, University of Wisconsin, Madison
NANCY D. CONNELL, Johns Hopkins Center for Health Security
SEAN M. DECATUR, Kenyon College
JOSEPH R. ECKER, NAS,[1] Howard Hughes Medical Institute
SCOTT V. EDWARDS, NAS, Harvard University
GERALD L. EPSTEIN, National Defense University
ROBERT J. FULL, University of California, Berkeley
MARY E. MAXON, Lawrence Berkeley National Laboratory
ROBERT NEWMAN, Independent Consultant
STEPHEN J. O'BRIEN, NAS, Nova Southeastern University
LUCILA OHNO-MACHADO, NAM,[2] University of California, San Diego
CLAIRE POMEROY, NAM, The Albert and Mary Lasker Foundation
MARY E. POWER, NAS, University of California, Berkeley
SUSAN R. SINGER, Rollins College
LANA SKIRBOLL, Sanofi
DAVID R. WALT, NAE[3]/NAM, Brigham and Women's Hospital
PHYLLIS M. WISE, NAM, University of Colorado

Staff

FRAN SHARPLES, Director
KATHERINE BOWMAN, Senior Program Officer
JESSICA DE MOUY, Senior Program Assistant
ANDREA HODGSON, Program Officer
JO HUSBANDS, Scholar/Senior Project Director
STEVEN MOSS, Associate Program Officer
KEEGAN SAWYER, Senior Program Officer
AUDREY THEVENON, Program Officer
KOSSANA YOUNG, Senior Program Assistant

[1] Member, National Academy of Sciences.
[2] Member, National Academy of Medicine.
[3] Member, National Academy of Engineering.

BOARD ON RESEARCH DATA AND INFORMATION

ALEXA T. McCRAY, NAM,[1] Harvard Medical School, *Chair*
AMY BRAND, Massachusetts Institute of Technology Press
STUART FELDMAN, Schmidt Futures
SALMAN HABIB, Argonne National Laboratory
JAMES HENDLER, Rensselaer Polytechnic Institute
MARY LEE KENNEDY, Association of Research Libraries
BAREND MONS, Leiden University Medical Centre
SARAH NUSSER, Iowa State University
MICHAEL STEBBINS, Science Advisors, LLC

Staff

TOM ARRISON, Director
EMI KAMEYAMA, Associate Program Officer
GEORGE STRAWN, Scholar
ESTER SZTEIN, Deputy Director

[1] Member, National Academy of Medicine.

Acknowledgment of Reviewers

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:

Helen Berman, Rutgers, The State University of New Jersey,
David Browdy, Fred Hutchinson Cancer Research Center,
Mercè Crosas, Harvard University,
Brandi Davis-Dusenbery, Seven Bridges,
Mark Ellisman, National Center for Microscopy and Imaging Research,
Adam Ferguson, University of California, San Francisco,
Aaron Friedman, Amazon Web Services,
Peter Jones, NAS,[1] Van Andel Institute,
Brian Lavoie, OCLC Research,
Margo Seltzer, NAE,[2] University of British Columbia,
Douglas Sicker, Carnegie Mellon University,
Sharon Terry, Genetic Alliance, and
Carol Thompson, Allen Institute.

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 Susan J. Curry, NAM,[3] University of Iowa. She was responsible for

[1] Member, National Academy of Sciences.

[2] Member, National Academy of Engineering.

[3] Member, National Academy of Medicine.

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.

Contents

APPENDIXES

Summary

Biomedical research results in the collection and storage of increasingly large and complex data sets. Preserving those data so that they are discoverable, accessible, and interpretable accelerates scientific discovery and improves health outcomes but requires that researchers, data curators, and data archivists consider the long-term disposition of data and the costs of preserving, archiving, and promoting access to them. All involved in data management throughout the data life cycle need to consider how data-related choices affect the costs of future preservation, management, and use. All need to be informed about the costs of retaining versus replacing data, the value of retained data, the costs of data curation and storage, and potential costs borne by future data users. These are integral to data preservation, archiving, and access promotion. Attention to and quantitative estimates of such costs will facilitate better allocation of resources and planning by those charged with guiding and investing in the production of scientific knowledge such as researchers, research-performing institutions, and funders.

The mission of the National Library of Medicine (NLM) within the National Institutes of Health (NIH) is to acquire, organize, and disseminate health-related information. At the request of NLM, the National Academies of Sciences, Engineering, and Medicine convened a committee to examine and assess approaches and considerations for forecasting costs for preserving, archiving, and promoting access to biomedical research data. This report provides a comprehensive conceptual framework for cost-effective decision making that encourages data accessibility and reuse for researchers, data managers, data archivists, data scientists, and institutions that support platforms that enable biomedical research data preservation, discoverability, and use. The framework can be adapted by anyone responsible for managing data at any point in the data life cycle, but the analysis conducted during its application by researchers, data, data repository hosts, and funding institutions will vary greatly. Its purpose is to make the forecaster think of all the elements that could affect life-cycle costs so that costs can be understood and total costs be more accurately calculated. Other than the forecasting framework itself, the report does not include recommendations. Rather it describes the kind of environment conducive to forecasting the cost of sustainable data management, and provides strategies that could be applied by different members of the biomedical research community for creating those environments.

THE STUDY CHARGE

As part of its charge to develop and demonstrate a framework for forecasting long-term costs for preserving, archiving, and accessing various types of biomedical research data, the study committee evaluated economic factors to be considered when examining the life-cycle costs for data acquisition, curation, and preservation; the cost consequences for various practices related to accessioning and deaccessioning data sets; economic factors if data are designated as high value; anticipated technological disruptors and future developments in data science in a 5- to 10-year horizon; and critical factors for successful adoption of data-forecasting approaches by research and program management staff. Per the statement of task provided to the committee by NLM, the framework was applied to two case studies in different biomedical contexts relevant to NLM data resources. The committee also organized a 2-day workshop to gather input on tools and practices that NLM could use to help researchers and funders better integrate risk-management practices and considerations into data preservation, archiving, and accessing decisions; methods to encourage NIH-funded researchers to consider, update, and track lifetime data costs; and burdens on the academic researchers and industry staff to implement these tools. A summary of workshop proceedings was published in a separate document (NASEM, 2020).

THE COST-FORECASTING FRAMEWORK

The framework for forecasting costs presented in this report first describes the different data environments in which data may be placed (herein referred to as "data states"; Box 2.1) and the various activities associated with those data states (Tables 2.1-2.3), and steps in the framework process are identified (Table 4.1). The cost drivers that may be important for each of those activities (Table 4.2) and questions that lead critical decision points related to those cost drivers are described in Chapter 4 through a series of questions to be answered by the forecaster. The committee tabulated those questions in a template that can be modified and used to inform a cost analysis (Appendix E). The forecasting framework does not offer computational models for quantifying costs because those applying the framework will have diverse interests in the framework's application and diverse resources. Instead, it provides a comprehensive conceptual framework which the forecaster can use to identify what costs need to be quantified.

States of the Data Life Cycle

The data life cycle begins when data are collected during primary research and continues through data analysis, preservation and curation, reuse, storage, and potentially to deaccession. The data life cycle is not necessarily linear—data may be reused and repurposed, combined with other data, and analyzed in a variety of ways and for different purposes throughout their existence. How actively data are used during the data life cycle may change: they may be used often when initially collected and then only periodically after placed in a repository. At some point, they may become dormant and be placed in an archive for long-term preservation. They may be rediscovered at any time and again be actively used. Ideally, the data states in which the data are placed throughout their existence allow for different types of activities. Data may be moved from one data state to another as needs arise, the data may transition in a nonlinear manner, or some data may not ever transition into all the data states.

Digital data may transition among three states in the data life cycle:

- *State 1:* The primary research and data management environment where data are captured and analyzed. It is possible that no one managing or using a given State 1 data environment is focused on standardizing, documenting, sharing, or preserving data and algorithms.
- *State 2:* An active repository and platform where data may be acquired, curated, aggregated, accessed, and analyzed. Such a repository is an active information system that usually provides services to a wide range of users. Where data are complex, confidential, or very large, it may be a platform for controlling access. Support may be provided for analyzing and processing data.
- *State 3:* A long-term preservation platform in which content is preserved across changes in governance, assessment of data value, and technology. The platform may include an extract of data from a single data set, multiple data sets, or an information system in a system-agnostic format. In this state, data are neither directly analyzable nor easily accessible.

These data states were conceptualized by the committee to communicate the characteristics of different environments, with different purposes, and having different data storage and preservation costs. The data states can be represented by Figure S.1, which also illustrates the major activities associated with each state. Tables 2.1-2.3 in the main body of the report provide more detail about the activities shown in the figure, as well as various subactivities that may occur and the personnel required to conduct them.

The Cost-Forecasting Process

Every data resource and management situation has unique characteristics and considerations, but there are commonalities in the cost-forecasting process. This report does not present an instrument for cost forecasting but rather a framework to help the cost forecaster build the instrument that is suitable for the particular application. The framework identifies many of the commonalities and should be considered a foundation for a detailed analysis that can be tailored for specific circumstances. Regardless of the application, the forecaster is encouraged to think about the entire life cycle of the data rather than of just the life of the data resource being developed or managed. It is more cost efficient in the long term if decisions are made in light of their impacts on future costs of management and data access. Table S.1 summarizes the steps necessary to understand the cost drivers that are important for

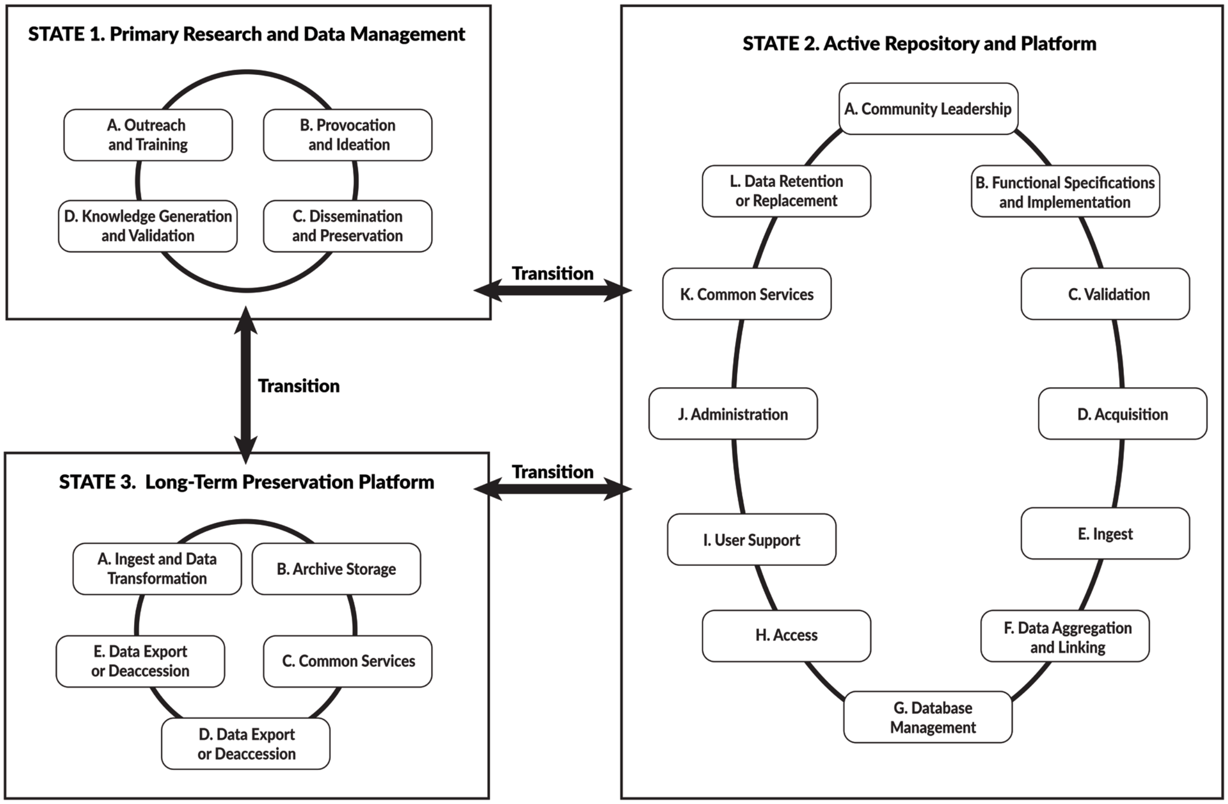

STATE 1. Primary Research and Data Management
A. Outreach and Training
B. Provocation and Ideation
D. Knowledge Generation and Validation
C. Dissemination and Preservation
STATE 2. Active Repository and Platform
A. Community Leadership
L. Data Retention or Replacement
B. Functional Specifications and Implementation
K. Common Services
C. Validation
J. Administration
D. Acquisition
I. User Support
E. Ingest
H. Access
F. Data Aggregation and Linking
G. Database Management
Transition
Transition
Transition
STATE 3. Long-Term Preservation Platform
A. Ingest and Data Transformation
B. Archive Storage
E. Data Export or Deaccession
C. Common Services
D. Data Export or Deaccession

FIGURE S.1 Conceptual diagram showing the three data states, the principal activities associated with each state, and how data may transition between states. Note that the transition arrows between states are bidirectional, indicating that data already existing in repositories can transition back into a primary research environment when new data are incorporated, data are aggregated with other data, or data are used in new ways.

TABLE S.1 Steps for Forecasting Costs of a Biomedical Information Resource

1. Determine the type of data resource environment, its data state(s), and how data might transition between those states during the data life cycle. The data states are defined in Box 2.1: State 1: primary research environment State 2: active repository State 3: long-term preservation and archive	• Decide the goals and objectives for the data resource. • Consider how the resource is likely to be used now and in the future. • Identify available guidance that defines the type of resource to be created or managed (e.g., requests for application, community standards, or institutional requirements). • Compare the above with the activities defined for each of the data states (see Figure S.1) and decide which data state(s) best align(s).
2. Identify data characteristics (Chapter 4), data contributors, and users.	• Fill in the cost-driver template (Appendix E) o Complete category A of the template to help to identify the size, complexity, metadata requirements, depth versus breadth, processing levels and fidelity, and replaceability of the data. o Complete category E of the template to help to identify the life-cycle issues. o Complete category F of the template to help to identify data contributors and users.
3. Identify the current and potential value of the data and how the data value might be maintained or increased with time.	• Consult with the institution hosting the data resource, the project funders, and the broader research community to develop appropriate metrics for assessing the value of the data. • Identify decisions that affect data value in the shorter and longer terms (see Chapter 3 for different methodologies). • Consider how data generation methodologies affect short- and long-term data value in terms of data contributors and users and the data life cycle.
4. Identify the personnel and infrastructure likely necessary in the short and long terms.	• Identify the major activities and subactivities associated with the information resource, including activities related to potential transitions between data states (Tables 2.1, 2.2, and 2.3). • Identify short- and long-term staffing requirements for the current state and transition between states. • Identify the infrastructure requirements and available resources.
5. Identify the major cost drivers associated with each activity based on the steps above, including how decisions might affect future data use and its cost.	• Identify the major cost drivers and associated uncertainties for each of the activities identified above by completing the cost-driver template (Appendix E). • Identify likely relative costs (e.g., using Table 4.2). • Consult with institutional experts (e.g., at the institution hosting the resource, library resources) and determine available personnel and infrastructure resources. • Work with experts at the host institution to quantify short-term costs and to bound uncertainties in longer-term forecasts.
6. Estimate the costs for relevant cost components based on the characteristics of the data and information resource.	• Identify which cost drivers are important for each cost component of the information resource (e.g., labor, information technology infrastructure and services, media, licenses and subscriptions, facilities and utilities, outside services, travel, and institutional overhead; Box 3.2). • Estimate costs for the current funding period. • Estimate costs and cost uncertainties for future funding periods, including costs to transition data to other states.

a given information resource. The framework will assist the forecaster in identifying the characteristics of data and the biomedical information resource, the near-term and future data management needs, and the activities and decisions that are likely to be important drivers of near-term and future costs. The steps outlined in the table will not necessarily be performed in the order presented. Forecasting activities may occur concurrently, and they may need to be revisited as new information unfolds during analysis. The cost forecast can be quantified when decisions pertaining to them are made. Chapter 4 defines the primary cost drivers, listed below:

- Content (e.g., data size, complexity, and diversity; metadata requirements, depth versus breadth, processing level and fidelity; and replaceability of the data);
- Capabilities (e.g., user annotation, persistent identifiers, citation, search, data linking and merging, use tracking, and data analysis and visualization);
- Control (e.g., content, quality, access, and platform);
- External context (e.g., resource replication, external information dependencies, and distinctiveness);
- Data life cycle (e.g., anticipated growth, updates and versions, useful lifetime, and offline and deep storage);
- Contributors and users (e.g., contributor base, user base and usage scenarios, training and support requirements, and outreach);
- Availability (e.g., tolerance for outages, currency, response time, and local versus remote access);
- Confidentiality, ownership, and security (e.g., data privacy issues and licensing);
- Maintenance and operations (e.g., periodic integrity checking, data-transfer capacity, risk management, and system-reporting requirements); and
- Standards and regulatory compliance and other governance concerns.

Table 4.2 (in Chapter 4) illustrates which of these drivers are likely to be important for the major activities in the three data states. A series of questions related to each cost driver is provided in Chapter 4—and compiled in a template in Appendix E—to assist the forecaster in his analysis. The questions may need to be modified for a specific application of the cost framework: not all the guiding questions may be relevant to a given application, and not all relevant questions may be included. Through work with experts from within the institution that will host the data resource (e.g., the researcher's university), relative costs may be estimated for activities for the data life cycle, and shorter-term costs may be quantified. Working through the guiding questions will also help the forecaster identify uncertainties in forecasted costs.

In most cases in which data are shared, the costs of long-term data preservation are not borne by a single individual or institution. Responsibility may be transferred, for example, from a researcher to a data platform host or between platform hosts. Understanding where costs will be accrued, who pays those costs, and who has managerial responsibility for them will inform decision makers for all data states. The cost-forecasting framework guides the forecaster through identification of those who hold responsibility for those factors.

CREATING AN ENVIRONMENT CONDUCIVE TO COST FORECASTING OF SUSTAINABLE DATA MANAGEMENT

Approaches to building and managing data repositories differ across institutions and among researchers, but regardless of where biomedical information resources are hosted, costs associated with personnel are likely to dominate total life-cycle costs. Storage, computing, and networking services also contribute to total cost. The ability of individual researchers to forecast and manage those costs depends on how well they understand service-provider costs and prices—whether those services are rendered by the research institution or by commercial providers. The lack of visibility regarding the true costs of data storage and access in individual laboratories, institutions, and community resources often hampers reliable cost forecasting.

Costs associated with long-term preservation, archiving, and access to biomedical research data will likely rise as data sets increase in size and complexity. Being able to forecast those costs is critical to the success of sustainable data preservation and access. Successful cost forecasting and sustainable data management require that those making decisions about data have the necessary information and incentives to recognize the full costs of data management borne by all parties throughout the data life cycle. This is true whether decision makers are researchers, data scientists, research institution officials, data resource managers, or program managers at funding agencies or federal agencies that host and manage data on behalf of the broader research community.

To foster the scientific environment necessary to conduct better long-term cost forecasts now and into the future, a series of strategies, actions, and advances is presented below. The reader will need to determine how best to apply the strategies based on her role in the scientific endeavor and on the data environments under consideration.

Strategies

Efficient long-term data management and effective cost forecasting are more likely if data resource managers, cost forecasters, and institutions that support them apply the following strategies:

- **Create data environments that foster discoverability and interpretability through long-term planning and investment throughout the data life cycle.** Data sharing is not equivalent to data reuse, and developing processes that allow efficient data preservation, archiving, and access to facilitate data reuse could benefit scientific discovery.
- **Incorporate data management activities throughout the data life cycle to strengthen data curation and preservation.** Up-front costs may be increased, but data value may also increase, and the overall cost of research may be reduced.
- **Incorporate the expertise and resources needed to create and curate metadata throughout the data life cycle, and in the transition between data states into the cost forecast.** Data discoverability and reusability depend on adherence to community-accepted data and metadata standards.
- **Weigh the benefits, risks (e.g., data loss), and costs (both up-front and anticipated) of data storage and computation options before selecting among options.** A service may look attractive from an immediate-financing perspective, but service-provider strategies deserve vetting and verification, including examination of exit or transition strategies and costs. Long-term costs need to be informed by a provider's risk-management strategies.

Actions

Individuals within specific biomedical sectors may collaborate to increase the efficiency of data management efforts, but there is little guidance available from funding agencies and the institutions that support biomedical data resources on practices for long-term management and cost forecasting for the biomedical research community. The following actions, especially if taken by funding agencies and institutions that support data resources, could expand the capacity of data producers and managers to make sound management decisions and cost forecasts:

- **Explicitly recognize the value of State 2 data resources (i.e., active repositories) to the enhanced curation, discoverability, and use of data.** This recognition is absent among the funding entities, researchers, and institutions supporting research, most of which apply the more traditional data management approach of transitioning data directly from the primary research environment (i.e., State 1) to long-term archiving (i.e., State 3).
- **Structure cost forecasts for State 2 resources around communities and research programs rather than individual research efforts.** Because State 2 resources serve communities of researchers, it may not be appropriate to allocate the costs of managing data in a State 2 resource back to the individual data contributor.
- **Support standardization efforts, including developing tools and methodologies to estimate the cost of standards development, encouraging the use of those tools and standards as part of the funding programs where appropriate, and explicitly supporting metadata preparation.** Support could take the form of funding and the provision of tools. Issuing clarifying language about the use of federal funds for data preservation beyond the performance period of the project that collected them would also help assist in the development and promotion of the use of community standards and metadata preparation.

- **Identify incentives, tools, and training for adopting good data management practices, including cost-forecasting practices, which facilitate sustainable long-term data preservation, curation, and access.** Such activities would benefit the entire biomedical research community, including the institutions and funding entities that support research. To support these endeavors, funding entities need to better understand research-community needs, help the community to define desired outcomes, support training, develop realistic and actionable metrics for success, and provide near-term incentives for success.
- **Understand the charges associated with storage and computation in a data resource, regardless of who "pays the bill," when making decisions about data and workflows.** Institutions supporting research might develop mechanisms to inform researchers of the actual costs paid for the services rendered to them and encourage them to limit those costs.

Advances for Practice

Data are of little use without services to support them. Institutions that support primary research (State 1), or the development and management of State 2 (active) or State 3 (long-term preservation) repositories, face challenges understanding and providing the resources necessary to build, maintain, or otherwise acquire access to the systems necessary for a sustainable data-preservation platform. There is often confusion regarding who bears ultimate ownership (i.e., intellectual rights) and responsibility for data and data policies at the institutional level. Successful long-term data stewardship cannot be an ad hoc endeavor but rather needs to be planned in advance. Methodologies to forecast life-cycle costs for preserving, archiving, and accessing biomedical data are immature, and few tools and resources are available for those to quantify long-term costs with confidence. Making people aware of and accountable for their costs—and helping them understand that their actions generate costs for someone—might help researchers reduce resource consumption with more efficient workflows, experiment design, and data tracking.

The following activities, likely to be enabled at an agency or research-institution level, could advance practices and drive future improvements in the ability to forecast costs:

- **Recognize explicitly that scientific data constitute an asset and that data stewardship requires support.** Biomedical research data and data resources are vital to the delivery of good science and, ultimately, to the public good. The universities and institutions that support or enable research and host data resources, in turn, benefit from the recognition of that support.
- **Systematically collect data on costs associated with the biomedical research data enterprise to allow the translation of the framework outlined in this report into resources and methodologies that would benefit individual researchers and repository institutions.** A clear locus of responsibility for compiling this information systematically is necessary.
- **Develop easier mechanisms for creating and maintaining data management plans (DMPs), automatically incorporating data and metadata into resources, and improving citations for data to work together with other research products.** By providing these mechanisms, funders and research institutions could help improve efficiency, return value for stakeholders, and increase the likelihood that stakeholders will make sound data-related decisions.

POTENTIAL DISRUPTORS

Disruptors are considered anything that may cause radical changes to the ways research is conducted and data are collected, used, archived, or preserved. Disruptors may be positive, negative, or mixed, and may either raise or lower the cost of data management and preservation. There is no way to fully anticipate potential disruptor impact, but remaining aware of it and building flexibility into data can help to mitigate the effects. There are numerous issues that could lead to disruptions, including issues such as the evolving open data practices and the application of "findable, accessible, interoperable, and reusable" (FAIR) data principles for research data; developments in cybersecurity (both regulatory and legal requirements that may interact with privacy and human-subjects regulations, and in terms of changing threat environments); major changes in funding levels and flows; more general changes in the vendor landscape (e.g., bankruptcies, mergers, and acquisitions); technology production and supply chains; and environmental or geopolitical developments. Many of these are discussed throughout this report. Chapter 7 of this report includes discussion of the following potential disruptors:

- *Biomedical data volume and variety:* Sudden orders-of-magnitude increases in data collection in domains such as imaging and multiscale high-performance computing simulations have moved biomedical research into the realm of "big data." This has been observed, for example, given recent advances in genomics research. Biomedical research will experience growth that tends to add dimensions to the data space or to extend a dimension by an order of magnitude.
- *Advances in machine learning and artificial intelligence (AI):* The increased use of machine learning and AI techniques has accompanied the increases in data volumes. Automatic annotation of data and metadata generation using AI that allow regular updates to volumes of data increases the need for active storage approaches and requires programmatic access to data; this may also have implications for metadata requirements. Certain compliance and regulatory processes may be automated, but AI may also give rise to new challenges as it upends assumptions about data identifiability and data security.
- *Changes in storage technologies and practices:* While costs per byte stored per year continue to drop, although more slowly than in the past, the aggregate size of the data being stored and managed is growing quickly. Some sites face physical facilities constraints on the amount of storage they can support locally; the capital investment costs of purchasing and upgrading storage are also challenging. Cloud providers offer greater flexibility in terms of expansion, and costs of storage and compute services continue to become more attractive. On the other hand, if there is a need to change providers, moving large amounts of data is technically challenging and could involve a variety of costs of which the information resource manager must be aware. In other words, vendor lock-in could be a risk.
- *Future computing technologies:* The scale and speed of the adoption of emerging technologies in the next 5 to 10 years is uncertain. For example, the ability to move computation to data rather than data to computation can change practices and costs. New cloud cost models may be the determinant factor for the overall cost of data. Emerging edge computing models, reliance on non–von Neumann architectures, and specialized hardware such as machine learning accelerators may also reshape how data are stored and reused, and the associated costs of storage and usage.
- *Workforce development challenges:* It is difficult and expensive to attract and sustain the needed size and quality of workforce when industry offers higher wages than those offered in the public sector and academia.
- *Legal and policy disruptors:* Changes in legislation and policy related to issues such as data sharing; data identifiability; permissible data collection, storage, and sharing; and human-subjects research may require changes in the way data are stored, shared, and accessed.

ENSURING LONG-TERM STEWARDSHIP

It is not common practice to think beyond a current funding period when developing a data management budget, and the current system for funding research is not conducive to data life-cycle cost forecasting.

At present, cost forecasting is typically short term and is often conducted only at the onset of an endeavor when many issues are uncertain (e.g., data quantities, quality, and format). Planning horizons are dictated by funding streams (e.g., federal budget allocations, grant levels) and thus extend only for the life of the project, excluding post-project data-preservation issues. Many researchers think about the disposition of their data after their primary research is complete and strive to make those data public. DMPs (see Appendix B) today are typically static documents prepared as a mandatory—but not necessarily influential—part of the funding process. Placing more emphasis on quantified cost forecasts during the award process may be one way to incentivize early planning and communication, even if cost forecasts are uncertain. However, placing greater emphasis on cost forecasting at that time does not mean that the forecasts will become precise estimates; they could be considered accurate reflections of uncertainties. Cost forecasts and DMPs need to evolve as research progresses and as associated data and the resources and technologies available to manage those data evolve. Monitored evolution of a DMP (e.g., at mid-term evaluations or at the end of the award period) might inform eligibility for future funding.

The cost of long-term data stewardship is better considered systematically by the funding institution rather than research by staff. Researchers working in a State 1 environment typically are not responsible for costs or data management beyond the grant performance period. Managing data in States 2 and 3 generally becomes an institutional responsibility, but planning at the institutional level is typically over 1- to 2-year time horizons rather than over the many years required to realize the promise of current and future repositories. A forecaster will focus on costs associated with the resource under development or being managed but needs to be aware of how early planning decisions can affect long-term costs of data curation and use in future states (e.g., by increasing the efficiency of future curation and use or by making future curation prohibitively expensive).

Treating cost estimation as an important agency priority and investing in training, recognizing success, critiquing failures, and encouraging assembly of cost-related data are increasingly important. However, evidence is needed to understand costs. The federal government has an important role in preserving data resulting from scholarly activity. The systematic collection of cost data related to the biomedical-research-data enterprise by an organization that owns that responsibility could provide evidence necessary to translate the cost-forecasting framework presented in this report into a set of tools that can be used by the biomedical-research and data-preservation community. This development could encourage institutions to focus on costs, facilitate future cost forecasting, and help refine cost-forecasting models. The ultimate beneficiaries of such efforts, of course, will be the scientific enterprise and our nation's citizens, whose well-being science seeks to advance.

REFERENCE

NASEM (National Academies of Sciences, Engineering, and Medicine). 2020. *Planning for Long-Term Use of Biomedical Data: Proceedings of a Workshop*. Washington, D.C.: The National Academies Press. https://doi.org/10.17226/25707.

1

Introduction

Biomedical researchers generate, collect, and store more research data than ever. They do so in an environment that requires increasing levels of data openness and sharing as well as ever greater attention to the privacy of those from whom the data derive. Preserving those data in discoverable and accessible ways is increasingly important; however, it is no longer reasonable to collect data with the expectation that they will be stored "forever" at no cost. Data often may be worth preserving only if they have been integrated and aggregated with other related data. Resource constraints make it necessary for researchers and data archivists to consider the long-term disposition of data, data sets, and data streams, as well as to consider the long-term costs for preserving, archiving, and promoting access to those data. In many, if not most, cases, responsibility for data management shifts to different individuals and institutions as the data are collected, analyzed, curated, archived, and potentially reused for other purposes.

Given resource constraints, researchers will need to be able to estimate the quantity of data that they will collect over the course of a project as well as determine the likelihood and feasibility of data reuse when the original project has ended. Archivists might need to consider what archival principles might be imposed on the data as well as the ramifications of those principles. Archivists might also need to determine whether greater value should be placed on particular data types and to consider how risk management might inform data-preservation and archiving strategies. To support the scientific endeavor, all involved in managing the data throughout the data life cycle need to consider how choices regarding their data affect the costs of future preservation, management, and use, regardless of who bears those costs. Their decisions will require information about costs for curation and storage, revenue prospects associated with data generation or future data use, and estimated value of retained data (or, alternatively, the cost of replacing the data). While information alone will not result in the internalization of costs incurred elsewhere in the data life cycle, attention to and quantification of such costs will facilitate better allocation of resources and planning by those charged with guiding and investing in the production of scientific knowledge. Some of those costs may or may not be expressed in monetary units. Market valuation may place more value on, for example, clinical versus preclinical data in some contexts, or on one type of biomedical data versus another. Such valuations may not align with the priorities and mission of the organization bearing the cost. An understanding of how to make and use such valuations will inform data-preservation decisions.

The mission of the National Institutes of Health's (NIH's) National Library of Medicine (NLM) is to acquire, organize, and disseminate health-related information. NLM's strategic plan includes accelerating discovery and advancing health through data-driven research, reaching more people through enhanced dissemination and engagement (NLM, 2018). As the largest biomedical library in the world, NLM understands that meeting the needs of the biomedical research community requires a community-wide understanding that recent technological advances in data-collection technologies and data science require commensurate increases in resources for data curation, preservation, and discoverability (NLM, 2017). Efforts have been undertaken to understand the issues related to improving data discovery and access to NIH-funded data (e.g., Read et al., 2015), and NLM wants to strengthen a research community's capability to value future data reuse and to estimate the cost of making reuse possible, which requires tools, methods, and practices.

At the request of NLM, the National Academies of Sciences, Engineering, and Medicine (National Academies) have prepared this report, which examines and assesses approaches and considerations for forecasting costs for preserving, archiving, and promoting access to biomedical research data. The report first identifies the different data environments (called "states" in this report) in which data must be managed and the activities that take place in those environments. The report then identifies cost drivers and the states and activities in which costs are incurred. Critical decisions that might be made during each of the data states that could influence long-term costs of curation, preservation, and access to data are identified. The purpose of this report is to provide a general framework for cost-effective decision making that encourages data accessibility and reuse for researchers, data managers, data archivists, data scientists, and institutions that support platforms for biomedical research data preservation and use. The framework is not itself a cost model; rather, it provides the set of activities that must be considered in constructing a cost model. The wide range of possible activities associated with any particular biomedical data set precludes constructing a single generic model that all readers of this report could employ. Costs will be a function of the data set characteristics, the activities that will be undertaken with the data set, and the duration of those activities. Moreover, readers of this report will be addressing cost forecasts at different times in the life cycle and with different interests.

However, the cost-forecasting framework presented in this report can be adapted by anyone responsible for managing data at any point in the life cycle, and use of the framework itself is the primary recommendation in this report. Specific recommendations regarding the application of the framework are not provided because researchers, data, repository hosts, and funding institutions and their requirements vary greatly. The report does describe the kind of environment conducive to forecasting the cost of sustainable data management, but making recommendations to specific institutions about how to create those environments is beyond the scope of this report.

THE CHARGE TO THE NATIONAL ACADEMIES AND THE STUDY COMMITTEE

NLM asked the National Academies to develop and demonstrate a cost-forecasting framework and estimate potential future benefits to research. The charge to the National Academies is provided in Box 1.1. To meet its charge, the National Academies convened an ad hoc committee of experts in areas such as biomedical sciences; biomedical informatics; cybersecurity; data science; data storage, archiving, and architectures; database systems; decision making under uncertainty; economics; ethics; health care; information theory; mathematical modeling of reliability; cost forecasting; and statistics. The members were nominated by their peers and selected by the National Academies in consideration of the balance of individual expertise and to avoid any unnecessary conflict of interest and bias. Brief biographies of the committee members are included in Appendix G.

COMMITTEE INFORMATION GATHERING AND APPROACH TO ITS TASK

Given the complexity of the task, the committee employed a series of public meetings, a workshop, multiple site visits, and one-on-one communication to hear from the numerous stakeholders in the biomedical research community. These stakeholders included biomedical-science researchers at academic or nonprofit institutions; data scientists and institutional administrators within academic, private, and public sectors; data archivists; software engineers; data platform managers; and many others—all individuals who make decisions about data throughout the entire data life cycle. The committee consulted the published literature and communicated with individuals and institutions with responsibility for managing data in different formats to understand any cost-forecasting models that might be applied and those factors that influence decision making. The committee members also drew from their collective expertise and experiences, and many of the report conclusions are based on their own observations.

BOX 1.1
Statement of Task

A National Academies of Sciences, Engineering, and Medicine–appointed ad hoc committee will develop and demonstrate a framework for forecasting long-term costs for preserving, archiving, and accessing various types of biomedical data and estimating potential future benefits to research. In so doing, the committee will examine and evaluate the following considerations:

- Economic factors to be considered when examining the life-cycle cost for data sets (e.g., data acquisition, preservation, and dissemination);
- Cost consequences for various practices in accessioning and deaccessioning data sets;
- Economic factors to be considered in designating data sets as high value;
- Assumptions built into the data collection and/or modeling processes;
- Anticipated technological disruptors and future developments in data science in a 5- to 10-year horizon; and
- Critical factors for successful adoption of data forecasting approaches by research and program management staff.

The committee will provide two case studies illustrating application of the framework to different biomedical contexts relevant to the National Library of Medicine's data resources. Relevant life-cycle costs will be delineated, as well as the assumptions underlying the models. To the extent practicable, the committee will identify strategies to communicate results and gain acceptance of the applicability of these models.

As part of its information gathering, the committee will plan and organize a 2-day workshop to gather input on the following topics:

- Tools and practices that NLM could use to help researchers and funders better integrate risk management practices and considerations into data preservation, archiving, and accessing decisions;
- Methods to encourage NIH-funded researchers to consider, update, and track lifetime data costs (e.g., through data management plans and project renewals, or other interactions with NIH); and
- Burdens on the academic researchers and industry staff to implement these tools, methods, and practices.

The committee held five meetings in Washington, D.C.; three of these meetings included open sessions in which speakers and guests were invited to respond to questions relevant to the committee's statement of task (Box 1.1). Agendas for the committee's open session meetings, the workshop, and site visits appear in Appendix A.

The committee's first meeting included presentations and a panel discussion with leadership and staff at NIH and NLM. Those individuals were asked to describe NIH's institutional priorities and primary objectives for data management, the largest cost issues encountered to meet those objectives, and the financial mechanisms employed or being considered to meet data-related goals. The second meeting included panelists with expertise in research technology, methodologies, and workflows across the data life cycle, and in research-data collaboration from the private sector. They were asked to describe their respective methods for anticipating long-term uses and costs of data management and emerging issues that could affect data management in the future. The committee's third meeting included discussions with the director of digital preservation from the U.S. National Archives to learn about that agency's approaches to develop budgets for data preservation and with the deputy project leader of the World Wide Computing Grid of the European Organization for Nuclear Research (CERN) to learn about lifetime data management at CERN. A principal project manager from a commercial cloud vendor discussed the changes in technologies, data volumes and types, and data uses in the near and distant future. The committee queried all of these individuals about the positive and negative developments anticipated in the next 5 to 10 years that are likely to affect the cost of data preservation, archiving, and access. The committee's workshop included 15 speakers and more than 50 invited guests, and focused on elements as described in Box 1.1. Workshop participants had the opportunity to discuss (1) tools and practices that NLM could use to help researchers and funders better integrate risk-management practices and considerations into data preservation, archiving, and accessing decisions; (2) methods to encourage NIH-funded researchers to consider, update, and track lifetime data costs; and (3) burdens on the academic researchers and industry staff to implement these tools, methods, and practices. The workshop was summarized in a set of proceedings (NASEM, 2020) that was released separately from the present report.

To engage in open discussion with a number of additional stakeholders and service providers, the committee made a total of 10 site visits to a variety of institutions. To maximize efficiency, the committee first identified the types of institutions with which it should engage, identified specific institutions of those types and their locations, and then identified regions where a number of such institutions could be visited within a few days. Ultimately, the committee visited institutions in San Diego, California; Washington, D.C.; Boston, Massachusetts; and Seattle, Washington. These included visits to the National Center for Microscopy and Imaging Research; the University of California, San Diego/San Diego Supercomputer Center Advanced CyberInfrastructure Development Lab; NIH; Dana-Farber Cancer Institute; Harvard Medical School; The Broad Institute; Amazon Web Services; Institute for Systems Biology; Allen Institute; and the Fred Hutchinson Cancer Research Center. The committee also met with NIH staff across institutes that run various data repositories. Ultimately, the committee members met with more than 100 individuals with different perspectives and expertise who engaged with research data at different states within the data life cycle. Visits to any number of other institutions could have informed the committee's deliberations, but the mix of private- and public-sector data resource managers, researchers, data scientists, and service providers provided the committee the base of information it needed. Information gaps were then filled via literature searches and personal communications between committee members and external experts.

The committee found that few think about data preservation beyond the state in which the data currently exist. Only a few of the individuals with whom the committee interacted consider how their data-related decisions affect the long-term costs associated with data preservation, curation, and access. Data regarding the costs of data resource management do not seem to be collected in any systematic way to inform future efforts. Planning for the longer term seems nascent: there are few tools or community standards available to assist with long-term planning, and such planning is at best short term in nature. Planning horizons are dictated by funding streams (e.g., federal budget allocations, grant levels) and thus extend only for the life of the project, excluding post-project data-preservation issues. Given these findings and the apparent lack of suitable examples on which to base the committee's cost-forecasting model, the committee concluded that it would be necessary to develop an original framework on which to base a forecasting model per its charge. Once the framework was developed, the committee demonstrated its use by applying it to two use cases, per the statement of task.

FEDERAL CONTEXT

Some federal agencies, such as the National Aeronautics and Space Administration and the National Oceanic and Atmospheric Administration, are already committed to data preservation and understand that proper data preservation is a complex endeavor requiring dedicated resources over the long term. Data preservation is integrated into their cultures, and the cost of long-term preservation is included when costing a project. Other agencies have traditionally attached less importance to data preservation, and increased efforts on that front may require major cultural shifts within those agencies, a challenge that should not be underestimated. Regardless, planning horizons for agencies are often short given annual budget appropriations and that there may be legal prohibitions to planning beyond the appropriation period. These factors affect the way agencies fund external research.

Researchers who receive federal funds may be required or encouraged to do some level of planning for the disposition of their data at the end of their research projects, but research funding does not generally extend beyond the performance of the research and often does not cover data preservation. As a result, data-preservation activities are often minimal and may be conducted only as an afterthought to the research. Research grants provided by U.S. funding agencies, for example, are generally awarded only for 2- to 3-year performance periods, although some awards may be longer. Foreign funding agencies often fund research over longer performance periods with explicit requirements related to data curation. For example, Canadian[1,2] and European[3,4] funding agencies offer different types of awards for periods ranging up to 7 years. German agencies have provided research funding for up to 12 years. Information-technology infrastructure and data management are explicitly incorporated into these grants.

BIOMEDICAL DATA LANDSCAPE

To provide some context for its work, the committee presents an overview of some of the current infrastructure and stakeholders that comprise the biomedical landscape, focusing mostly on infrastructure funded by NIH. The biomedical data landscape is diverse, distributed, and dynamic, characterized by an array of data repositories, databases, and platforms, which host data and make them available for reuse. These repositories are the database infrastructures where long-term stewardship, preservation, and access to research data are made possible. For the purposes of this report, the committee uses the terms "data repository" and "data archive" to refer to data infrastructure that host primary research data rather than to refer to knowledge bases that extract and aggregate analyzed data from the scientific literature. A similar distinction is adopted in the *NIH Strategic Plan for Data Science* (NIH, 2018). Examples of primary research data repositories include the Protein Data Bank,[5] the National Institute of Mental Health Data Archive, and the National Archive of Computerized Data on Aging;[6] examples of knowledge bases include UniProt[7] and the Monarch Initiative.[8] The distinction between a database and a knowledge base is not always clean—many digital repositories serve dual purposes—but NIH defines the primary function of a data repository to "ingest, archive, preserve, manage, distribute, and make accessible the data related to a particular system or systems."[9] The primary function of a knowledge base, according to NIH, is to "extract, accumulate, organize, annotate, and link growing bodies of information related to core datasets."[10] A third type of digital artifact that does not fit neatly into these categories is the many digital spatial atlases that cover structures such as the nervous system (e.g., Brainspan Atlas of the Developing Human,[11] Cell Atlas of Mouse Brain-Spinal Cord Connectome[12]), urogenital system, heart, and other organs (e.g., Genito-Urinary Molecular Anatomy Project).[13] An important component of biomedical data and knowledge resources is that they are usually not simple platforms for hosting data; they also comprise many software tools and services that make the data and knowledge usable and useful. Consideration of challenges in hosting physical samples (i.e., biospecimen or biosample repositories) is beyond the scope of this report.

The biomedical repository landscape spans accessible data repositories hosted by government agencies, national laboratories, research consortia, institutions and hospitals, patient advocacy organizations, researchers, journals, and commercial entities, including consortia of study sponsors. The exact number of these repositories is difficult to estimate. re3data,[14] a registry of research data repositories, returns a list of 483 for the search term "basic biological and medical research," of which 38 are in the United States. The California Digital Library, a digital research library founded by the University of California, lists more than 600 biomedical source databases

[1] Information regarding the Canadian Social Sciences and Humanities Research Council is found at https://www.sshrc-crsh.gc.ca/funding-financement/programs-programmes/partnership_development_grants-subventions_partenariat_developpement-eng.aspx, accessed May 12, 2020.

[2] Information regarding Canada's Partnership Grants (stage 1) is found at https://www.sshrc-crsh.gc.ca/funding-financement/programs-programmes/partnership_grants_stage1-subventions_partenariat_etape1-eng.aspx, accessed May 12, 2020.

[3] Information regarding the European Research Council (ERC) consolidator grants is found at https://erc.europa.eu/funding/consolidator-grants, accessed May 12, 2020.

[4] Information regarding the ERC synergy grants is found at https://erc.europa.eu/funding/synergy-grants, accessed May 12, 2020.

[5] The website for the Worldwide Protein Data Bank is https://www.wwpdb.org/, accessed December 2, 2019.

[6] The website for the National Archive of Computerized Data on Aging is https://www.icpsr.umich.edu/icpsrweb/NACDA/, accessed January 2, 2020.

[7] The website for UniProt is https://www.uniprot.org/, accessed December 2, 2019.

[8] The website for the Monarch Initiative is https://monarchinitiative.org/, accessed December 2, 2019.

[9] For example, see https://grants.nih.gov/grants/guide/pa-files/PAR-20-089.html, accessed February 20, 2020.

[10] For example, see https://grants.nih.gov/grants/guide/pa-files/PAR-20-097.html, accessed February 20, 2020.

[11] The website for the Brainspan Atlas of the Developing Human is http://brainspan.org/, accessed December 2, 2019.

[12] The website for the Cell Atlas of Mouse Brain-Spinal Cord Connectome project is https://projectreporter.nih.gov/project_info_description.cfm?aid=9583948&icde=0, accessed December 2, 2019.

[13] The website for the Genito-Urinary Molecular Anatomy Project is https://www.gudmap.org/, accessed December 2, 2019.

[14] The website for the re3data registry of research data repositories is https://www.re3data.org/, accessed May 12, 2020.

(Wimalaratne et al., 2018). An analysis by the ELIXIR project lists more than 500 data resources available in Europe (Durinx et al., 2017). NLM currently lists 82 repositories on the NLM Data Sharing Repositories page.[15] The Neuroscience Information Framework (NIF) has maintained a registry for online biomedical data resources since 2006. NIF lists more than 200 such primary research data repositories serving biomedicine. Approximately 120 of these repositories explicitly list support from NIH, although information was not available or recorded for many. The total number of data resources, which includes data repositories, databases, knowledge bases, and atlases in NIF, is greater than 6,000.[16]

While some institutes are more invested in data repositories than others, as shown in Figure 1.1, the majority of NIH institutes and centers have funded one or more repositories. Reflecting the diversity of NIH institutes and other funding sources, the repositories cover a wide range of data types (Figure 1.2) and topics. At the same time, many generalist repositories exist that are hosted by institutions, nonprofits, and commercial entities that host data of all types, often deposited in the context of a published scientific paper. Some repositories, mostly institutional repositories and data centers supporting various research consortia, restrict data deposition, and sometimes access, to specific constituents.

Researchers generally have a choice as to where they store their data, whether the data are for private or public use. Many researchers have access to data repositories within their home institutions, or they can take advantage of specialist or generalist community repositories. In the absence of specific requirements coming from the funder or journal, the general recommendation from community organizations promoting data sharing is to use a community data repository specialized for a particular type of data (see OpenAIRE, 2019; e.g., protein-structure data might be deposited in the Protein Data Bank,[17] microarray data might be deposited in the Gene Expression Omnibus). Specialist repositories generally enforce community standards and have software available to help researchers comply with these standards. They also generally provide visualization and analysis tools that work with these specialized data types (e.g., the National Center for Biotechnology Information [NCBI] Basic Local Alignment Search Tool [BLAST][18]).

Most repositories do not charge the depositor a fee for submitting or hosting data (although this may be true only up to a specified size limit). For example, the Dryad Digital Repository[19] is operated by a not-for-profit organization and originally hosted data associated primarily with Earth-science publications. It now functions as a cross-disciplinary data repository that is integrated as part of the manuscript submission process for more than 1,000 journals, many in biomedicine. To offset costs, Dryad charges $120 for data deposition for data sets smaller than 20 gigabytes (GB). Depositors are charged $50 for each additional 10 GB.[20]

Much of the ecosystem of data repositories described above has been designed to share data with third parties for the purposes of transparency and reuse. However, not all data hosted by these repositories are open—that is, made available for anyone to use and distribute. A consortium data center, for example, may make the data available only to others in the consortium, and even open repositories may have requirements (e.g., they may require approval by the Institutional Review Board that governs data use). Institutional repositories, many of which are hosted by research libraries, generally provide services for private management or public sharing of research data only for researchers within their home institutions. Not all published data are hosted within a repository. Many journals allow authors to publish small data sets that live on the journal website as supplemental materials to their papers.

The biomedical data landscape is dynamic in that new data are constantly being generated, and new data infrastructures are continually coming into existence while older ones may migrate, merge, grow stale, or be taken down. The ELIXIR project of the European Union has defined different phases of a data repository—developing, mature, and legacy—and provides some characteristics of each phase (Durinx et al., 2017). The legacy phase refers to the state in which the repository is still online but no longer growing. The NIF project shows that, of the 200 data repositories listed, only 18 have gone out of service or have merged with other entities, and approximately 10 others

[15] The website for NIH's Data Sharing Repositories is https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html, accessed August 13, 2019.

[16] The website for the Neuroscience Information Framework is https://neuinfo.org/, accessed December 2, 2019.

[17] The website for the Protein Data Bank is http://www.rcsb.org/, accessed December 2, 2019.

[18] The website for the NCBI BLAST is https://blast.ncbi.nlm.nih.gov/Blast.cgi, accessed December 2, 2019.

[19] The website for the Dryad Digital Repository is https://datadryad.org/, accessed December 2, 2019.

[20] The website showing Dryad's data publishing charges is https://datadryad.org/stash/publishing_charges, accessed May 27, 2020.

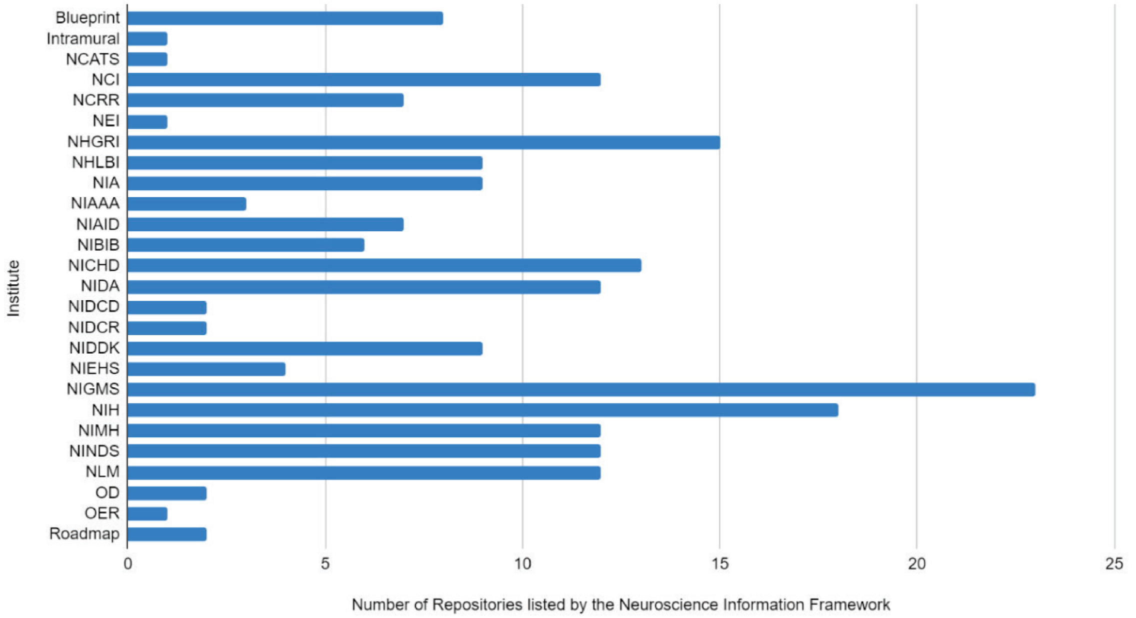


FIGURE 1.1 The number of primary biomedical research data repositories supported by institutes within the National Institutes of Health, excluding the resources hosted by the National Center for Biotechnology Information. NOTE: Some institutes may be under-represented. Based on data obtained from the Neuroscience Information Framework (www.neuinfo.org, accessed March 18, 2020).

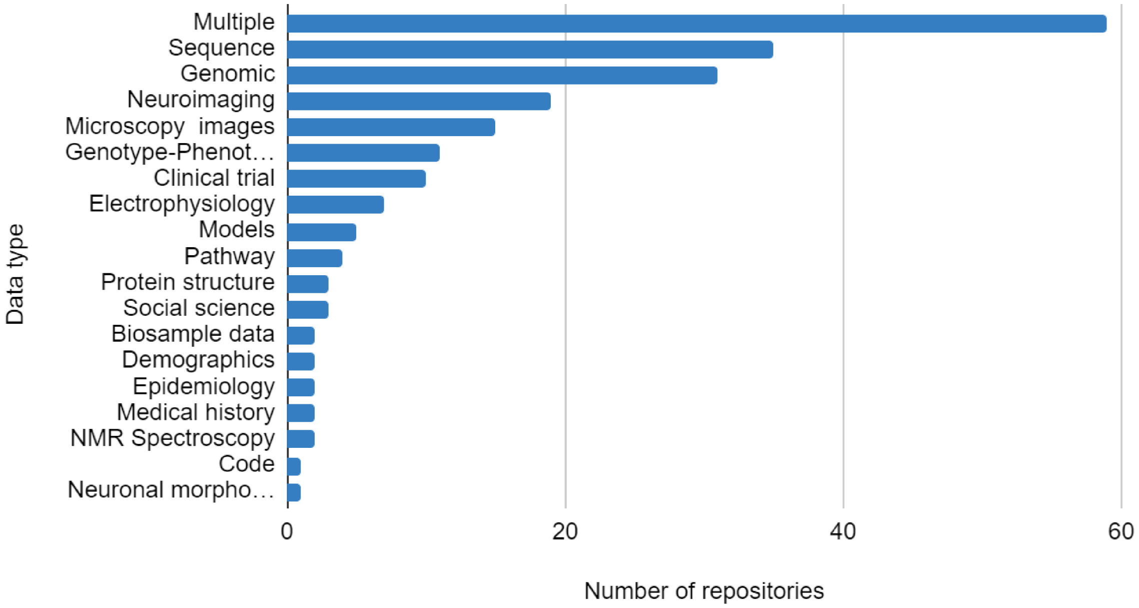


FIGURE 1.2 The number of repositories in the Neuroscience Information Framework Registry classified by data type. Based on data obtained from the Neuroscience Information Framework (www.neuinfo.org, accessed March 18, 2020).

remain online but appear to be no longer actively maintained. At the same time, more data repositories are being created. A search of the NIH Research Portfolio Online Reporting Tools[21] (NIH RePORTER), an online database of NIH-awarded grants, identifies another 128 data centers or data coordinating centers funded in 2018-2019 to support individual projects or consortia, many of which may be too nascent to be listed in repository catalogues.

Little information is available about what happens to data from resources that have been decommissioned. Occasionally, a resource is taken down and a plan for disposition and continued access to the data for a period of time is displayed on the website—for example, the Beta Cell Biology Consortium[22] (BCBC), a project funded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The BCBC data were transferred to the NIDDK Information Network[23] before the resource went out of commission.

Less understood and studied are the practices of researchers and institutions for managing their research data before they are made available in a repository or how much such repositories and services are used. Although data management practices in the laboratory are at the front line of eventual data sharing and long-term data access, there is often a lack of incentive for researchers to think about long-term curation and preservation needs, as they do not recognize a personal benefit (see Box 1.2). The policies around data stewardship and retention at universities may not have kept pace with the digital-data revolution. Although NIH and the National Science Foundation require researchers to have data management plans specified in their grant proposals, there are no requirements for how those plans are to be formulated (see Appendix B). The committee was not able to locate any research on how much of the data generated by researchers are transferred to a more stable entity for long-term stewardship, although funder mandates and requirements are assumed to play a role in populating public archives.

The variety of expertise and types of infrastructures and services required to work with diverse data make it unlikely that biomedicine will ever be served by a single large data resource; multiple archives and data repositories will continue to exist, even for the same type of data. The advantages of such an approach are specialized tools and services and a certain amount of robustness and innovation in the ecosystem. If these repositories use the same standards, federated search across them becomes possible. Nevertheless, multiple repositories impose a cost in that separate infrastructures, staff, and tools must be maintained at each site and may, in some cases, result in less value, and less data discovery, than might otherwise accrue from a more unified resource, particularly if different standards and formats are imposed by different repositories. Therefore, after a period of innovation and separate development, it may make sense to consolidate some resources. Merging of two active data resources will lead to costs of data transfer, harmonization of data, and adaptation of technologies.

In an effort to reduce redundancy and increase functionality, the Alliance of Genome Resources[24] has been formed as part of the *NIH Strategic Plan for Data Science*. The goals of the Alliance are to “establish a common infrastructure and software platform for data from all the [Model Organism Databases]; adopt updated data management practices; better integrate content, software, and user interfaces; improve interoperability; exchange best practices; and reduce redundancies of operation and maintenance.”[25]

In summary, the diversity and dynamism of data repositories and other infrastructures and the “black hole” of dark data (i.e., unpublished data; see Box 1.3) or data that are otherwise discoverable (e.g., Read et al., 2015) makes a true landscape analysis difficult. As indicated above, reliable data on the number and locations of such repositories, particularly institutional repositories, are difficult to determine,[26] and few may be set up for large volumes of data. Thus, a complete accounting of biomedical data infrastructures and their content is not possible, even for publicly accessible data. To date, there is no equivalent of PubMed for biomedical data sets, although

[21] The website for the NIH RePORTER is https://projectreporter.nih.gov/reporter.cfm, accessed December 2, 2019.

[22] The website for BCBC is https://www.betacell.org/, accessed December 2, 2019.

[23] The website for the NIDDK Information Network is https://dknet.org/, accessed December 2, 2019.

[24] The website for the Alliance of Genome Resources is https://www.alliancegenome.org/, accessed December 2, 2019.

[25] The website for the National Human Genome Research Institute is https://www.genome.gov/Funded-Programs-Projects/Computational-Genomics-and-Data-Science-Program/The-Alliance, accessed May 12, 2020.

[26] University research data policies provide little guidance. See, for example, https://doresearch.stanford.edu/research-scholarship/research-data, https://research.columbia.edu/research-data-columbia, https://libraries.mit.edu/data-management/, http://guides.library.jhu.edu/c.php?g=813898&p=6281112, and https://ogc.umich.edu/frequently-asked-questions/research/, accessed December 12, 2019.

BOX 1.2
Practices and Attitudes Related to Data Management in the Laboratory

A survey was conducted of 140 researchers, spanning different career levels, in neuroimaging (Borghi and Van Gulick, 2018). Respondents were asked about the types of data collected, tools used for data storage, organization and analysis, and the degree to which practices are documented and standardized within their research group. Survey results suggested that only about 25 percent of researchers engaged with their institutional resources and services for data management, although 45 percent reported using a technical infrastructure for managing their data during the course of the study. Most indicated that they used a common file structure (70 percent) and file-naming conventions (67 percent) for organizing their data. Overall, researchers were unequivocal in their belief that good data management practices were necessary to protect against loss of data and to ensure present and future access for at least their collaborators. Few indicated that practices they used were fully documented and mature.

Results identified several barriers to data management and sharing (see Table 1.2.1). The biggest limitation was the amount of time it took, followed by lack of best practices, and lack of training. More than half believe lack of training and best practices are constraints (assuming that the lack of best practices reflects a lack of training). This study suggests that regardless of whether there is data management infrastructure in the laboratory, more training is needed to improve the knowledge and application of best practices. If one infers that "lack of time" implies the absence of the ability to outsource tasks (at some cost) to campus or community resources, then it is possible there is a lack of capacity on campuses and within community resources and a lack of funds within project budgets.

TABLE 1.2.1 Research Data Management and Limitations

		Data Collection	Data Analysis	Data Sharing
Limits	The amount of time it takes	69.60	71.30	79.46
	Lack of best practices	43.20	48.70	49.11
	Lack of incentives	36.80	32.18	37.50
	Lack of knowledge/training	32.80	40.87	41.07
	The financial cost	17.60	8.70	22.32
	Other	7.20	6.09	5.36
Motivations	Prevent loss of data	100.00	85.83	78.57
	Ensure access for collaborators	76.80	73.33	70.53
	Openness and reproducibility	63.20	64.17	66.96
	Institutional data policy	52.00	39.17	47.32
	Publisher/funder mandates	35.20	28.33	41.96
	Availability of tools	12.00	9.17	8.93
	Other	3.20	3.3	0.0

NOTE: Limits and motivations for RDM during the data collection, analysis, and sharing/publishing phases of a research project. All values listed are percentage of total participants. More than one response could be selected. For limitations, "Other" responses included changes in personnel, differences in expertise within a laboratory, differences in preferences between laboratory members, lack of top-down leadership, and concerns about future cost. For motivations, "Other" responses included ensuring continuity following personnel changes, keeping track of analyses, preventing error, and maximizing efficiency. (Data collection: n = 125 [limits/motivations]; Data analysis: n = 115 [limits], 120 [motivations]; Data sharing: n = 112 [limits/motivations]).

SOURCE: Borghi and Van Gulick (2018).

continued

BOX 1.2 Continued

The survey revealed that data collection is not limited to measurements derived from the scanners; it also includes ancillary types of data (e.g., demographic data, study design, code). High percentages of respondents indicated that those data should be preserved for the long term (and therefore must be accommodated). Most participants indicated that they do not use formal tools to document their activities (e.g., data-analysis pipelines) but rather use a word processing program or "read me" files. Only 25 percent used version-control systems, and 20 percent used electronic notebooks such as Jupyter.[a] The use of laboratory management tools such as LabGuru[b] or Open Science Framework[c] was low (2.5 percent). Approximately 10 percent admitted they do not document their activities in any systematic way.

Overall, this and other studies (e.g., Barone et al., 2017) suggest that data management in the laboratory is an area that will require more attention and training. It is difficult to compensate later for lack of good documentation and organization during State 1 (the primary research environment).

BOX 1.3
Dark Data

Data stewards often do not have the domain expertise to understand the value of data to the long-term scientific enterprise, and yet they are often called on to make decisions about the disposition of data. If data are not properly documented with informative metadata, the scientific context of the data may not be understood. They become "dark data" that can neither be used nor disposed of and are therefore kept indefinitely. Committee members heard anecdotes of stewards "inheriting" data of unknown provenance or quality, and of servers housing multiple similar versions of the same data sets without any documentation describing what might have been done to them. As data sets become larger, these dark data will represent greater costs indefinitely. Dark data are good candidates for deaccessioning, but no standards exist for data stewards to do so.

there have been several efforts launched by NIH and others to construct them (e.g., DataMed;[27] Chen et al., 2018). None of these has been fully populated.

FAIR DATA

The research community has designed a set of principles to assist reuse of data by third parties (i.e., by those other than the data producers) and drive science discovery. The findable, accessible, interoperable, and reusable (FAIR) data principles offer guidance on how to design data and data systems to make data more reusable by both humans and machines (Wilkinson et al., 2016). These principles have seen rapid endorsement by funders in both Europe and the United States. Although the principles themselves are not recommendations for implementation, they do lay out a set of 15 attributes (see Box 1.4).

Communities attempting to implement FAIR principles recognize associated costs accrued by both the data provider and those providing data access. Some costs are short lived, but others are recurrent and long lasting. Some obligations imposed by FAIR can even be viewed in perpetuity—for example, the requirement that (meta) data be assigned persistent identifiers (e.g., Digital Object Identifiers). There are short-term costs associated with providing these identifiers and then long-term costs associated with ensuring that links between the identifier and

[27] The website for DataMed is https://datamed.org/, accessed December 2, 2019.

BOX 1.4
The 15 Data Attributes for FAIR Data

The FAIR data principles, formulated during a 2014 workshop and published by Wilkinson and colleagues (2016), offer guidance on how to design data and data systems to make data more reusable by both humans and machines. The principles and their attributes as defined by Wilkinson and others (2016) are listed below.

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, p. 4).

data are maintained. Similarly, the FAIR requirement that metadata are accessible, even when the data are no longer available, represents a small cost for an individual data set. However, in aggregate, the requirement imposes a societal obligation on ensuring that there are entities to maintain access to these metadata in perpetuity. Thus, implementing FAIR principles requires both technical infrastructure and organizational infrastructure. As data are transferred across entities and moved across the different stages of maturity, these services must be maintained.

On the other hand, although perhaps too early to tell, implementation of the FAIR principles also has the potential to contribute to long-term data sustainability, as agreement on and adherence to standards and best practices can conceivably lower the cost of porting data from one archive to another.

REPORT ORGANIZATION

The statement of task requests a general cost-forecasting framework that is applicable to all data resources and throughout the data life cycle. It also asks the committee to evaluate an array of considerations. To provide the basis for forecasting long-term costs for preserving, archiving, and accessing various types of biomedical data, Chapter 2 explores the three states in the data life cycle and their associated activities: (State 1) the primary research and data management environment; (State 2) an active repository and platform where data may be acquired, curated, aggregated, accessed, and analyzed; and (State 3) a long-term preservation platform. Chapter 3 describes the economics of cost forecasting. Chapter 4 presents the cost-forecasting framework and highlights the important cost drivers, the decisions about which may affect costs throughout the data life cycle. In Chapters 5 and 6, the committee demonstrates the application of the cost-forecasting framework in biomedical contexts. Chapter 5 applies the framework of the cost forecast for a new repository and platform for biomedical research data, and Chapter 6 applies the framework to forecasting costs for new research in the primary research environment. Chapter 7 discusses potential economic, technology, policy, and legal disruptors that could affect data costs in the future. In Chapter 8, the committee offers a set of strategies, actions, and needed advances that would foster an environment conducive to responsible long-term data management decisions and cost forecasting.

The cost-forecasting framework itself is not an instrument for calculating the dollars necessary to develop or manage an information resource, but rather it is a framework that assists the cost forecaster in developing his own instrument. Each application of the framework will be unique depending on the nature of the information resource, its users and contributors, available resources, and the point of view of the forecaster. Box 1.5 outlines the major elements of the framework found throughout the report.

BOX 1.5
Key Report Elements of the Cost-Forecasting Framework

Elements of the cost-forecasting framework are found in Chapters 2, 3, and 4 of this report, and in Appendix E. Below is a list that describes portions of the report that are fundamental to understanding and applying the cost-forecasting framework.

Chapter 2

- *Box 2.1* provides an overview definition of each of the three data states (environments) in which data might exist throughout their life cycles. Details about each of the data states are provided in sections following Table 2.1.
- *Tables 2.1-2.3* (describe each of the three data management environments, the activities associated with each of the three states, and the personnel that would generate staffing costs.

Chapter 3

- *Box 3.2* lists major cost components of a biomedical information resource.

Chapter 4

- *Table 4.1* outlines the steps for forecasting the costs of a biomedical resource.
- *Table 4.2* identifies the major cost drivers for the major activities that occur in each of the data states. More detailed descriptions of each of the cost drivers is provided in subsequent sections of Chapter 4.
- *Box 4.2* provides a description of how to use Table 4.2.

Appendix E is a template to assist the cost forecaster in identifying the characteristics of the data resource and data. It will direct the forecaster through decisions related to each of the cost drivers. Using the information identified through the completion of this template, the cost forecaster might then develop a decision tree or conduct other analysis to quantify costs for each of the cost components (listed in Box 3.2).

BENEFICIARIES OF THIS REPORT

Through its interactions with a variety of stakeholders, the committee determined that considering costs of preservation, archiving and allowing access to data beyond a short 1- to 2-year planning horizon is not part of common practice. Many researchers think about the disposition of their data after their primary research is complete and strive to make those data public. They may struggle, however, owing to the lack of financial and technical resources available once the performance period of the original research funding has ended. There are also researchers who considered the outcomes of their research to be journal articles rather than data and therefore put little thought into the long-term disposition of their data beyond that required as a condition of their research funding or journal policy. The responsibility of managing data is often transferred to the manager of a data-archiving platform, and managers of those platforms face new challenges associated with, for example, maneuvering within commercially available cloud storage and associated fee structures. Further, those managers may not be domain experts and may not understand how to retain the value of the data if the data are not properly standardized or documented.

While NLM commissioned this study, the cost-forecasting framework presented as part of this report is intended to be useful across multiple stakeholder groups, including the following:

- Researchers who need to estimate the costs involved in acquiring data, managing them effectively in the laboratory, and preparing them for submission to an archive;
- Graduate students and other shorter-term research staff who may not see or appreciate the long-term cost benefits of good decision making related to data collection and curation;
- Institutional officials at the researchers' home institutions—these institutions bear significant shared and shifted operating and capital costs to maintain data infrastructure and supporting staff;
- Archive managers who need to estimate costs when determining the amount of funding required to fulfill their mission and who may need to transfer their archives to platforms receiving greater or lesser use;
- Program officers or other funding agency staff who are launching new programs and need to anticipate costs across the different stages of the project, including long-term preservation and access; and
- Data preservationists who will need to estimate the costs for long-term preservation ahead of procuring or accepting data.

Expanding this conversation among these and other stakeholders will not only advance data preservation, archiving, and access, but it will also foster rich scientific discovery.

REFERENCES

Barone, L., J. Williams, and D. Mickloset. 2017. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLOS Computational Biology* 13(10):e1005755. https://doi.org/10.1371/journal.pcbi.1005755.

Borghi, J.A., and A.E. Van Gulick. 2018. Data management and sharing in neuroimaging: Practices and perceptions of MRI. *PLoS ONE* 13(7):e0200562. https://doi.org/10.1371/journal.pone.0200562.

Chen, X., A.E. Gururaj, B. Ozyurt, R. Liu, E. Soysal, T. Cohen, F. Tiryaki, et al. 2018. DataMed: An open source discovery index for finding biomedical data sets. *Journal of the American Medical Informatics Association* 25(3):300-308. https://doi.org/10.1093/jamia/ocx121.

Durinx, C., J. McEntyre, R. Appel, R. Apweiler, M. Barlow, N. Blomberg, C. Cook, et al. 2017. Identifying ELIXIR core data resources [version 2; peer review: 2 approved]. https://f1000research.com/articles/5-2422/v2.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2020. *Planning for Long-Term Use of Biomedical Data: Proceedings of a Workshop*. Washington, D.C.: The National Academies Press.

NIH (National Institutes of Health). 2018. *NIH Strategic Plan for Data Science*. https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf.

NLM (National Library of Medicine). 2017. Synopsis of *A Platform for Biomedical Discovery and Data-Powered Health: Strategic Plan 2017-2027*. https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport_Synopsis_FINAL.pdf.

NLM. 2018. NLM Launches 2017-2027 Strategic Plan. https://www.nlm.nih.gov/news/NLM_Launches_2017_to_2027_Strategic_Plan.html.

OpenAIRE. 2019. Guides for Researchers: How to Select a Data Repository? https://www.openaire.eu/opendatapilot-repository-guide.

Read, K.B., J.R. Sheehan, M.F. Huerta, L.S. Knecht, J.G. Mork, B.L. Humphreys, and NIH Big Data Annotator Group. 2015. Sizing the problem of improving discovery and access to NIH-funded data: A preliminary study. *PLoS One* 10(7):e0132725. https://doi.org/10.1371/journal.pone.0132735.

Wilkinson, M.D., M. Dumontier, I. Jan Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018.

Wimalaratne, S.M., N. Juty, J. Kunze, G. Janée, J.A. McMurry, N. Beard, R. Jimenez, et al. 2018. Uniform resolution of compact identifiers for biomedical data. *Scientific Data* 5:180029.

2

Framework Foundation: Data States and Associated Activities

The data life cycle begins when data are collected during the conduct of primary research and continues through data analysis, preservation and curation, reuse, storage, and potentially to deaccession. The data life cycle is not necessarily linear, and data may be reused and repurposed, combined with other data, and analyzed in a variety of ways and for different purposes throughout the existence of the data. How actively data are used during the data life cycle may change: they may be used often when initially collected, then see only periodic use after being placed in a repository. At some point, they may become dormant and be placed in an archive for long-term preservation. They may be rediscovered at any time and once again see active use. The environments in which the data are placed throughout their existence allow for different types of activities, and they may be moved from one environment to another as the need arises. The committee calls these environments "data states" and recognizes that the data may move from one state to another in a nonlinear manner. These data states were conceptualized by the committee to communicate the characteristics of different environments with different purposes, and different data storage and preservation costs. Note that they do not map directly to the data life cycle.

Digital data transition among three states over the research life cycle is described in Box 2.1.

BOX 2.1
The Three Data States

Digital data transition among three states over the research life cycle. The three states provide the framework for forecasting data storage, preservation, and archiving costs presented in this report. Data take a different form in each state, and each state includes different activities with different personnel, hardware, and management requirements. The labor and computation required to transform data from one state require significant resources.

- *State 1:* The primary research and data management environment where data are captured and analyzed. It could be possible that no one working in the data environment is focused on standardizing, documenting, sharing, or preserving data and algorithms.
- *State 2:* An active repository and platform where data may be acquired, curated, aggregated, accessed, and analyzed. This is an active information system that usually provides services to a wide range of users. Where data are complex, confidential, or very large, it may be a platform for controlling access and may also provide support for analyzing and processing data.
- *State 3:* A long-term preservation platform in which content is preserved across changes in governance, assessment of data value, and technology. The platform may include an extract of data from a single data set, multiple data sets, or an information system in a system-agnostic format. In this state, data are neither directly analyzable nor easily accessible.

Because research activities related to data may not occur sequentially, data might not transition through the three states sequentially during the life cycle of a research project or the life of a repository. A research laboratory may maintain data in State 1 for analysis, while transforming the data into State 3 for other purposes. An active State 2 repository may coexist with the same data stored on a State 3 long-term preservation platform, or the same data may be stored in more than one State 2 environment. A new research project may require a new State 1 environment with inputs resulting from transformations of multiple State 2 or 3 resources, in addition to capturing data from novel sources.

The three states and the activities involved in each of the states are summarized in Figure 2.1.1. The activities and subactivities in each state will be described in greater detail in later sections. The enumeration of the activities and subactivities have benefited from previous work, including the recent National Academies study *Open Science by Design* (NASEM, 2018) and the models proposed by "Keeping Research Data Safe" (Beagrie, 2019), *The Open Archival Information System* (Lavoie, 2014), *The LIFE^2 Final Project Report* (Ayris et al., 2008), the National Aeronautics and Space Administration's Cost Estimation Toolkit (Fontaine et al., 2007), and Palaiologk et al. (2012).

continued

BOX 2.1 Continued

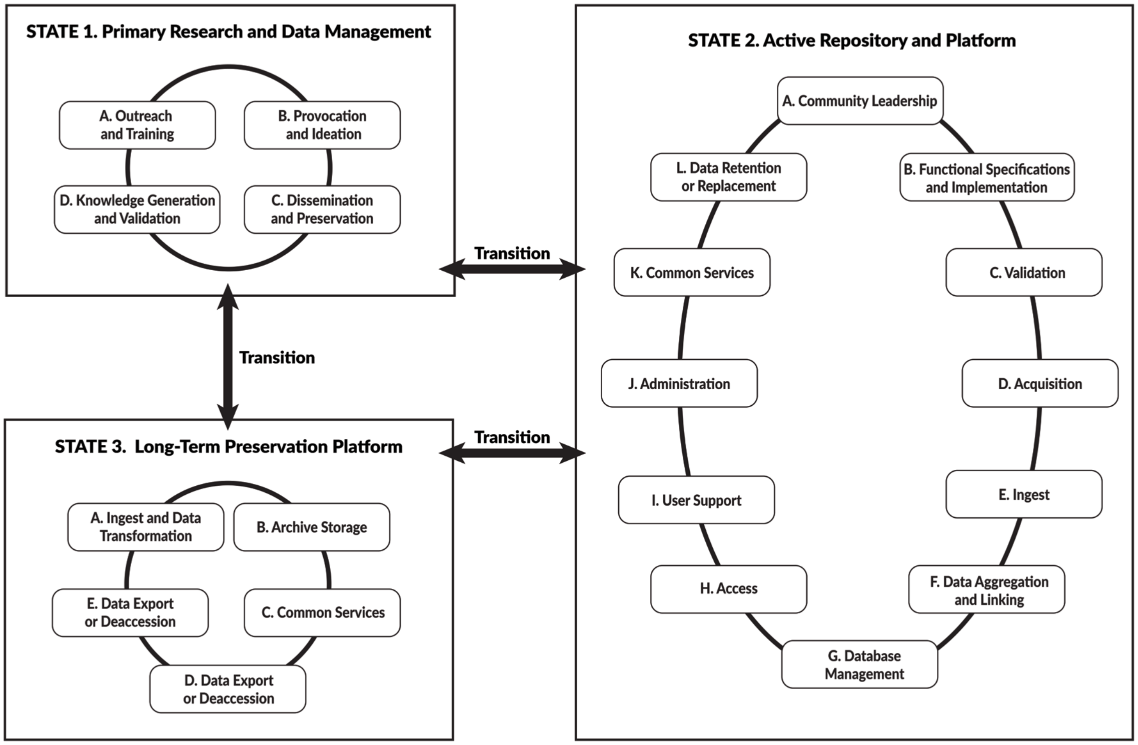

STATE 1. Primary Research and Data Management
A. Outreach and Training
B. Provocation and Ideation
D. Knowledge Generation and Validation
C. Dissemination and Preservation
Transition
Transition
Transition
STATE 2. Active Repository and Platform
A. Community Leadership
B. Functional Specifications and Implementation
C. Validation
D. Acquisition
E. Ingest
F. Data Aggregation and Linking
G. Database Management
H. Access
I. User Support
J. Administration
K. Common Services
L. Data Retention or Replacement
STATE 3. Long-Term Preservation Platform
A. Ingest and Data Transformation
B. Archive Storage
C. Common Services
D. Data Export or Deaccession
E. Data Export or Deaccession

FIGURE 2.1.1 Conceptual diagram showing the three data states, the primary activities within each state, and how data may transition between states. Note that the transition arrows between states are bidirectional, indicating that data already existing in repositories can transition back into a primary research environment when new data are incorporated, data are aggregated with other data, or data are used in new ways.

STATE 1: THE PRIMARY RESEARCH AND DATA MANAGEMENT ENVIRONMENT

The first state is the form that the data take in the primary research environment. The data are actively captured in this environment as they are created—for example, as digital sampling of electrical current, image and voice signals, text, or binary data. Computing ahead of storage (e.g., processing as data are generated) is generally fast enough to synchronously capture the data stream and to manage its conversion to data structures for quality assurance and initial analysis. The data management systems in this environment ideally include software features to manage disruptions in logical work units (if, for example, there is a disruption in electrical current as data are being transferred, the data flow needs to be corrected before completing the transfer). Multiple generations of backup may be needed to provide time to detect corruptions resulting from the addition of new data before those new data cascade across older backups.

Table 2.1 describes State 1 activities and subactivities as well as the types of individuals who carry out those activities. Personnel are specifically noted because personnel costs often account for the largest expenditures in data management activities. Relative salary levels of personnel costs are discussed in a later section following the discussions of the three data environments.

TABLE 2.1 State 1: Primary Research and Data Management Environment Activities and Personnel

Activity	Subactivities	Personnel
A. Outreach and Training Guidance on best practices in collecting and archiving data	1. Obtain support for creating funding proposals and data management plans (DMPs). 2. Obtain support for creating and describing research data. 3. Identify tools available for optimal data sharing.	Researcher, records management specialist, data scientist, data librarian, information technology (IT) systems engineer, education specialist, policy specialist
B. Provocation and Ideation Activities involved in exploring existing data resources and initiating the research activity	1. Explore and mine existing data resources for possible use and augmentation. 2. Design project with data sharing in mind. 3. Prepare funding application and explicit DMP (including estimates of costs of data storage and access). 4. Negotiate intellectual property rights. 5. Obtain ethics and regulatory approvals (e.g., Institutional Review Board [IRB], privacy office/Health Insurance Portability and Accountability Act, information security protocols).	Researcher, data scientist, software engineer, research domain project manager, IT security specialist, policy specialist, administrative staff
C. Knowledge Generation and Validation Activities involved in creating shareable research data	1. Evaluate and use tools for data collection, curation, and analysis. 2. Generate data and metadata using community-accepted standards. 3. Manage and document project data. 4. Validate data and code (including version). 5. Maintain active DMP/records.	Researcher, metadata librarian, data scientist, research domain project manager, research domain curator, software engineer
D. Dissemination and Preservation Activities involved in the disposition of the data	1. Prepare data and algorithms for submission to an active repository or long-term archive. 2. Transform data and algorithms as necessary in line with repository/archive submission requirements.	Researcher, research domain project manager, IT project manager, software engineer, data wrangler, research domain curator

STATE 2: THE ACTIVE REPOSITORY AND PLATFORM

The second state is the active repository and platform. Data are acquired from the primary research environment or from another active repository, or may be revived from archival storage for active use. Acquisition is asynchronous, either in near real time or in a batch form. Data are less volatile during acquisition in this state than they are in the primary research environment. In the ideal case, data may be curated as they are acquired to add metadata describing the data's provenance (i.e., the context that is implicit in the primary research environment and must be made explicit to accommodate use across research environments). Depending on the depth and quality of the data curation before it enters State 2 (including adherence to community data standards), the transition to State 2 may require extensive curation. Data sets are merged and aggregated with other data already in the active repository, which includes formatting, applying standards, and validating the data. The storage is fast enough to accommodate the search and analysis compute platforms used to make the data accessible. The data management systems in this environment necessarily handle much more data than the primary research environment because they aggregate data from multiple research projects. It is important to note that many State 2 activities will need to be repeated each time a new data set is added to the existing system. It is crucial that versioning and its documentation be controlled and curated. Failure to document and curate versions as they are created can lead to scientific errors with significant negative consequences. Costs incurred through activities in this state may reduce the efforts of future users of the data and for those transitioning data to other states or platforms.

Table 2.2 describes State 2 activities and subactivities as well as the types of individuals who carry out those activities.

TABLE 2.2 State 2: Active Repository and Platform

Activity	Subactivities	Personnel
A. Community Leadership Engagement with the broader community in the development of tools, standards, and best practices	1. Develop community data standards and best practices and policies. 2. Share lessons from development of repository systems and tools. 3. Identify community needs through community outreach.	Researcher, informatician, records management specialist, data librarian, communication specialist
B. Functional Specifications and Implementation Processes involved in designing or modifying and implementing the system for access and use	1. Design or modify and implement the repository infrastructure. 2. Consult with stakeholders on proposed design. 3. Design or modify and implement analytic tools. 4. Design or modify and implement search capabilities. 5. Design or modify and implement visualization tools. 6. Design or modify and implement authentication/authorization methods for secure access. 7. Design and implement user interfaces for data submission and access. 8. Design or modify and implement services for programmatic access to the data. 9. Design or modify and implement a private data enclave for researcher and collaborator use before access by other users of the repository. 10. Address findable, accessible, interoperable, and reusable (FAIR) compliance.	Senior staff, software engineer, informatician, research domain project manager, IT project manager, IT security specialist
C. Validation Processes involved in supporting the researcher in ensuring compliance with repository requirements	1. Provide a sandbox for researchers to test data sets for compliance with repository standards. 2. Test compliance with repository submission requirements. 3. Resolve errors. 4. Release data for submission.	Research domain curator, research domain project manager, software engineer
D. Acquisition Processes involved in acquiring the data	1. Apply selection policy to incoming data. 2. Provide support for and negotiate submission agreements with depositors. 3. Assess compliance with legal, ethical, and other policies (e.g., determination that secondary use is consistent with consent terms). 4. Revise selection policy as necessary.	Senior staff, data librarian, policy specialist
E. Ingest Processes involved in receiving and preparing the data for insertion in the repository	1. Receive submission. 2. Conduct quality assurance of submitted data. 3. Transform data into a format suitable for deposit and access (including possible deidentification). 4. Curate data: generate, validate, or upgrade descriptive metadata and documentation. 5. Assign unique identifiers. 6. Generate administrative metadata.	Research domain curator, research domain project manager, metadata librarian, data wrangler, IT project manager, software engineer
F. Data Aggregation and Linking Processes involved in merging and aggregating new data with existing data, and processes involved in linking to external databases	1. Integrate data with existing data in the data repository. 2. Link new data to external repository data, if relevant (e.g., link data to publications). 3. Link data to external data sets through database federation.	Software engineer, informatician, data scientist, research domain curator, research domain project manager

TABLE 2.2 Continued

Activity	Subactivities	Personnel
G. Database Management Services and functions for managing the repository	1. Maintain the integrity of the database. 2. Generate administrative reports from the database. 3. Back up data at additional storage sites. 4. Plan for potential disaster recovery.	Software engineer, IT project manager, IT security specialist
H. Access Services and functions for making the data available to users	1. If applicable, confirm identity or eligibility of user as a qualified user (e.g., IRB approval, Collaborative Institutional Training Initiative training). 2. Determination that specific proposal for secondary use is consistent with consent terms. 3. Design or modify and deploy search algorithms. 4. Prepare data for dissemination to user. 5. Deliver search results.	Software engineer, IT security specialist, IT project manager, informatician, policy specialist
I. User Support Services for making the repository useful to users	1. Develop or modify and implement training materials. 2. Staff a help desk. 3. Publicize the repository.	Software engineer, education specialist, communication specialist
J. Administration Functions that control the overall operation of the repository	1. Provide general management and oversight. 2. Develop and review policies and standards. 3. Monitor use. 4. Provide support for security assessment and audit. 5. Provide administrative support including billing for submission and usage, if required.	Senior staff, research domain project manager, IT security specialist, policy specialist, administrative staff
K. Common Services Shared supporting services	1. Provide operating system, network, and network security services. 2. Provide and renew software licenses. 3. Provide hardware maintenance. 4. Ensure physical security and disaster management. 5. Supply utilities.	IT systems engineer, IT project manager, facilities manager
L. Data Retention or Replacement Determining whether the data will be retained, replaced, transferred, or destroyed	1. Retain data, or 2. Replace data, or 3. Prepare data for transfer and transfer data and any transformation code to long-term archive, or 4. Destroy data.	Senior staff, research domain project manager, software engineer

STATE 3: THE LONG-TERM PRESERVATION PLATFORM

The third state is the long-term preservation platform. Content (e.g., data and code) are preserved in such a platform when it is anticipated that the data will not be actively used for the foreseeable future or if the resources are not available to maintain an active repository. For example, data from an active repository may be transformed into text, delimited strings, images, or other forms that may be viewed or processed without the content of the data management systems of States 1 and 2. This transformation enables preservation over tens to, perhaps, hundreds of years through changes in governance and computational technologies and may include compression (although compression could hinder preservation if corresponding decompression routines are also not preserved). Storage may be offline. Data may be rehydrated (see Box 2.2) as needed and moved back into an active environment, where it can be accessed and be more easily discovered.

BOX 2.2
Data Dehydration and Rehydration

Data dehydration and rehydration are terms used in this report as shorthand for the processes of transitioning data from one state to another. Data are said to be dehydrated when transitioning from an active platform (States 1 or 2) to a less active platform (usually a State 3 platform). A decision to dehydrate may be made, for example, when a sophisticated, high-function software platform reaches the end of its funding cycle and additional funds cannot be found to sustain it. Commonly, data are moved to some file repository as a series of flat files accompanied by metadata descriptions.

The following should be considered for data in a given State 2 (active repository) resource that are to be dehydrated:

1. What data should be preserved? Not all data can realistically be preserved in perpetuity. Decisions will need to be made about the potential value of data and the criteria that warrants their preservation.
2. Granularity: How should data be mapped to files? A general rule would be that any data object with an externally referenced identifier that might be found in the literature should be recoverable from the files and metadata exported from the State 2 platform in a reasonably straightforward way (e.g., the identifier corresponds to a file or a group of files).
3. What metadata should be exported to accompany the files? Note that the response leaves room for curatorial decision making. Some State 2 platform metadata about data objects may not export meaningfully or usefully (e.g., detailed editing histories attributed to specific users of the platform).

Hydration is not a tidy inverse of dehydration. It could be viewed as building a new State 2 hosting platform (or adapting an existing one) and importing data sets existing from one or more State 3 repositories. More broadly, it could be viewed as the process that a researcher (or group of researchers) needs to go through to move data from a State 3 environment to one that is directly useful and usable. This description is necessarily vague; to give one example of the kinds of issues here, when in a State 2 platform, computational or analytic tools might be part of the platform and thus provide a set of readily available capabilities for researchers reusing data on the platform. When the data transition to State 3, these tools are no longer there. Rehydrating data from State 3 may require a specific subset of these analytic capabilities for their intended use of the data; the reuser may need to rebuild these capabilities, or may be able to obtain them from existing tools. Note also that standards are an important enabler: to the extent that there are standardized file formats for classes of biomedical data—and tools that understand those standards—the barriers to some kinds of rehydration may be considerably reduced.

It will become important to collect and understand best practices about dehydration of data in State 2 platforms as they are decommissioned. These will evolve, continuously informed by subsequent attempts to reuse and rehydrate data from State 3 repositories. Importantly, best practices will also be informed by decisions *not* to reuse data given that the costs of rehydration are too high (and borne by the reuser). It will be valuable to understand these decisions, perhaps through reviews of research proposals that choose to collect new data or to ignore certain existing State 3 data for these reasons.

There will naturally be overlap in some activities in all the data states. The distinction between States 2 and 3 helps focus on the different issues that arise as one moves from facilitating active use to long-term retention. Those managing a State 2 information resource may make decisions related to a State 3 resource, and the movement from State 2 to State 3 could potentially be seamless. Following good archival practice, State 2 resource managers may automatically create preservation copies of the data as they are accessioned, or those data may be stored in a preservation format. Drawing a boundary between States 2 and 3 helps to ensure that decision-making processes also consider the challenges of long-term data preservation and their associated costs.

Table 2.3 describes State 3 activities and subactivities as well as the types of individuals who carry out those activities.

TABLE 2.3 State 3: Long-Term Preservation Platform

Activity	Subactivities	Personnel
A. Preservation Planning Services and functions for ensuring that the archive remains accessible over the long term	1. Develop preservation policies, strategies, and standards with particular attention to possible future data rehydration. 2. Develop preservation-metadata specifications. 3. Engage with and monitor the designated user community. 4. Monitor technology. 5. Develop migration plans.	Senior staff, records management specialist, curator, IT project manager, software engineer
B. Ingest and Data Transformation Processes involved in receiving and preparing the data for insertion in the archive	1. Receive data for long-term storage. 2. Check for errors in data transfer. 3. Transform data into a format suitable for deposit. 4. Generate administrative metadata.	IT project manager, records management specialist, curator, software engineer, data wrangler, data scientist
C. Archive Storage Services and functions for long-term data storage	1. Store data. 2. Replace media as needed.	Software engineer, IT project manager, IT security specialist
D. Common Services Shared supporting services	1. Provide hardware maintenance. 2. Ensure physical security and disaster management.	IT systems engineer, facilities manager
E. Data Export or Deaccession Functions involved in transferring custody of or deaccessioning data	1. Prepare data for transfer of custody, or 2. Deaccession data.	Senior staff, software engineer, research domain curator

PERSONNEL AND THEIR RELATIVE SALARY LEVELS

Based on published case studies (e.g., Palaiologk et al., 2012) and experience of individual committee members, personnel salaries often account for the largest expenditures in data preservation, curation, and access. Appendix C provides data drawn from occupational employment statistics for the relative salary levels shown in Table 2.4. Table 2.4 defines the roles of the personnel shown in Tables 2.1-2.3 and indicates a relative salary level (VH, very high; H, high; M, medium) for each of them based on information from Appendix C.

REFERENCES

Ayris, P., R. Davies, R. McLeod, R. Miao, H. Shenton, P. Wheatley, S. Grace, et al. 2008. *LIFE^2 Final Project Report*. http://discovery.ucl.ac.uk/11758/1/11758.pdf.

Beagrie, C. 2019. Keeping research data safe: Cost-benefit studies, tools, and methodologies focussing on long-lived data. https://beagrie.com/krds.php.

Fontaine, K., G. Hunolt, A. Booth, and M. Banks. 2007. Observations on cost modeling and performance measurement of long-term archives. NASA research paper in *PV2007 Conference Proceedings*. http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf.

Lavoie, B. 2014. *The Open Archival Information System (OAIS) Reference Model: Introductory Guide*, 2nd ed. (Charles Beagrie, Ltd, eds.). Digital Preservation Coalition. https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, D.C.: The National Academies Press.

Palaiologk, A., A. Economides, H. Tjalsmaand, and L. Sesin. 2012. An activity-based costing model for long-term preservation and dissemination of digital research data: The case of DANS. *International Journal on Digital Libraries* 12:195-214.

TABLE 2.4 Personnel Categories with Definitions and Relative Salary Levels

Personnel	Definition	Relative Salary Level
Administrative staff	Provides a variety of support functions for a project or program	M
Communication specialist	Trained in effective methods for publicizing and disseminating information to a broad audience	M
Curator	Often an archivist, trained in methods to describe and add value to data	M
Data librarian	Trained in the technical aspects of data management	M
Data scientist	Trained in quantitative methods for collecting, analyzing, and interpreting data	H
Data wrangler	Trained in methods for transforming data from one format into another and data cleansing for improved data interpretation	H
Education specialist	Trained in design, modification, and implementation of training materials relevant to data management and use	M
Facilities manager	Oversees and handles matters relating to the physical environment	M
Informatician	Trained in biology, medicine, or other health-related field and in quantitative methods for collecting, analyzing, and interpreting data in those fields	VH
IT project manager	Responsible for planning, executing, and overseeing a project; trained IT specialist	H
IT security specialist	Trained in methods to protect IT systems against inadvertent or malicious attacks	VH
IT systems engineer	Trained in implementing, monitoring, and maintaining IT systems	VH
Metadata librarian	Trained in the technical aspects of data standards	M
Policy specialist	Trained in relevant ethical, legal, and regulatory requirements	H
Project manager	Responsible for planning, executing, and overseeing a project	M
Records management specialist	Often an archivist, trained in managing data throughout the data life cycle	M
Research domain curator	Domain expert trained in methods to describe and add value to data	H
Research domain project manager	Domain expert responsible for planning, executing, and overseeing a project	H
Researcher	An individual who generates potentially shareable data while conducting research	H
Senior staff	Has a supervisory and decision-making role within an organization or program	VH
Software engineer	Trained in the design, implementation, testing, evaluation, operation, and maintenance of computer programs or databases	VH

NOTE: H, high; M, medium; VH, very high.

3

Cost and the Value of Data

Cost refers to the resources necessary to accomplish an objective. A cost forecast (i.e., a prediction of costs) is usually expressed in monetary terms and will be based on the quantity and nature of the items and services needed. Cost forecasts inform decision makers about the resources needed to transform an idea into a reality. Costs and cost forecasts come in a variety of forms, and there is a rich set of issues to consider in thinking about them. A cost forecast can inform actions needed to assemble necessary resources or to resolve any issues that would impede success. That may include identifying less costly solutions if the forecast suggests that original plans will face financing difficulties. Indeed, cost forecasts are vital in comparisons of alternative courses of action. Cost is not a measure of value (i.e., the benefit the objective produces), and cost forecasts may be unwelcome to the extent that they raise questions about the merits of an undertaking relative to the resources required. Decision makers focus on the best use of available resources; however, they will want good cost forecasts as the basis for their choices. A good cost forecast allows focus on the idea that resources expended in one use are unavailable for other potentially high-value uses. This "opportunity cost" is an important element of data management. Expending resources on keeping existing data sets available means fewer resources for funding new research activities. There is a trade-off in balancing the allocation of resources to maintain the by-products of past research and allocating resources toward new research.

Controversies created by cost forecasts help explain why so many governments in the United States and abroad use cash budgets (i.e., finance 1 year of costs at a time) versus acknowledging the total commitment implied by today's decisions through full funding of projects. A variant of this difficulty is created when a fixed horizon (e.g., a "5-year plan") is used to forecast costs. Large costs may be encountered just beyond that fixed horizon. As described later, this problem arises in planning for the long-term preservation and use of biomedical research data (States 2 and 3, in the terminology of this report), since typical previous State 1 primary research grants provided only for the performance period of the grant (a policy within at least the National Institutes of Health [NIH] that may change [DHHS, 2019]). To the extent that the biomedical research enterprise wants to ensure that good decisions about data management and data access are made at research project inception, or at any point in of the data life cycle (see Chapter 2), it is critical to address all cost components across the full life cycle of the data system.

Because many researchers and data scientists have had no formal education in economics, this chapter begins with a primer that introduces basic economics terms and concepts that may be encountered by the cost forecaster. The text delineates the principal economic issues in creating cost forecasts and the significant variables (i.e., "cost drivers") affecting the forecast in the biomedical research data life cycle. It then assesses which are most significant in each element of States 1, 2, and 3 and follows with a review of how the properties of these drivers influence costs. The chapter concludes with an illustration of how the cost drivers might affect forecasts for States 1, 2, and 3.

ECONOMIC ISSUES IN FORECASTING COSTS

When developing a cost forecast, it is important to understand that (1) all goods and services will incur costs to someone—even if a cost seems "free" (e.g., services provided by a university to a researcher, or the "free" access to data repositories managed by a government institution), (2) many costs are not incurred immediately, (3) many costs are not easily anticipated, and (4) cost burdens may shift. The first step when considering costs is to define precisely what one is trying to accomplish—in other words, identify what one is "buying."

Whose Costs?

From the perspective of the individual or organization that makes management decisions (e.g., a researcher, research institution, or repository host), the costs that matter most are those that must be financed from its budget. Those costs may be less than the total costs of a project or responsibility, distorting comparisons among competing courses of action. For example, a government agency that manages a data repository may underfund pension costs and omit overhead items such as facilities cost that a private organizations must include. A public agency or other organization that manages research or a repository may benefit from services such as a computing environment that is financed outside of its budget.

Parent institutions such as a university that provides services to component units such as research departments may insist that the costs of those services be incorporated into decision making. From the parent institution's perspective, the decision to proceed will trigger payment of those costs by the institution even if the immediate project manager (e.g., the researcher) does not have to finance them.

Sunk Versus Marginal Costs

It is important to distinguish between sunk costs (i.e., costs that have already been sustained and cannot be recovered, such as previously purchased computer equipment) and marginal costs (i.e., future costs, including costs for the next increment of effort, such as additional servers for data storage). Some sunk costs might be derived from reusing or redeploying previously developed software or infrastructure paid for by others. For example, existing open-source software might be incorporated as a component of a new data information resource, so some amount of development costs for that software would not be included in the present forecast. However, there will still be marginal costs for adapting, maintaining, and integrating that existing software that would need to be incorporated into the cost forecast. Marginal costs tally what costs beyond the fixed overhead that is already financed or investments that do not need to be repeated will be incurred if a project proceeds. Marginal costs might also change if there are savings derived from greater efficiencies incorporated into later project stages, for example, as a result of experience gained managing data or improvements in hardware technologies. Decisions are best informed based on marginal costs because the incremental resources required for a project may be better understood. This dictum is true even if an institution requires a budget to be prepared otherwise (e.g., to amortize a building that has already been constructed).

Because marginal costs may be difficult to calculate, many institutions rely on an average of past costs for their forecasts. If there are significant fixed costs for an activity (e.g., to create the data typology), average costs could decline as additional data sets are acquired. However, in that situation, the average will exceed the marginal cost, thus potentially overstating resource needs and unduly discouraging the next increment of activity. It is also possible that marginal costs could exceed the historical average, for example, if a new facility is required to accommodate a major expansion, and the historical average is based on a building for which the construction cost has not been adjusted for inflation. Presumably, an appropriate adjustment will be made, but the best safeguard against misforecasting is to invest the additional effort to understand costs at the margin.

Cost Versus Price

Institutions or individuals often base their cost forecasts on the prices charged for goods and services. Prices will vary with the extent of services provided. Prices for in-house services (e.g., on-premises data storage) may not cover all the elements needed for the data function (e.g., power, data center overhead), whereas a cloud provider's price is more likely to bundle these elements (along with a certain level of data security). A price reflects the amount of money needed to purchase a product or service, but it may not always accurately reflect the true cost to provide those inputs (i.e., all the resources that society must use). If, for example, another part of the institution or a different institution subsidizes a research program or data repository by providing "free" or discounted computer services or by lending staff to a project, the prices may understate production costs. This situation also arises if larger social effects are omitted that should be (but are not) reflected in prices (e.g., environmental effects of electricity generation and use required by data repositories). Again, some of these may not be incurred directly as monetary costs, or they may not be incurred immediately. They may result in future costs (e.g., power companies may need to charge higher rates as they become responsible for mitigating environmental impacts). On the other hand, market forces might lead to prices that overstate actual costs for goods and services (e.g., "excessive" download charges from a cloud service provider).

The issue of cost versus price is especially important to consider when projecting the cost of commercial services. Service providers may benefit from much greater economies of scale and thus lower cost than an individual institution or researcher, but their lower costs will not necessarily translate into lower prices for the science community. Even if prices accurately reflect past (marginal) costs, there is no guarantee that they will do so in the future. For example, the widespread adoption of new data practices (or even an adoption by a single large enterprise) could shift demand sufficiently to affect future prices (e.g., for a particular skill) in a way that is not captured by studying the past pricing history. Recent increases in salaries for data scientists is an example (see Box 3.1).

Institutions acting in the public interest may be instructed to include the full cost of producing a result or to avoid practices that impose social costs not reflected in market prices. They may be instructed to finance some of the subsidies (e.g., in the form of scholarships). They may be directed to reduce the impact of imperfections through how they procure an item (e.g., the statutory direction that the Department of Defense use Veterans Administration preferential drug prices).

Investment Versus Operating Costs and Their Time Profiles

The time profile of costs matters when comparing courses of action. That case certainly arises if funds for the immediate budget period are more difficult to obtain than those further in the future, when fewer commitments are perceived to be fixed and there is more discretion about how funds might be employed. Some one-time costs (i.e., investments) may be necessary to begin a project or a project's next phase. Sometimes, such costs must be expended periodically (e.g., the cost of hardware or software refreshment). These expenditures are often followed by a period in which operating costs require a lower level of continued resources.

BOX 3.1
Data Scientist Compensation

Based on interviews with more than 2,000 individuals described as predictive analytic professionals and data scientists, a 2019 study by Burtch Works Executive Recruiting (Burtch, 2019) reports a median entry-level base salary for a person with a bachelor's degree and without significant computer skills as $80,000. The salary of an experienced person with a Ph.D. and significant computer skills is reported to be $180,000. Meanwhile, the median base salaries of experienced managers are reported to be $250,000. These salary levels reflect the scarcity of the required advanced skills of the sort delineated in this report but also substantially exceed what research-oriented staff typically earn in the public sector and academia.

Two projects may have the same (total) forecast costs but very different time profiles. The standard solution to comparing these different cost streams is discounting (e.g., Mankiw, 2017)—for example, using a discount (interest) rate to price everything as a single payment made immediately (i.e., the "present value"). This present value can be thought of as a corpus that pays not just for first-period costs but for future expenses as well, using a combination of the principal and the interest theoretically (or actually) earned in the meantime. A "discounted value" recalculates future payments as the equivalent of payments made today on which the earnings at the discount (i.e., interest) rate plus the first-period principal would be exactly enough to cover future obligations. Controversy arises among stakeholders about the choice of discount rate, which affects how courses of action rank. A high discount rate diminishes the present value of future costs, a low one vice versa. Discount rates for U.S. federal agencies are usually mandated by the Office of Management and Budget (usually the rate on Treasury obligations).[1]

Buildings and other physical facilities present a special problem: they represent large expenditures in short periods, and there is an issue of cost recovery. If the institution already owns them and no refurbishment is needed, their costs may be viewed as sunk and thus omitted from forecast marginal costs. If another entity either owns or is renting the building on the institution's behalf (e.g., for a federal agency, the General Services Administration), the rent becomes part of the forecast cost. If the institution constructs the facilities for the purpose of the project under consideration, those costs become part of the project's early-period expenses.

Forecasting costs may be reasonably straightforward when investment costs occur early in the life of an individual project. Forecasters might draw on experience with similar recent projects, or they might even be based on bids for the specific project (with due allowance for how the contracting strategy and other factors might affect the actual price eventually paid). If periodic future investments are needed to sustain the project, forecasting could be more difficult, given changes in the marketplace that affect costs. Some costs may increase (e.g., owing to suppliers leaving the business), while others may shrink (e.g., from technological improvements such as those that have characterized computing power). As a result, there may be substantial uncertainty about future costs.

In reality, investment costs are often underestimated at inception, in part owing to the cost of developing necessary new technology (e.g., new software and hardware), the procurement costs of which may thus be greater than anticipated. Such inaccurate cost forecasts may reflect excessive optimism about what can be achieved, a lack of clarity or precision regarding what is to be accomplished, or deliberate "lowballing" on the part of a proposer seeking to win approval for an initiative.

Principal Elements of Operating Costs

For most public and private enterprises, the principal elements of operating costs are consumable inputs (e.g., power, vendor services) and direct labor (i.e., personnel). Both present interesting forecasting challenges. Box 3.2 lists the major costs to establish and operate a biomedical information resource. Uncertainty may arise from potential changes in the marketplace for non-direct-labor inputs. For example, what is the likelihood of changes in the cost of materials-based inputs (e.g., reductions in energy costs as a result of hydraulic fracturing versus any increase that a carbon tax would impose)? Are vendor prices for services likely to be stable? If not, how plausible are the mechanisms that drive change?

Estimating the cost of direct labor (i.e., for the personnel employed by the organization) may appear to be straightforward, but the institution must allow for the reality that wages are likely to increase over the longer term, driven by general inflation and productivity growth. Moreover, fringe benefits (e.g., health care) account for a major part of direct labor costs, and their costs may be driven by factors outside the institution's control. Also challenging is forecasting how much direct labor will be required for new or different projects (e.g., for activities such as curating a new data set, and for data security, integrity checking, and addressing the impacts of disruptions to access of those data sets). Collecting early data on what specific activities will be required in any of the data steps will help to improve estimates.

The institution will also need to consider changes in the composition of its direct labor force in its forecasts. A higher proportion of specialized skills would likely increase costs (e.g., more data scientists), whereas new

[1] See OMB Circular A-94, December 18, 2018.

BOX 3.2
Cost Components of a Biomedical Information Resource

The following list includes major costs borne by an organization charged with constructing and operating a biomedical information resource. These components will necessarily be visited during the application of the cost-forecasting framework (see Table 4.1) once the activities associated with the development and management of an information resource, as well as the major cost drivers associated with them, are identified.

- *Labor*—direct salaries and benefits.
- *Information technology (IT) infrastructure*—costs associated with the purchase, upgrade, and replacement of computers, storage servers, networking equipment, and software purchases.
- *IT services*—costs incurred in conjunction with installing, operating, and maintaining IT infrastructure, such as network connectivity fees, cloud provider fees, and repair costs.
- *Media*—costs of consumable storage formats, such as tapes and DVDs.
- *Licenses and subscriptions*—one-time or periodic payments for the use of data and software; subscriptions to or memberships in organizations that provide access to needed data or services.
- *Facilities and utilities*—costs of space for people and IT infrastructure, the power to run them, and other utilities; in some cases, these costs might be incorporated into institutional overhead.
- *Outside services*—amounts paid to or for entities outside the hosting organization, such as consultants, external auditors, off-site media storage, and training.
- *Travel*—costs for outreach activities, to convene governing boards, and so on.
- *Institutional overhead*—indirect costs for administrative and other support that are not direct costs in a unit's budget but that might be allowed to a greater or lesser extent in a contract or grant.

There are other "soft" costs that are more difficult to capture quantitatively (but that may nevertheless be compared across design and operation alternatives). One soft cost is the time users must expend to make use of the data. Such time costs for users may prove significant if data discovery is difficult. See Appendix D for a discussion of these types of costs.

data processing approaches (e.g., the application of artificial intelligence) or the ability to employ a more junior workforce might decrease them, after allowing for any required initial software or process investments. Since changes in the experience composition of the workforce are a product both of organizational and individual decisions, forecasts need to look at a distribution of potential developments for those elements not under direct control (e.g., retirement rates).

Relative Costs of Storage Media and Hardware

A difficult issue in forecasting costs for a data-intensive enterprise is how to deal with the information infrastructure (i.e., the storage media and hardware). First, this infrastructure may be provided by others (e.g., a university or other host institution), and the repository may be charged in such a way that prices and costs diverge substantially. Indeed, the repository may see only an operating cost—that is, charge for services—because the providing entity is making the actual investments. It will nonetheless be useful for repository managers to understand those underlying costs, if only to judge their reasonableness, especially in deciding whether reliance on another provider or investing in some or all of the infrastructure itself is a better course of action. Second, as is widely appreciated, IT changes rapidly, with implications for both the nature of the services the repository is providing (e.g., users wanting the latest level of capability) and for the costs the repository faces, as has been true for storage.

The choice of data storage media may also affect the short- and long-term costs of data storage. Data might be sitting in the potentially volatile main memory of the computer system (e.g., if they just came off an instrument or were output by a simulation model) or in online storage (e.g., solid-state or mechanical disk drives), from where they can be readily transferred to the information resource. They might be held in an offline storage medium, such as removable disk drives, compact discs, or tapes, in which case there will be costs in bringing the data back online by either automated or manual means. In some cases, the data may be stored in a deprecated medium (e.g., a Zip-drive disk), where finding a device to access the data could be a challenge. The data might even be in nondigital format, such as paper or photographs, which entails high costs for scanning or manual transcription (see, e.g., Nielson et al., 2015). The repository may face a one-time cost to shift to contemporary storage media, which may be less expensive on a life-cycle basis, or an ongoing challenge with cost implications for maintaining access to data using storage approaches for which commercial and technical support are shrinking. The relative costs of hardware and storage media for long-term preservation need to be compared in a systematic way, especially in light of how quickly options evolve. Example approaches to such comparison can be found in the literature (e.g., Merrill, 2017; Rosenthal, 2017).

Forecast Reliability

The reliability of a cost forecast is an important consideration. Procuring something new typically involves substantial uncertainty, which should be communicated (Manski, 2019). For example, during its information gathering for this report, the committee heard how the switching of service providers resulted in unexpected costs (see Box 3.3). Throughout the data life cycle, there will be a distribution of estimates for what is needed to sustain the activities in each data state. The Department of Defense, for example, recognized this reality in the call to budget to the "most likely cost" in 1981 (Greene, 1981) and in the later development and evolution of its "Better Buying Power Initiatives," which acknowledge that there is a distribution of potential cost outcomes (Kendall, 2017). Doing so quantitatively may be difficult, owing to a lack of data, and require substantial additional effort. But at a minimum, the cost forecaster owes decision makers a warning and discussion about the existence of those uncertainties—even if they cannot be precisely characterized.

BOX 3.3
Changing Behaviors Given Changing Storage and Compute Scenarios

The committee heard from researchers about their ability to "experiment" with data-intensive computations, at no additional cost to them, when data resources were hosted by their research institutions. However, when their data were moved to a commercial cloud, the same levels of experimentation resulted in unexpected and large computational bills at the end of the month. Once the cost consequences of their behaviors became transparent (requiring compute bills to be sufficiently granular)—and especially when they were responsible for some or all of those costs—the researchers learned to be more thoughtful and efficient. For example, they began to pilot their analyses before performing them on full data sets. Making people responsible for their costs, helping them understand that their actions generate costs for someone, and providing appropriate training might help reduce resource consumption with more efficient workflows. The information resource platform manager might develop a compelling narrative to alert researchers to storage and computational cost structures and the empowering benefits to researchers of forecasting their costs (Chodacki, 2019). The narrative could properly stress the researchers' larger responsibility to the research community.

ASSESSING THE VALUE OF DATA

Data constitute a different type of asset than physical infrastructure, and biomedical research data constitute a different type of data than those which are readily monetized in the commercial sector. The biomedical research community will want data valuation models that are able to attach value to the public good that a data resource can generate, and that can recognize the value society places on the institutions that support the data resources. The value of a single data set reflects factors such as its uniqueness, the number of times it is used, the cost per use, and the impact of each use (e.g., the change in prior probability of a hypothesis). Value differs from cost—it reflects worth in terms beyond the monetary. The value of data might vary for different purposes (e.g., initial discovery versus repeatability). And it may change with time, even if the data themselves do not change (e.g., a data set may lose value if it is superseded by another that is more precise or accurate, or it may gain value if better analysis techniques allow new knowledge to be obtained). Value may accrue in different states to different actors for reasons. A data set may be considered valuable while in regular use in a State 1 (primary research) or State 2 (active repository) environment (see Box 2.1), especially if seen to contribute to advancing science. However, it is difficult to forecast the value of data into the future because it is difficult to know how the data may be used or aggregated and repurposed.

A larger aggregated data resource has the potential to increase the value of a data set by increasing the number of uses and by increasing the benefit per use through linkage to other data sets. A data set increases the value of the aggregate data resource by contributing to its breadth (e.g., variety of data types added by the data set) and depth (e.g., number of instances of a data type or granularity of data in an instance). The degree to which the aggregate resource delivers on the potential to increase the value of a data set depends on how well it handles factors such as accessibility, discoverability, and analysis. The value of a data resource compounds if it sparks connections among diverse users. This compound value reflects factors such as the distribution of user backgrounds, geographic origins, and purposes, including research and nonresearch purposes. In the long term, the greatest value may be realized through the multiplier effect as heterogeneous data sets are aggregated and linked on novel computational platforms in ways that are impossible to predict at the time a data set is created. Uniqueness of a data set may be the best long-term predictor of value.

When determining the value of data from small, individual studies, an important factor is the extent to which they can be combined with other similar studies to increase statistical power. Studies that yield small sample sizes by the end of the study may be considered exploratory. If rigorous community standards and good data management practices have already been implemented by a laboratory, then submitting the data to a specialist repository will require less effort. The inherent value of the data may have increased, as it is more likely that they can be used with other similar data. On the other hand, if the data require significant formatting to meet community standards, the laboratory would have to expend significant resources preparing the data for submission. In this case, the data might be considered only moderately valuable and the researcher may choose to submit the data to a repository with less onerous requirements.

The lack of statistical power in smaller data sets is a key factor in current reproducibility problems (e.g., Ioannidis, 2005). Large efforts like the Human Connectome Project, the Alzheimer's Disease Neuroimaging Initiative, and the Adolescent Brain Cognitive Development (ABCD) initiative are producing large, well-aligned data sets, but these types of projects are not able to sufficiently sample the phenotype space either within or between conditions. Promising results are starting to emerge, however. It is possible and perhaps even advantageous at times to aggregate data from smaller studies to increase statistical power and to train new machine learning algorithms to take advantage of heterogeneous data. Aggregating heterogeneous data allows a more complete and robust model of preclinical research to emerge, as each individual laboratory samples a small slice of a larger, multidimensional picture (Ferguson, 2019; Williams, 2019). The work of Alan Evans (Moradi et al., 2017) in neuroimaging with multicenter data from the Autism Brain Imaging Data Exchange (ABIDE) database[2] also shows the importance of making available multiple independently acquired data sets. The ABIDE initiative includes two large-scale neuroimaging data collections, ABIDE I and ABIDE II, created through the aggregation of data sets

[2] The website for ABIDE is https://fcon_1000.projects.nitrc.org/indi/abide/, accessed December 12, 2019.

independently collected across more than 24 international brain imaging laboratories studying autism. Moradi et al. (2017) showed that machine learning algorithms that are trained across independently acquired data are more robust and generalizable than when trained on data from a single site. This work is consistent with that discussed at the Workshop on Forecasting Costs for Preserving and Promoting Access to Biomedical Data by Ferguson (2019) and Williams (2019), in which investigations using data from multiple laboratories or multiple genetic strains lead to more robust clinical predictions than investigations using more limited data. These results suggest that while a small individual data set on its own may be of limited value, when aggregated with other data, it can potentially increase the value of the pool of data. So to the extent that data are "multiplicatively integrative" through adherence to the findable, accessible, interoperable, and reusable (FAIR) principles and exposure through platforms that make them FAIR, their value increases. If, however, the data are shared through a platform where their discoverability is limited and where standards and curation are not enforced, then their value will be diminished.

To retain data value, reanalysis of data by either a researcher or a repository may be necessary to make them compatible with new data and to ensure that the results derived from the reanalysis are valid. For example, a widely discussed and controversial paper claimed that the statistical methods used in functional magnetic resonance imaging (fMRI) data analyses were leading to an overinflation of false positives in neuroimaging studies (Eklund et al., 2016). Detailed comparisons across major software packages in structural fMRI find major differences in the way that these packages calculate parameters such as cortical thickness (summarized in Kennedy et al., 2019). Software bugs or system dependencies may be uncovered that invalidate results drawn from older studies (Kennedy et al., 2019). Such reanalysis will entail costs that cannot be estimated in early cost forecasts because the evolution cannot always be predicted.

While technological volatility can make data obsolete as higher-quality or higher-resolution data become available, it can also increase the value of data in the long term, provided that the underlying data are valid. Their long-term availability ensures that these data may be reanalyzed with newer algorithms and approaches. As Eklund and colleagues note, "Due to lamentable archiving and data-sharing practices, it is unlikely that problematic analyses can be redone" (PNAS, 2016). In an analysis reported in 2019, Eklund and colleagues estimated that, of the total of 23,000 studies published in neuroimaging, up to 2,500 were likely affected by the misapplication of statistics. If the average cost of a neuroimaging NIH Resarch Project Grant Program (RO1) award to an investigator is $400,000 (Kennedy, 2014), then the total value of these 2,500 studies that must be discarded or reperformed is $1 billion. Thus, many in the neuroimaging community called for more long-term storage of primary neuroimaging data (Eklund et al., 2016, 2019; Kennedy et al., 2019).

Data value will also be related to the quality of the data. Higher-quality data are likely to be more valuable than lower-quality data, although quality control metrics for data are not always known and algorithms in the future may be able to account for suboptimal data characteristics (e.g., motion artifacts) and "rescue" these data for future use. The quality of the data is likely to be affected by the platform and standards used and how well they are supported by automated and human curation. Dr. Greg Farber (committee site visit to NIH, September 18, 2019) and Dr. Russ Poldrack (personal communication with M. Martone, September 18, 2019) described that automated pipelines catch many errors such as inconsistently named files that are difficult for human curators to identify, improving the overall quality of data and metadata submitted. However, human curators, using their human knowledge and insight, can catch discrepancies that software misses. So the same data set might be significantly increased in value when submitted to a repository with both automated and human curation in comparison to one that is submitted to a generalist repository that has minimal curation support. However, if the researcher carefully documents her data and adheres to community standards and best practices independently, data deposited in such repositories can be quite FAIR.

The perceived value of data influences preservation, access, and archiving decisions as well as decisions made regarding transition of data from state to state. Characterizing value for decision making might be related to the number of different tasks or decisions that the data support, and it might be possible to compare values of data sets without quantifying them. For example, if data set "A" supports all tasks that data set "B" supports, then one could assert that data set B's value is no greater than that of data set A. Some may attempt to relate the value of data with the cost of obtaining them, but data value does not necessarily correlate with the financial investment made to collect them. Decisions made about the disposition of data may be based on the cost to replace the data, but those decisions should also be informed by the data quality, use, and replacement costs. A data set can have many anticipated uses and thus be viewed as high value, but if those uses do not occur, the value is not realized. The cost to replace the data may change. The day may come, for example, when technology advancements make it less expensive to resequence an organism than to download its genome. On the other hand, some data may be irreplaceable (e.g., surveys done in the past with time-dependent results) or take a long time to re-create (e.g., the Framingham Heart Study [see, e.g., Tsao and Vasan, 2015]).

Identifying data as high value when in any of the data states will have cost ramifications. For data value to be realized, the data need to be discoverable and usable. Metadata are required to make the data more easily discoverable. Without at least some minimal standard for metadata tags, the data are destined to become "dark data"—data that are undiscoverable and, hence, unused (see Box 1.3). Services around the data may be required for data to be usable, and significant labor will likely be necessary to implement and provide those services, including maintaining data standards. From a scientific point of view, data have no value without proper standardization and documentation. Preserving the value of data in any state, particularly in the State 3 environment, requires using accessible formats and keeping high-level context information (e.g., "all these data came from the same clinic") so that the data are discoverable at reasonable levels of effort that make the search time worthwhile.

APPROACHES TO DATA VALUATION

Before considering the consequences of designating data as high value, it helps to consider how the different facets of value might be assessed for data. It may be useful to look at some commercial approaches to data valuation (see Box 3.4) to see what insights they offer. Information valuation is still a nascent discipline in the commercial world, in part because standard accounting principles do not permit data to be listed as an asset on a company's balance sheet (Laney, 2017). Nevertheless, commercial information clearly has a value, as witnessed by the stock market valuations and sales prices of information-intensive companies relative to their balance sheets. In particular, the committee finds the taxonomy of valuation approaches set forth by the Gartner Group to be informative (Laney, 2017). Some of the approaches are not well suited to biomedical research data. For example, the "market value of information" approach is of limited use, as biomedical data sets are generally not bought and sold in public marketplaces. However, at least three of the approaches seem relevant in the biomedical-information setting.

1. *Cost value of information (CVI)*: This approach equates the value of data with the expense of obtaining them. While this approach ignores a multitude of factors about a data set, it is useful in setting a target for how much it makes sense to invest in preserving a data set. Spending more than the replacement cost of the data should raise questions, although there are a few caveats. The replacement cost of the data set may differ from the original cost of obtaining it—perhaps dropping as technology improves or rising with labor costs. Also, replacing data takes time, so it may make sense to spend more than the CVI of a data set to preserve it so as to avoid gaps in availability.
2. *Intrinsic value of information (IVI)*: IVI is a nonmonetary metric based on the quality (i.e., correctness and completeness), scarcity, and expected lifetime of a data set. While IVI does not determine appropriate costs for preserving data sets, it can be used to prioritize expenditures among data sets when funds are limited.
3. *Business value of information (BVI)*: BVI is another nonmonetary metric based on the goodness and relevance of a data set for a specific purpose. In a biomedical research or clinical setting, BVI might be interpreted as the range of tasks or investigations that a data set enables and its adequacy to those ends.

With all of these approaches, it is important to recognize that the value of data need not stay fixed over time. As noted with CVI, the replacement cost may change with developments in technology or trends in labor costs. For IVI, quality assurance activities or the retirement of similar data sets can increase value. BVI is especially amenable to change. A data set may become useful in a broader range of tasks (or better suited to current uses) if it is combined with other data sets or new analysis methods are developed that can work with it.

BOX 3.4
Commercial Approaches to Assessing Data Value

Data valuation has received increased attention in the commercial domain, as more and more companies see their data assets as major drivers of revenue. While these companies are often focused on the use of data to achieve their "bottom line," the biomedical domain is focused on the value of data for research and clinical benefits. Nevertheless, it is useful to draw insights from several commercial approaches.

O'Neal (2012) posits that when data are not actively maintained, their value depreciates at a rate of 10 percent per year. That loss is based on data becoming outdated (e.g., old customer addresses and e-mails). While much biomedical data, such as gene sequences, will not become obsolete, other types, such as survey results or disease demographics, might. Another valuable approach from O'Neal (2012) relates to the analysis of sources of data-productivity loss (with notional formulas) for manual reconciliation of data, data access and retrieval, and project delays.

Schmarzo (2016) points out that data are an unusual currency: the same data can be used across multiple use cases and thus do not have the "transactional limitations" of money. He introduces the concept of developing formulas to "express the intangible but quantifiable prudent value" of data assets and considers key business initiatives as the basis for valuing data (Schmarzo, 2016). Accordingly, in the biomedical research setting, it might be possible to identify key goals that an information resource could enable.

Short and Todd (2017) present cases where data value is not fully retained during transfer (i.e., the utility to one entity is not what another might realize). Results from a survey of 30 companies and nonprofits revealed that most had no formal data valuation policies, but some had data-classification efforts (e.g., "critical," "important," "other"). Short and Todd (2017) explained that "value may be based on multiple attributes, including usage type and frequency, content, age, author, history, reputation, creation cost, revenue potential, security requirements, and legal importance. Data value may change over time in response to new priorities, litigation, or regulations." Thus, data value should be a composite of three sources of value: (1) *Asset value* concerns direct or indirect monetization. Direct monetization includes buying, selling, or trading data. Indirect monetization refers to when a new product or service is based on data, but the data do not change hands. (2) *Activity value* concerns the value of data in use. Unlike tangible assets, data value generally does not decrease with use and might, in fact, become more valuable with use. The benefit of different uses of the same data can also vary greatly. Capture, storage, and maintenance are the main costs for data, and the marginal cost of use can be very small. (3) *Future value* concerns how the value of the data would reflect in a balance sheet. This value could be quantified in several ways: based on transactions on similar data, the income or savings they produce, or development or replacement cost (Short and Todd, 2017).

Laney (2015) presents perhaps the most comprehensive treatment of information valuation. He describes two classes of measures for data, each with three information valuation methods: (1) *Foundational measures* focus on improving information management discipline, with consideration for the intrinsic, business, and performance values of information. (2) *Financial measures* focus on improving information's economic benefits, with consideration for the cost, market, and economic values of information.

REFERENCES

Burtch, L. 2019. *The Burtch Works Study: Salaries of Data Scientists and Predictive Analytics Professionals*. Evanston, Ill.: Burtch Works Executive Recruiting. https://www.burtchworks.com/wp-content/uploads/2019/06/Burtch-Works-Study_DS-PAP-2019.pdf.

Chodacki, J. 2019. Forecasting the Costs for Preserving and Promoting Access to Biomedical Data. Presentation to the National Academies Workshop on Forecasting Costs for Preserving and Promoting Access to Biomedical Data, July 12.

DHHS (Department of Health and Human Services). 2019. Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance. 48 *Federal Register* 60398 (November 8, 2019). https://www.govinfo.gov/content/pkg/FR-2019-11-08/pdf/2019-24529.pdf.

Eklund, A., H. Knutsson, and T.E. Nichols. 2019. Cluster failure revisited: Impact of first level design and physiological noise on cluster false positive rates. *Human Brain Mapping* 40(7):2017-2032.

Eklund, A., T.E. Nichols, and H. Knutsson. 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America* 113(28):7900-7905.

Ferguson, A. 2019. The Burden and Benefits of 'Long-Tail' Data Sharing. Presentation to the National Academies Workshop on Forecasting Costs for Preserving and Promoting Access to Biomedical Data, July 11.

Greene, R.D. 1981. DARCOM's new program and cost control system. *Army Research, Development*, *and Acquisition*, July-August. https://asc.army.mil/docs/pubs/alt/archives/1981/Jul-Aug_1981.PDF.

Ioannidis, J.P.A. 2005. Why most published research findings are false. *PLoS Medicine* 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124.

Kendall, F. 2017. *Getting Defense Acquisition Right*. Fort Belvoir, Va.: Defense Acquisition University Press.

Kennedy, D.N. 2014. Data persistence insurance. *Neuroinformatics* 12(3):361-363. http://doi.org/10.1007/s12021-014-9239-0.

Kennedy, D.N., S.A. Abraham, J.F. Bates, A. Crowley, S. Ghosh, T. Gillespie, M. Goncalves, et al. 2019. Everything matters: The ReproNim perspective on reproducible neuroimaging. *Frontiers in Neuroinformatics* 13:1.

Laney, D. 2015. Why and How to Measure the Value of Your Information Assets. Gartner Research. https://www.gartner.com/en/documents/3106719/why-and-how-to-measure-the-value-of-your-information-assets.

Laney, D. 2017. *Infonomics*. Abingdon, UK: Routledge.

Mankiw, N.G. 2017. *Principles of Economics*, 8th ed. Boston, Mass.: Cengage Learning.

Manski, C.F. 2019. Communicating uncertainty in policy analysis. *Proceedings of the National Academy of Sciences of the United States of America* 116(16):7634-7641.

Merrill, D. 2017. Economic perspectives for long-term digital preservation: Achieve zero data loss and geo-dispersion. White Paper, Hitachi Data Systems.

Moradi, E., B. Khundrakpam, J.D. Lewis, A.C. Evans, and J. Tohka. 2017. Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. *NeuroImage* 144(Pt A):128-141.

Nielson, J., J. Paquette, A.W. Liu, C.F. Guandique, C.A. Tovar, T. Inoue, K.-A. Irvine, et al. 2015. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature Communications* 6:8581.

O'Neal, K. 2012. The First Step in Data: Quantifying the Value of Data. https://dama-ny.com/images/meeting/101112/quantifyingthevalueofdata.pdf.

PNAS (*Proceedings of the National Academy of Sciences*). 2016. Correction to Eklund et al., "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates." *Proceedings of the National Academy of Sciences of the United States of America* 113(33):E4929.

Rosenthal, R.S.H. 2017. The medium-term prospects for long-term storage systems. *Library Hi Tech* 35(1):11-31.

Schmarzo, B. 2016. Determining the economic value of data. *InFocus*, June 14. https://infocus.dellemc.com/william_schmarzo/determining-economic-value-data/.

Short, J.E., and S. Todd. 2017. What's your data worth? *Sloan-MIT Management Review Magazine* 58:3.

Tsao, C.W., and R.S. Vasan. 2015. Cohort Profile: The Framingham Heart Study (FHS): Overview of milestones in cardiovascular epidemiology. *International Journal of Epidemiology* 44(6):1800-1813.

Williams, R. 2019. Presentation to the National Academies Workshop on Forecasting Costs for Preserving and Promoting Access to Biomedical Data, July 11.

4

The Cost-Forecasting Framework: Identifying Cost Drivers in the Biomedical Data Life Cycle

Thus far, this report has provided foundational information. This chapter organizes that information into a framework for identifying the major cost drivers for any biomedical research information resource. It can be applied by anyone who generates, collects, or manages data at some point in the data life cycle, or it may be applied by a funding or institutional official. The framework walks the cost forecaster through the various characteristics of data and information resources to determine which of those are likely to represent major cost drivers in the short and long terms. Cost forecasters will likely need to consult with multiple individuals with varied expertise to minimize uncertainty in the forecast.

The framework presented herein should be considered the basis of a cost forecast rather than a one-size-fits-all analytical tool for all applications. How it is applied in any situation depends on the circumstances, needs, and resources available to those involved. The activities, decisions, and cost drivers will be situationally dependent, and the framework provided herein will need to be modified to suit the specific purpose. In whatever application, however, the forecaster is encouraged to think beyond the costs associated with the specific data state being developed or managed. In the long term, it is more efficient to think early about how decisions may affect the costs of data management and access in future data states, the transitions to those states, and the future value of data to the scientific enterprise.

Making the right decisions about infrastructure can help to minimize many costs, as can taking advantage of economies of scale. But decisions need to be weighed against each other to understand their short- and long-term cost implications and their effects on data value. Many costs incurred over time are related to the curation, management, and preservation of different types of data, from different sources, that are generated by different evolving technologies and that are to be aggregated, accessed, and used in new ways. The cost-forecasting framework will help the forecaster identify the decisions to be made about the variables that can impact costs and the value of data in the short and long terms. The forecaster will necessarily focus on costs associated with the resource under development or being managed but will need to be aware of how decisions made in the earliest planning stages might affect long-term costs of data curation and use. Decisions made early in the planning process might increase the efficiency of future data curation and use or make future data curation prohibitively expensive.

Table 4.1 provides a framework for conducting a cost forecast. Subsequent sections of this chapter describe how to accomplish the steps. It should be noted that although these are presented as “steps,” they are actually activities that may occur concurrently, or iteratively as new information is gathered.

TABLE 4.1 Steps for Forecasting Costs of a Biomedical Information Resource

1. Determine the type of data resource environment, its data state(s), and how data might transition between those states during the data life cycle. The data states are defined in Box 2.1: State 1: primary research environment State 2: active repository State 3: long-term preservation and archive	• Decide the goals and objectives for the data resource. • Consider how the resource is likely to be used now and in the future. • Identify available guidance that defines the type of resource to be created or managed (e.g., requests for application, community standards, or institutional requirements). • Compare the items above with the activities defined for each of the data states (see Figure S.1) and decide which data state(s) best align(s).
2. Identify data characteristics (Chapter 4), data contributors, and users.	• Fill in the cost-driver template (Appendix E): o Complete category A of the template to help to identify the size, complexity, metadata requirements, depth versus breadth, processing levels and fidelity, and the replaceability of the data. o Complete category E of the template to help to identify the life-cycle issues. o Complete category F of the template to help to identify data contributors and users.
3. Identify the current and potential value of the data and how the data value might be maintained or increased with time.	• Consult with the institution hosting the data resource, the project funders, and the broader research community to develop appropriate metrics for assessing the value of the data. • Identify decisions that affect data value in the shorter and longer terms (see Chapter 3 for different methodologies). • Consider how data generation methodologies affect short- and long-term data value in terms of data contributors and users and the data life cycle.
4. Identify the personnel and infrastructure likely necessary in the short and long terms.	• Identify the major activities and subactivities associated with the information resource, including activities related to potential transitions between data states (Tables 2.1, 2.2, and 2.3) • Identify short- and long-term staffing requirements for the current state and transition between states. • Identify the infrastructure requirements and available resources.
5. Identify the major cost drivers associated with each activity based on the steps above, including how decisions might affect future data use and its cost.	• Identify the major cost drivers and associated uncertainties for each of the activities identified above by completing the cost-driver template (Appendix E). • Identify likely relative costs (e.g., using Table 4.2). • Consult with institutional experts (e.g., at the institution hosting the resource, library resources) and determine available personnel and infrastructure resources. • Work with experts at the host institution to quantify short-term costs and to bound uncertainties in longer-term forecasts.
6. Estimate the costs for relevant cost components based on the characteristics of the data and information resource.	• Identify which cost drivers are important for each cost component of the information resource (e.g., labor, information technology [IT] infrastructure and services, media, licenses and subscriptions, facilities and utilities, outside services, travel, and institutional overhead; Box 3.2). • Estimate costs for the current funding period. • Estimate costs and cost uncertainties for future funding periods, including costs to transition data to other states.

CONSULTING WIDELY TO CONDUCT A COST FORECAST

The cost forecaster may be a researcher but could also be a funding or institutional official. The steps outlined in Table 4.1 require the cost forecaster to develop a narrative regarding the biomedical information resource and the data contained within it that considers the entire life cycle of the data. To envision the data life cycle, which likely extends beyond any single funded performance period, the forecaster may need to consult an array of experts to understand how decisions made about data and the information resource affect the data life cycle and costs. When identifying how the information resource is to be used both in the present and future (step 1 in Table 4.1), the forecaster may refer to the request for application (RFA) from a funding agency, consider the goals and objectives of relevant research, and consult experts within the institution that will host the information resource about available assets. An RFA may require or provide guidance regarding specific data curatorial or preservation activities that will define the type of resource to be created or managed, or the researcher or research community may have specific needs or standards to be met. Aligning those activities with the activities associated with States 1, 2, or 3 (see Tables 2.1, 2.2, and 2.3) will help the forecaster determine the data state of the information resource. Because the scientific enterprise benefits from preserving the long-term value of the data and from increasing the efficiency and effectiveness of long-term data curation and use, activities related to eventual transfer of data to other states need to be considered. To identify data characteristics, data contributors, and data users (step 2 in Table 4.1), the cost forecaster will need to work with her institution, project funders, and perhaps the broader research community to identify or develop appropriate metrics to better understand and manage costs.

In spite of the fact that metrics for determining the costs and value of biomedical research data are immature, the framework explicitly makes identification of the current and potential value of the data an integral part of the cost-forecasting process (step 3 of Table 4.1). The long-term value of data comes with their being discovered, aggregated, and reused. Repositories that commit to archiving all data submitted for some designated minimum period (i.e., to satisfy a research funding agency's data archiving requirements), regardless of the current or future value of those data, risk using valuable resources for little return on investment. Chapter 3 provides information regarding the economics of cost forecasting and for data valuation.

Identifying the personnel and infrastructure necessary in the short and long terms (step 4 of Table 4.1) requires an identification of the activities and subactivities associated with the desired data state. The forecaster will refer to Tables 2.1, 2.2, and 2.3 that describe the high-level activities and subactivities associated with each of the data states as well as the personnel who might be required for each activity to develop and maintain an efficient and sustainable data resource. The forecaster will need to consider the goals for the information resource under consideration to identify the appropriate activities. Table 2.4 describes many of the categories of personnel required for all the activities and subactivities, as well as their relative salary levels. The forecaster can work with the institution to determine how personnel resources might be acquired and then with those staff to help determine how physical infrastructure needs may be met. Some design or modification and implementation costs related to the information resource capabilities (e.g., persistent identifiers, citation management, and search) might be avoided if open-source software can be used. However, use of open-source software still incurs costs, such as those related to integrating with other repository components, updating the software as new versions are available, and harmonizing the user interface with the overall look and feel of the repository. Consulting with IT professionals, metadata librarians, software engineers, and many others may be necessary to compile the information necessary to identify the major cost drivers (step 5 in Table 4.1).

MAPPING COST DRIVERS TO ACTIVITIES IN EACH DATA STATE

The fifth step in the forecast is identifying the cost drivers and decision points associated with each anticipated activity and how those decisions might affect the ways data may be used, as well as the cost of those uses. If one were to forecast the costs of manufacturing a physical product (e.g., a digital camera), one would want to know how it will be used and distributed, its specific features (e.g., megapixels, memory capacity), and desired characteristics (e.g., long battery life, small form factor). It is also desirable to understand the properties of the components that go into the product (e.g., microprocessor power consumption, defect rate on lenses). Similarly, in costing a biomedical information resource,[1] its intended content, capabilities, and context—and the properties of data that will populate it—must be understood. In this section, those dimensions of a biomedical information resource likely to have the greatest effects on cost are considered. In each case, choices are laid out along the dimensions or range of variation in the data and the manner in which they may influence costs. Box 4.1 provides some examples in the biomedical research field regarding decisions and actions that affect costs in the short and long terms.

Table 4.2 is a generalized matrix developed by the committee that shows which cost drivers, identified by the committee and described later in this chapter, are most likely to affect the costs of specific activities in each of the data states described in Chapter 2. Although individual research activities, databases, and archives may generate costs differently than as depicted based on requirements for particular data sets or research platforms, Table 4.2 provides useful information when conceptualizing costs into the future. In most cases, the cost of long-term data preservation will not be accrued by a single individual or institution, but rather responsibility at different stages may be transferred from, for example, a researcher to a data platform host. Understanding where costs will be accrued and who has managerial responsibility for them will inform decision makers for all data states. Box 4.2 provides guidance on how to use Table 4.2. The individual cost drivers and decision points that can affect those cost drivers will be discussed in later sections.

Table 4.2 is a high-level summary of the main cost drivers influencing each activity. Some of the activities, listed in the columns in Table 4.2, are affected by a large number of cost drivers. People engaged in those activities (or costing of them) need to be sure not to focus too narrowly on one or two of the cost drivers (rows in Table 4.2) in decision making and planning. Activities affected by the most cost drivers are listed below. Definitions for all of these, and questions to guide decisions around them, are provided later in the chapter.

- *I.C Knowledge Generation and Validation*. This item encompasses subactivities for creating shareable research data. These activities are critical for promoting use and controlling preservation costs at later states in the data life cycle. Many of the cost drivers, such as metadata requirements, persistent identifiers, and quality control, reflect up-front work that benefits downstream use.
- *II.B Functional Specification and Implementation*. It is not surprising that this activity is influenced by many cost drivers, as it includes a large number of subactivities involving design or modification and implementation of all of the main repository components.
- *II.F Data Aggregation and Linking*. The large number of major cost drivers for this activity may indicate two general conclusions. One is that the nature, quality, and amount of data in a repository strongly influence the effort required for successful aggregation and linkage. The second is that data linkage, especially to external sources, creates dependencies that must be managed whenever data at either end of a link change.
- *II.L Data Retention or Replacement*. The number of cost drivers influencing these activities perhaps highlights the complexity of decisions about data retention, encompassing characteristics of the data, users and uses of the data, and constraints that regulate the data.
- *III.B Ingest and Data Transformation*. This activity has many cost drivers in common with Knowledge Generation and Validation. That similarity is not surprising—it reflects that decisions during the generation of data have a major influence on activities at the end of their lifetime (and about where costs are borne). For example, rigorous metadata requirements on the initial collectors of data require more effort of those people but simplify the job of those charged with archiving the data later.

It also is evident that certain cost drivers (rows in Table 4.2) affect many activities (columns in Table 4.2). When specifying and scoping a biomedical information resource, special attention should be given to these cost drivers because the ramifications of decisions related to them will strongly influence costs. The cost drivers that affect the most activities are listed below.

[1] "Biomedical information resource" is used in this chapter as a generic term for a system for storing and accessing biomedical information, across all the states introduced in Chapter 3. It might be a group workspace in a single laboratory (State 1), a public repository (State 2), or a cold-storage archive (State 3), among other possibilities.

BOX 4.1
Actions That Affect the Cost of Data

The cost of preserving and providing access to data depends on choices made at a number of points across the data life cycle and on the presence of certain kinds of tools, institutional support, and incentives that affect the choices made at those points. These choices often predate the launch of an individual research project in which data are generated. A number of factors may affect the cost of preserving, archiving, and promoting access to data. Funder requirements (e.g., specific data-sharing or identifier requirements), data management mandates, Institutional Review Board specifications, federal regulations, and journal requirements all influence costs across the data life cycle. Data management plans that incorporate costs and value across the data life cycle may reduce the cost and time required for later data deposit and sharing.

Data for which a research community has developed shared standards may result in lower costs for the community and decrease the cost of repository governance and maintenance while creating the potential for those data to be of higher scientific value. In some cases, tools may already exist that help researchers prepare data according to those standards such as through the National Institutes for Health (NIH) National Library of Medicine (NLM) clinical trials database,[a] which requires researchers to standardize study-level metadata. The National Data Service Consortium Sustainable Environment/Actionable Data[b] and the Center for Open Science[c] both provide research platforms that allow researchers to analyze and publish data directly to an active repository, although they do not require data standardization. The National Institute of Mental Health (NIMH) Data Archive (NDA) houses several large neuroimaging studies as well as data collected from individual studies and has implemented standards that help to harmonize across them.[d] Colectica[e] and the Norwegian Centre for Research Data's Nesstar Publisher[f] both offer tools for managing metadata for rectangular, quantitative data published to community-normed standards.[g] The Stanford Center for Reproducible Neuroscience[h] provides tools to standardize neuroimaging data based on emergent community standards.

Capture of provenance metadata in the primary research environment for potential future reuse may reduce longer-term curation costs for an active repository. Similarly, making data discoverable and interoperable may be less expensive if data collection is informed by prior practices and shared standards. Costs of data preservation later in the digital life cycle may be reduced if researchers are provided with tools that allow them to curate data according to community standards and if they use those tools in the creation and analysis of data. Purging data from the primary research environment, after it has been validated and accepted by the active repository, reduces the long-term costs of the primary environment. Transformation of data sets in an active repository into a state suitable for long-term content preservation (e.g., by data compression followed by error-correcting coding) may protect against data corruption. The overall cost of a repository in which data are actively added or improved or for which tools for working with the data are integrated may be high, but the potential scientific benefit may be large. Similarly, the transformation of an active data management platform to a novel computational platform requires reinvestment in curation at costs similar to initial platform development costs (see, e.g., the model of centralized versus decentralized curation discussed in this chapter). Purging the data sets from the active repository as access approaches zero reduces the cost of the active repository, but costs will be incurred at a future date if those data are to be once again made easily accessible.

[a] The website for NLM's clinical trials database is www.ClinicalTrials.gov, accessed December 4, 2019.

[b] The website for the National Data Service Consortium Sustainable Environment/Actionable Data is www.sead-data.net, accessed December 4, 2019.

[c] The website for the Center for Open Science is www.cos.io, accessed December 4, 2019.

[d] The website for NIMH's NDA is https://nda.nih.gov/, accessed January 11, 2020.

[e] The website for Colectica is https://www.colectica.com, accessed December 4, 2019.

[f] The website for the Norwegian Centre for Research Data's Nesstar Publisher is http://www.nesstar.com/software/publisher.html, accessed December 4, 2019.

[g] Both Colectica and Nesstar use the Data Documentation Initiative (DDI) metadata standard. The website for DDI is https://ddialliance.org, accessed December 4, 2019.

[h] The website for the Stanford Center for Reproducible Neuroscience is http://reproducibility.stanford.edu/resources/, accessed December 4, 2019.

- *A. Content*. That the size, complexity and diversity of data, and metadata requirements facets of content affect so many activities is not unexpected. The effort required in many activities scales directly with these aspects.
- *H. Confidentiality, Ownership, and Security*. The prevalence of these cost drivers across many activities—especially confidentiality and security—derives in a large part from the prevalence of human-subjects and animal-model data in the biomedical domain.
- *J. Standards, Regulatory, and Governance Concerns*. The applicable standards cost driver influences a number of activities, which reflects the two sides of standards: there is effort required to conform to them, but dealing with data that conform to standards often facilitates other activities. The regulatory and legislative cost driver also impinges on many activities, which, again, possibly arises from the extensive use of human and animal data in biomedicine.

INDIVIDUAL COST DRIVERS IN THE DEVELOPMENT AND OPERATION OF A BIOMEDICAL INFORMATION RESOURCE

There is a wide variety of biomedical information that is worth preserving and sharing, from genomic sequences to clinical outcomes. Because of the variation in the content and other aspects, the costs of constructing, maintaining, and accessing such information can differ greatly. This section describes the main ways biomedical information resources may vary and why each variation is likely to affect costs or utility. The variations are grouped into more general categories in the next subsections, which are numbered to correspond with the categories provided in Table 4.2. When considering costs of alternatives, total costs related to managing, accessing, and using data need to be considered—both those costs borne by the operators of the resource as well as those of the users. Decisions regarding quality, delivery, or stewardship of data that will populate the resource can all drive up costs (or limit the value of the resource) as well. The more ambitious the plans for a biomedical information resource, the more personnel and financial resources will be required to support it but the greater potential benefit for users of the information resource and to scientific discovery. Thus, understanding the properties of the data is essential for estimating the costs involved with a biomedical information resource so that the forecaster can understand the short- and long-term trade-offs necessary related to each of the cost drivers. Cost drivers related to issues with input data and their disposition are called out in some of the subsections below.

Table 4.2 will help the cost forecaster understand which information resource-related activities will likely be important cost drivers to short- and long-term costs. Questions to help the cost forecaster identify key decision points for each cost driver are provided. The questions are written at a high level and intended to help the forecaster identify areas where more detailed lines of inquiry are warranted. When forecasting costs, these are the types of questions a cost forecaster needs to ask about each cost driver. How the forecaster answers these questions affects not only the cost of managing a given data state but also future costs for users or future data state managers.

The questions have been compiled into a blank table in Appendix E that could be used as a template when considering long-term costs. The template could help the forecaster organize the detailed narrative necessary to realistically assess activities that promote efficient and effective data preservation and use. The narrative can then drive a detailed quantitative analysis of the costs based on the resources available to the forecaster. Examples of how the template can be applied are provided in Chapter 5. Appendix F compares cost drivers for three hypothetical biomedical information resources (one for each data state).

TABLE 4.2 Drivers Affecting Cost of Data-Related Activities in the Three Data States

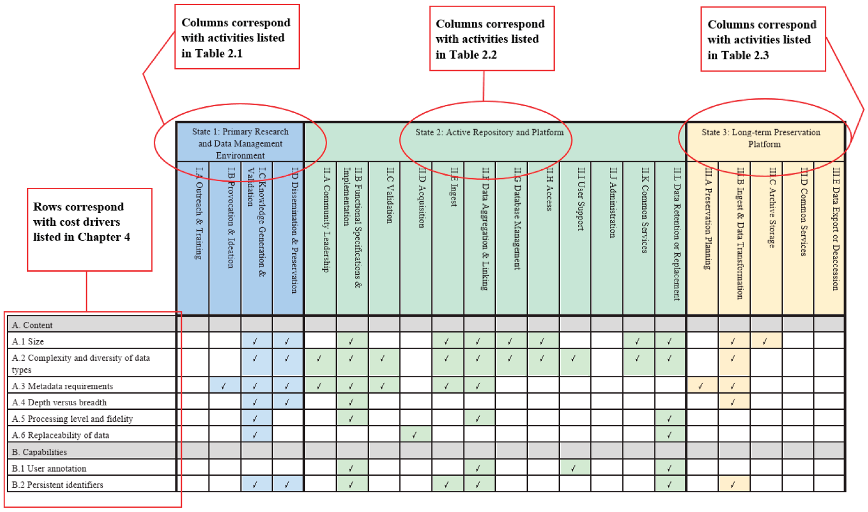
	State 1: Primary Research and Data Management Environment				State 2: Active Repository and Platform												State 3: Long-term Preservation Platform				
	I.A Outreach & Training	I.B Provocation & Ideation	I.C Knowledge Generation & Validation	I.D Dissemination & Preservation	II.A Community Leadership	II.B Functional Specifications & Implementation	II.C Validation	II.D Acquisition	II.E Ingest	II.F Data Aggregation & Linking	II.G Database Management	II.H Access	II.I User Support	II.J Administration	II.K Common Services	II.L Data Retention or Replacement	III.A Preservation Planning	III.B Ingest & Data Transformation	III.C Archive Storage	III.D Common Services	III.E Data Export or Deaccession
A. Content																					
A.1 Size			✓	✓		✓			✓	✓	✓	✓			✓	✓		✓	✓		
A.2 Complexity and diversity of data types			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓		✓			
A.3 Metadata requirements		✓	✓	✓	✓	✓	✓		✓	✓							✓	✓			
A.4 Depth versus breadth			✓	✓		✓												✓			
A.5 Processing level and fidelity			✓			✓				✓						✓					
A.6 Replaceability of data			✓					✓								✓					
B. Capabilities																					
B.1 User annotation						✓				✓			✓			✓					
B.2 Persistent identifiers			✓	✓		✓			✓	✓						✓		✓			
B.3 Citation		✓				✓			✓		✓					✓					
B.4 Search capabilities		✓				✓				✓		✓									
B.5 Data linking and merging						✓				✓						✓					
B.6 Use tracking						✓			✓	✓		✓		✓					✓		
B.7 Data analysis and visualization			✓			✓						✓									
C. Control																					
C.1 Content control					✓	✓		✓									✓				
C.2 Quality control			✓	✓		✓	✓		✓	✓								✓			
C.3 Access control			✓			✓				✓		✓									
C.4 Platform control			✓			✓									✓						
D. External Context																					
D.1 Resource replication			✓																	✓	
D.2 External information dependencies			✓			✓				✓						✓					
D.3 Distinctiveness								✓								✓					

	State 1: Primary Research and Data Management Environment				State 2: Active Repository and Platform												State 3: Long-term Preservation Platform				
	I.A Outreach & Training	I.B Provocation & Ideation	I.C Knowledge Generation & Validation	I.D Dissemination & Preservation	II.A Community Leadership	II.B Functional Specifications & Implementation	II.C Validation	II.D Acquisition	II.E Ingest	II.F Data Aggregation & Linking	II.G Database Management	II.H Access	II.I User Support	II.J Administration	II.K Common Services	II.L Data Retention or Replacement	III.A Preservation Planning	III.B Ingest & Data Transformation	III.C Archive Storage	III.D Common Services	III.E Data Export or Deaccession
E. Data Life Cycle																					
E.1 Anticipated growth						✓															
E.2 Update and versions				✓		✓				✓									✓		
E.3 Useful lifetime						✓										✓					
E.4 Offline and deep storage				✓												✓			✓		
F. Contributors and Users																					
F.1 Contributor base			✓			✓			✓	✓			✓			✓		✓			
F.2 User base and usage scenarios					✓	✓						✓	✓			✓					
F.3 Training and support requirements	✓			✓						✓			✓				✓	✓			
F.4 Outreach	✓				✓								✓				✓				
G. Availability																					
G.1 Tolerance for outages			✓			✓			✓	✓		✓			✓					✓	
G.2 Currency						✓			✓	✓		✓									
G.3 Response time						✓	✓			✓		✓	✓		✓						
G.4 Local versus remote access			✓			✓						✓									
H. Confidentiality, Ownership, and Security																					
H.1 Confidentiality		✓	✓	✓		✓				✓		✓	✓			✓		✓			
H.2 Ownership		✓		✓	✓	✓										✓					
H.3 Security		✓	✓	✓		✓			✓					✓	✓	✓		✓		✓	
I. Maintenance and Operations																					
I.1 Periodic integrity checking									✓	✓	✓					✓				✓	
I.2 Data-transfer capacity						✓			✓			✓			✓	✓		✓			
I.3 Risk management											✓			✓	✓				✓	✓	
I.4 System-reporting requirements											✓			✓	✓						
I.5 Billing and collections						✓			✓	✓		✓	✓						✓		
J. Standards, Regulatory, and Governance Concerns																					
J.1 Applicable standards		✓	✓	✓	✓	✓							✓				✓	✓			
J.2 Regulatory and legislative environment		✓		✓	✓	✓		✓		✓		✓		✓		✓	✓				✓
J.3 Governance			✓		✓			✓						✓		✓	✓				
J.4 External consultation						✓							✓		✓		✓				

BOX 4.2
Using Table 4.2

Understanding the characteristics of the data to populate an information resource and identifying the important cost drivers will help the forecaster understand the architecture ultimately needed to support the resource. Table 4.2 is a generalized matrix showing cost drivers most likely to affect short- and long-term costs related to preserving, archiving, and accessing various types of biomedical data. Figure 4.2.1 shows where information about the column and row labels are found in this report. The table might be used in multiple ways. For example, a cost forecaster working for an institution supporting a State 2 information resource may be interested in estimating the costs for one or more of the cost components listed in Box 3.2 (i.e., labor, IT infrastructure or services, media, licenses and subscriptions, facilities and utilities, outside services, travel, or institutional overhead). She might focus on labor because labor is likely to dominate operating costs. Referring to the columns in Table 4.2, the forecaster can identify the major activities often associated with State 2 active repository and platform information resources (highlighted in green in the table; more detail about the activities is found in Table 2.2), and then the likely influential cost drivers (checked boxes). This information, combined with the answers to questions related to the cost drivers posed later in this chapter (and tabulated in Appendix E) will give the cost forecaster some understanding to build a bottom-up estimate of labor needs for each of the State 2 activities. Experience with similar data resources or consultation with subject matter experts and institutional resources may inform the translation of those answers into labor hours. The forecaster can then price the labor hours by skill level.

Someone else at the institution, however, might use Table 4.2 to make decisions about system design and how to operate the State 2 information resource within a given budget. That person would focus on the rows of the table (representing cost drivers). For example, if that individual is interested in understanding how decisions related to search capabilities affect costs, he would look at row B.4 (search capabilities) and note that decisions related to search capabilities likely affect costs for three major State 2 activities (functional specification and design, data aggregation and linking, and access). Thus, he will want to estimate how different search capabilities affect the cost components identified in Box 3.2 for those activities and subactivities (as described in Table 2.2). That forecaster also will likely be challenged to locate data to inform his understanding of how, exactly, search capability choices affect cost components. He might draw on prior experience or the expertise of others.


Columns correspond with activities listed in Table 2.1
Columns correspond with activities listed in Table 2.2
Columns correspond with activities listed in Table 2.3
Rows correspond with cost drivers listed in Chapter 4

	State 1: Primary Research and Data Management Environment				State 2: Active Repository and Platform												State 3: Long-term Preservation Platform				
	I.A Outreach & Training	I.B Provocation & Ideation	I.C Knowledge Generation & Validation	I.D Dissemination & Preservation	II.A Community Leadership	II.B Functional Specifications & Implementation	II.C Validation	II.D Acquisition	II.E Ingest	II.F Data Aggregation & Linking	II.G Database Management	II.H Access	II.I User Support	II.J Administration	II.K Common Services	II.L Data Retention or Replacement	III.A Preservation Planning	III.B Ingest & Data Transformation	III.C Archive Storage	III.D Common Services	III.E Data Export or Deaccession
A. Content																					
A.1 Size			✓	✓		✓			✓	✓	✓	✓			✓	✓		✓	✓		
A.2 Complexity and diversity of data types			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓		✓			
A.3 Metadata requirements		✓	✓	✓	✓	✓	✓		✓	✓							✓	✓			
A.4 Depth versus breadth			✓	✓		✓												✓			
A.5 Processing level and fidelity			✓			✓				✓						✓					
A.6 Replaceability of data			✓					✓								✓					
B. Capabilities																					
B.1 User annotation						✓				✓			✓			✓					
B.2 Persistent identifiers			✓	✓		✓			✓	✓						✓		✓			

FIGURE 4.2.1 An excerpt of Table 4.2 showing where information about the column and row labels are found in this report. The columns list activities described in Chapter 2 in Tables 2.1, 2.2., and 2.3. The rows list cost drivers, described in more detail in the next several sections in this chapter.

A. Content

The aspects covered in this section deal with the amount, kinds, and qualities of data that a biomedical information resource is expected to host.

A.1 Size

There are at least two facets to size—overall size (e.g., volume of data in bytes) and number of identifiable items. The overall size affects media costs, time required to replicate and transfer data, and perhaps time to verify or index data. The number of identifiable items can affect the sizes of indexes and amount of metadata (e.g., the descriptive, structural, administrative, reference, or statistical information about data found in a database), as well as the time to curate the data. Identifying data at a finer granularity can help make searches more specific and might help a user avoid downloading large amounts of extraneous information.

Example decision points related to size:

1. How many files will be in a single data submission?
2. How large is an average data submission in total?
3. Are the data sizes likely to stay stable over the life of the resource?
4. What is the total amount of data expected?
5. In what kind of medium will data be captured in the short and long terms?

A.2 Complexity and Diversity of Data Types

Data in some biomedical information resources, such as The Cancer Genome Atlas,[2] were collected expressly for the resource. In such a situation, the resource managers have strong influence over the specific formats, standards, required fields, and other elements. The resulting homogeneity in the data makes them easier to process. In other situations, the data that end up in the resource are originally collected for other purposes, such as a specific research project or patient care. In that case, one expects that the data will be more heterogeneous and not necessarily conform to the conventions of the resource, thus requiring more effort to ingest and curate.

The items in an information resource might be structurally simple (e.g., deoxyribonucleic acid [DNA] sequences) or complex (e.g., patient medical histories). The resource might contain a single kind of data or several. The more data types and the greater their complexity, the greater the cost to design and maintain the storage schema, as well as the number and complexity of load scripts, quality control routines, query interfaces, and documentation. Cost-efficiently integrating multiple data types in a high-quality manner requires expertise in each of those data types.

The organization of data to be included in the resource will affect the effort required to assess whether the data should be included in the resource and the order in which they should be processed. For example, data items might be in separate files, organized into hierarchical collections, or perhaps grouped by species, chromosome number, patient identification, or phenotype. Such an arrangement can make it easier to select appropriate subsets for inclusion in the resource. In contrast, data items in a data set might all be in a single large file (e.g., the result of a backup), requiring a scan of the entire data set to extract the subsets appropriate for inclusion in the information resource.

Example decision points related to complexity and diversity of data types:

1. How complex is the underlying structure of the data?
2. How are the included data to be organized?
3. How complex is the experimental paradigm that produced the data?

[2] The website for The Cancer Genome Atlas is https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, accessed December 5, 2019.

4. What sort of additional files might be necessary to upload with the data to properly understand them?
5. How many different data types are being produced?
6. What are the relationships among these data types (e.g., are the data correlated)?

A.3 Metadata Requirements

An information repository might contain more or less metadata per identifiable item. Possible metadata elements include contributor, provenance (source and record of possession), lineage (derivation history), data uncertainty or quality, and search attributes. It may be possible to derive some of those metadata, such as summarizations, from the data themselves. Other parts could depend on external context and need to be supplied by the producers or curators of the data. Still other metadata might be in a form intended for human rather than machine understanding, in which case human effort will be necessary to interpret them. What portion of the metadata falls in each case affects costs, as the former case can be automated, while the latter two incur labor costs. Those labor costs might be borne by those that produce the data or by those who curate the information resource. The metadata (or some part of it) may need to be uploaded to a clearinghouse—a platform dedicated for publishing metadata for the purpose of discoverability.[3] Doing so will entail extra steps when adding data but will make them locatable by more people. Even small amounts of metadata can help with decisions about storage, archiving, or removal of data, thus saving long-term costs. One example is knowing if a data set represented is difficult or impossible to reobtain (i.e., "base" data)—either collected internally or obtained from an outside source—versus data that can be derived from base data. Another is whether the data have or were influenced by protected health information in any way. If the community has agreed on standards and tools, or if the targeted repository has data submission tools with standard data formats, the initial cost of managing data in the research environment may increase, but downstream costs may be reduced.

A data schema is a description of the data structure. At a minimum, the presence of a schema helps resource developers to understand the precise structure of the data they need to process. The absence of a schema is likely to require creation of special data processing software. The schema might be quite specific, such as a relational database schema, which lists the column names for each table, along with their data types, which columns must hold unique values, and other constraints, or it might include only column names or define possible structures. An Extensible Markup Language (XML) schema is an example of the latter. The schema might be held separately from the data, as in the case of relational or XML schemas, or embedded with the data in the form of, for example, column names in a spreadsheet or the header in a hierarchical data format (HDF) file. Some schemas can be supplied in machine-readable form, allowing for automated or semi-automated processing. The presence of a data schema can help biomedical-information-resource developers and managers in multiple ways. It might support an automated means to upload the data into the resource's data store or simplify scripts to do so. If data are guaranteed to conform to the schema, then less checking is required during data ingest.

The provenance of a data set is an account of how the data came to be in their current state. It might indicate who collected or generated the data, where, and when. Having such information with the data in the repository can improve trust in the data, thereby increasing the value of a biomedical information resource. The provenance could also include the processing history of the data, which might reveal biases in the data or indicate the appropriateness of particular further processing. For example, indicating if outliers were cleaned from the data might affect the suitability of certain statistical analyses of the data. The provenance could also contain parameters relevant to the collection or generation of the data, such as settings of an instrument used to collect the data, or the configuration file for a computational model. That type of information supports correct interpretation of the data by the biomedical-information-resource developers. Without such information, those managing the resource may need to reverse-engineer the values of the relevant parameters (or seek them elsewhere). If the parameters cannot be recovered, it might not be possible to determine if the data meet the conditions for inclusion in the resource, necessitating their rejection.

[3] An example of a metadata clearinghouse is the National Biological Information Infrastructure Metadata Clearinghouse hosted by the U.S. Geological Survey (see https://www.sciencebase.gov/catalog/item/4f4e48bee4b07f02db53a643).

Example decision points related to metadata requirements:

1. How much metadata must be stored with each data object to make them findable, accessible, interoperable, and reusable (FAIR)?
2. Will metadata be entered manually by the submitter/curator?
3. Will the data to be deposited include a data schema, or will one be generated?
4. Is the provenance of a data set sufficiently described, or does it need to be?
5. How much metadata can be extracted computationally?

A.4 Depth Versus Breadth

A biomedical information resource might be directed at a certain class of data (e.g., DNA sequences or cell images), regardless of the kind of study that generated them. Alternatively, the resource might target all types of data arising out of a particular domain of study or from a specific community. For example, a resource that hosts data from brain-damage studies might include functional magnetic resonance imaging images, optical images of brain slices, genomic and proteomic analyses, cognitive function tests, and clinician reports. A resource with responsibility to collect a wide range of data will likely be more expensive per data unit, as schemas, search capabilities, curation procedures, and other aspects of the resource design and management will need to be replicated on a per-data-type basis. Furthermore, as a field evolves and new experimental and computational techniques are developed, the resource potentially needs to extend its capabilities to handle data generated by those techniques. Decisions related to depth versus breadth of data will need to be informed by user needs and expectations. While it may be less expensive per data unit to have a resource focus on a single class of data, integration across multiple resources can be time intensive and costly for users. If researchers frequently use multiple data modalities, it may be more cost effective in the long term to design integration solutions early.

Example decision point related to depth versus breadth:

- Will the repository be restricted to certain data classes or types that the repository must support?

A.5 Processing Level and Fidelity

As data proceed from their raw form, through calibration, cleaning, and processing, to analysis, their volume generally shrinks. Thus, the point in this spectrum that a biomedical information resource targets for the data it collects can have a large effect on data volume. For example, storing all raw DNA reads from a sequencing run will require much more space than storing just a consensus sequence for the reads. Storing more-detailed versions of data incurs cost for the larger space and also for the effort of uploading and curating more data. However, the more-detailed data might support a larger range of uses. There are also potential space savings from storing approximate data rather than exact values. For example, DNA sequencing data can come with a quality score at each base position. Most current scoring schemes feature more than 40 different scores, which means that a quality score takes more space than the base it annotates. Replacing exact quality scores by 2-bit approximations could reduce storage by one-third.[4] Whether such an approximation is acceptable must be decided based on the intended use of the data.

Structure of the data also affects the processing level. Data intended for a biomedical information resource might be more or less explicitly structured and therefore may require different levels of processing to be incorporated into the resource. Highly structured data residing in a relational database management system, or in comma-separated-value files, may have easily discernible structures. Data might be in a semi-structured representation, such as in

[4] e.g., see Illumina, 2014, “Reducing Whole-Genome Data Storage Footprint,” Pub. No. 970-2012-013, Illumina, Inc., April 17, https://www.illumina.com/Documents/products/whitepapers/whitepaper_datacompression.pdf. Accessed May 27, 2020.

an XML data format, in which case more analysis and more complex scripts may be required to ingest the data. Data in simple text formats, or scans of text, may require more extensive processing to ingest the data.

Data intended for a biomedical resource might use a character or numeric encoding for values that differ from those used in the resource, in which case the data will need to be encoded after ingestion. A data source might also be compressed, which can have advantages and disadvantages. The smaller size of compressed data might reduce network-transfer times or intermediate storage of the raw data. On the other hand, the data will likely need to be decompressed, in whole or in part, to perform checks and manipulations at the resource.

Example decision points related to processing level and fidelity:

1. Do the raw data need to be stored?
2. Do processed data need to be stored?
3. Are there compression algorithms that can reduce the file size without compromising fidelity?
4. What kind of data structure requirements will the resource have?
5. Is the data contributor or the repository responsible for any restructuring necessary?
6. How is the data structure verified?

A.6 Replaceability of Data

A biomedical information resource might be the “official” home for certain data sets or could simply be a replica of data whose master copy resides elsewhere. Even if the resource is the official home, it might be possible to regather the data in it through repeating experiments or calculations. The cost of replacing data (or the impossibility of doing so) informs what are reasonable expenditures on redundancy and other means for loss prevention.

Example decision points related to replaceability of data:

1. Is the archive the primary steward of the data, or do copies exist elsewhere?
2. Can the data be easily recreated?

B. Capabilities

The previous section covered properties of the data themselves. This section covers aspects of a biomedical information resource that describe what information resource users are able to do with the data in the resource (i.e., without extracting the data into another environment).

B.1 User Annotation

A biomedical information resource might support comments, annotations, and corrections on data items beyond those originally submitted by the contributors of the data. If so, there is a cost for developing such a facility and for overseeing its appropriate use. Human or machine interventions in a data resource also need to be documented, authenticated, and retained as part of the metadata.

Example decision points related to user annotation:

1. Will the repository have to provide user annotation capabilities?
2. What is the nature of these annotations?
3. Are they provided by humans or machines, and how will they be authenticated?
4. Are permissions required to annotate the data?

B.2 Persistent Identifiers

A biomedical information resource might want or have to support persistent identifiers (PIDs) for data sets or data items, such as Digital Object Identifiers (DOIs).[5] The host of the resource may have to pay directly for the ability to assign such identifiers or indirectly in participation in an organization that has that capability. Support of such identifiers also carries a requirement to maintain a mapping from identifiers to data entities even as those entities are modified or moved.

Example decision points related to persistent identifiers:

1. What PID scheme will be used by the archive?
2. Is there a cost associated with using the PID?
3. How many objects need to be identified?
4. Who will be responsible for keeping the PIDs resolvable?

B.3. Citation

A biomedical information resource might support citation of data items (or sets of items) at a granularity smaller than entire data sets. If so, there might need to be a facility that, given an item or sets of items, generates a citation and, conversely, given a citation, locates the corresponding items.

Example decision points related to citation:

1. Will users be able to create arbitrary subsets of data files and mint a PID for citation?
2. Will the repository provide machine-readable metadata for supporting data citation?
3. Will the repository provide export of data citations for use in reference managers?

B.4 Search Capabilities

Efficient search for items in a biomedical information resource, beyond simple look-up by a reference or access number, usually requires construction and maintenance of indexes over the data. Those indexes take space beyond what is required for the base data, plus time to construct and maintain. If there are many indexes, they can slow updating of the base data. Different kinds of searches require different kinds of indexes. For example, an index that supports look-up of items by exact match to a value might not support searching on a range of values. Full-text searches and approximate matching (such as for genomic sequences) require specialized index structures to execute efficiently.

Example decision points related to search capabilities:

1. Will the repository provide a search capability for data sets?
2. How much of the metadata will be included in search?
3. How complex are the queries that will be supported?
4. What types of features for search will be provided?
5. Will the repository deploy services to search the data directly?

[5] The website for the DOI System is https://www.doi.org/, accessed December 4, 2019.

B.5 Data Linking and Merging

A biomedical information resource might supply users with the ability to navigate from a data item to related items, such as from a DNA sequence region to a protein coded by a region of that sequence, or from a medication to a list of clinical studies for that medication. The resource might combine information from multiple contributors into a single entry, such as a functional annotation from one contributor on a gene sequence from a different contributor. Such capabilities mean that the resource will need to create appropriate links and perform merging operations when new data are added.[6] Supporting such data interoperability is aided by the use of standards (e.g., ontologies and common data elements) but can be time consuming and expensive depending on the complexity of the data and their initial level of compliance.

Example decision points related to data linking and merging:

1. Will the data require/benefit from linkages to other related items?
2. Will the resource provide the ability to combine data across records based on common entities/standards?

B.6 Use Tracking

A biomedical information resource might track uploads, access, and downloads of data items to inform contributors and resource operators about their use. Statistics of such operations may incentivize researchers to contribute data by providing evidence of data use. They could inform life-cycle decisions such as when data could transition to another state, and they might be used to assess the long-term value of the data. In addition, the information could support billing and cost recovery. Tracking would likely have a minor effect on overall costs of operating the resource.

Example decision points related to use tracking:

1. Will the resource provide the ability to track uploads, views, and downloads?
2. If so, and if made available to users, how will this information be made available?
3. Will the resource track data citations to its data?

B.7 Data Analysis and Visualization

Generally, users of a biomedical information resource want to do more than view data. They might want to reformat data for use with a particular tool, visualize them in the context of other data, and conduct statistical analyses upon them, for example. The resource might provide services to perform such operations locally on the data. There might even be a requirement to provide general-purpose co-located computing, where a user can run arbitrary programs on the data. Providing such capabilities will incur costs for provisioning the computing cycles to support those operations. However, such costs might be offset by reductions in costs to resource operators by avoiding bulk downloads by users to perform computations locally. In considering costs to the wider scientific enterprise, supporting data analysis at the resource could avoid user costs of downloading and maintaining local copies, and could increase the value of the data by expanding the audience of researchers who could work with them. The operators of a resource might employ a credit- or token-based system to limit and track the use of computational resources. Such an approach can help control computation costs, although there will be costs associated with implementing and administering such a mechanism.

[6] In this case, the committee is talking about linking and merging at the resource rather than before deposit.

Example decision points related to data analysis and visualization:

1. What types of data analyses and visualizations will the repository support?
2. What types of other data operations will the repository support (e.g., file conversions, sequence comparison)?
3. Do these services require significant computational resources?
4. Who will pay for computational resources?

C. Control

This section covers aspects of a biomedical information resource that deal with control and oversight of the resource.

C.1 Content Control

A biomedical information resource can be more permissive or more restrictive in what it chooses to include in its contents. At the permissive end, a resource might allow open posting of any data of the appropriate type. At the other end, there may be a review process to determine whether submissions are to be included in the resource. That review could be minimal—for example, an automated check that the submission is properly formatted—or more extensive—for example, a human review of metadata to assess suitability for inclusion. A more intensive review process increases labor costs but provides a means to limit the amount of data that is hosted.

Example decision points related to content control:

1. Will all appropriate data be accepted, or will there be a review process?
2. Will the review process be automated, or will it require human oversight?

C.2 Quality Control

A biomedical information resource may exercise more or less rigorous control on the quality of the information within it. At one extreme, it might leave all quality control to be the responsibility of the data contributors. At the other, it might manually or automatically vet all incoming data to detect quality issues. There could also be quality assessments on derived products that are generated internally to the resource. More intensive quality control incurs higher costs, but it can increase the value of the resource for scientific and clinical use. Problems with quality, when encountered, entail increased review and (where possible) repair of data by someone. If repair is not possible, the value of the resource may be compromised. Quality-related properties that need to be verified are correctness, completeness, currency, and duplication.

Correctness is related to how accurate the data are; values in input data may be inaccurate for a variety of reasons (e.g., errors in data processing or transcription, noise in the instrumentation or method used to collect them, mislabeling of samples). There might be internal inconsistencies in the data or incompatibilities with external sources. Cross-links between records might be incorrect. The greater number of types and instances of correctness problems with the input data, the more effort is required by resource managers to address them. To the extent that such problems are not identified or addressed, the value of the resource can be compromised.

Validating the completeness of data means identifying missing items at the record or field level. Such gaps can entail costs for the resource because of added complexity in the processing to ingest the data and possible additional complexity in the data representations in the resource to cope with missing elements—for example, special flags for missing values.

The time sensitivity of data can impact the currency of data. Some kinds of biomedical data are relatively time insensitive (e.g., the amino-acid sequence of a particular protein). However, other data may have more value the more current they are (e.g., disease incidence). In that case, badly out-of-date data may limit the value of a biomedical information resource.

Duplicated information within or between data sets will require more review or processing to remove duplication. A particular instance of duplication is when a contributed data set is revised periodically with corrections and additions. If those changes are not explicitly flagged, then the managers of the resource will need to compare each new version of the data set with previously submitted versions to avoid loading duplicate data.

The biomedical information resource may promulgate guidelines or validation routines that indicate issues with data or conformance with resource expectations and standards. While prevalidation of data quality by the data contributor shifts costs to the data contributor, it could result in lower overall effort, as the providers may be able to incorporate checks into their normal data processing practices. Also, detecting a problem with the data on the provider side avoids back-and-forth communication between contributor and resource managers to point out problems and get corrected data.

Example decision points related to quality control:

1. What quality control process will the repository support?
2. Will these be automated or require human oversight?
3. What level of data correctness will be required, and how will it be validated?
4. What gaps in the data at the record or field level will be tolerable?
5. Will any of the data be time sensitive, and how will data currency be ensured?
6. How will duplication within or between data sets be addressed?
7. Will prevalidation guidelines or routines be distributed by the resource to the data contributors?
8. Will human curation be necessary?

C.3 Access Control

A biomedical information resource might place restrictions on which users can see which data—for example, if data are embargoed from general release for a certain length of time or the resource might provide private workspaces for individual users or groups. The data may also be consented for particular uses, in which case consent information will need to be linked to particular data items and consulted when deciding access permissions. Such control means having a mechanism to identify users (authentication) and to track which are allowed access to what data (authorization). This capability adds costs both for managing user identifications and for developing access-control mechanisms. Supporting collaborative workspaces, blind review, and mandatory release schedules all complicate those mechanisms.

Example decision points related to access control:

1. What types of access control are required for the repository (e.g., will there be an embargo period)?
2. At what level are they instituted (e.g., individual users, individual data sets)?
3. Does use of the data require approval by a data access committee?

C.4 Platform Control

There might be limitations on what computing platforms are allowed for running a biomedical information resource. Third-party hosting (e.g., commercial cloud providers) might be prohibited, permitted, or required, or there may be restrictions on hosting or mirroring data overseas (e.g., if overseas data privacy laws regarding human-subjects research data may not be aligned with domestic policies). Such restrictions constrain implementation alternatives, which in turn can influence costs.

Example decision point related to platform control:

- Are there restrictions on the type of platform that may or must be used?

D. External Context

This section considers the context of a biomedical information resource in relationship to other, external resources.

D.1 Resource Replication

There might be a requirement to replicate a biomedical information resource at other sites, with other groups operating "mirror" versions of the resource. Mirroring might be required, for example, to provide more convenient access to collaborators at a distant location. If there is such a requirement, then the original site will need to coordinate software updates and data releases with the mirror sites.

Example decision point related to resource replication:

- Is there a requirement to replicate the information resource at multiple sites (i.e., mirroring)?

D.2 External Information Dependencies

A biomedical information resource might have dependencies on other information sources. For example, a resource containing DNA sequences for an organism might depend on a reference sequence to provide position numbers to locate those samples. In another example, metadata records might require certain fields to come from a controlled vocabulary, such as the Medical Subject Headings (MeSH).[7] Such a dependence might engender maintenance costs when the external source is updated.

Example decision point related to external information dependencies:

- Will the resource be dependent on information maintained by an outside source?

D.3 Distinctiveness

There might be other biomedical information resources with similar content and that support some of the same tasks. In such a case, it might be that such information resources can substitute for the resource in question, albeit with some "degradation" in result. The type and amount of such degradation can help calibrate the soft cost of risk of loss (see Appendix D).

Example decision point related to distinctiveness:

- Are there existing resources available that provide similar types of data and services?

E. Data Life Cycle

This section deals with aspects of a biomedical information resource that concern how it is expected to evolve over time.

E.1 Anticipated Growth

The ultimate size of a biomedical information resource, and the rate at which it grows to reach that size, influence annual maintenance and expansion costs. Will the resource reach "maturity," where no new data are

[7] The website for the MeSH is https://www.nlm.nih.gov/mesh/meshhome.html, accessed December 12, 2019.

expected because of the end of a project or program that supplies the data, or is it expected to continue growing throughout the lifetime of the resource?

Example decision points related to anticipated growth:

1. Is the repository expected to continuously grow over its lifetime?
2. Is the likely rate of growth in data and services known?
3. Is the use of the repository likely to grow over time?
4. Is the likely growth of the user base known?

E.2 Update and Versions

The frequency of updates and the need to retain past versions for a biomedical information resource affect operating costs. Some resources provide periodic releases, which batch updates and apply them all at once, whereas other resources are revised incrementally as updates come in. In the case of the periodic-release model, past releases (versions) might be maintained, for example, to support replicability of a study that used a particular release. Retaining past versions obviously incurs storage costs over just providing the most recent version, and decisions need to be made about if and how prior versions will be made available. In the case of the incremental model, the frequency of update might be a cost driver, if updates entail manual review or curation activities.

Example decision points related to updates and versions:

1. Will the deposited data require updates (e.g., in response to new data or error corrections)?
2. Will prior versions of the data need to be retained and made available locally or in a different resource?
3. How frequently will individual data sets be updated?

E.3 Useful Lifetime

Some data in a biomedical information resource might have a limited period of usefulness, or their utility might decline with time. For example, a collection of cell images might be superseded by later images from a higher-resolution technology. Deaccessioning or archiving such data will reduce operating costs. If there is a predictable end date for a resource as a whole, that knowledge is useful in predicting lifetime costs. Useful lifetime of data can be difficult to predict because even data collected decades ago can still be used for analysis if properly documented (see Box 4.3).

Example decision points related to useful lifetime:

1. Are the data to be housed likely to have a limited period of usefulness?
2. Does the resource have a defined period of time for which it will operate?
3. Does the resource have to provide a guarantee that the data will be available for a finite period of time (e.g., 10 years)?

E.4 Offline and Deep Storage

If it is possible that some data in a biomedical information resource do not need to be available online but still need to be retained, then they might be migrated to a less expensive form of storage. This report distinguishes between offline and deep storage. Data in offline storage can be brought back online in the resource, albeit with some delay. Data in deep storage (typically State 3 data) are not intended to be brought back online in the same resource. Rather, they are preserved in the event that someone wants to "rehydrate" them in the future, either for individual use or as part of another information resource. In the case of offline storage, there can be costs for

BOX 4.3
Examples of Long-Lived Data

Deciding how long to keep the data (i.e., the useful life span of a data set) is difficult. According to Russ Poldrack (personal communication, April 15, 2019), imaging studies from the early 2000s are still used (e.g., as part of meta-analysis), even though the resolution is inferior to current data. Because neuroimaging early on invested in an open file format (the Neuroimaging Informatics Technology Initiative[a]), and because there is a dominant commercial file format (Digital Imaging and Communications in Medicine [DICOM][b]), data from the 1990s can still be read. In the 2000s, NIH established a pediatric neuroimaging database comprising several hundred data sets from normal children, the NIH Pediatric MRI Data Repository.[c] Although the database infrastructure is no longer independently maintained, the data sets are served through the NIMH NDA and have been ingested into additional infrastructures (e.g., the Montreal Neurological Institute[d]). These data are still being used for research purposes in 2019.[e]

[a] The website for the Neuroimaging Informatics Technology Initiative is https://nifti.nimh.nih.gov/, accessed December 12, 2019.

[b] The website for DICOM is https://www.dicomstandard.org/, accessed December 12, 2019.

[c] The website for the NIH Pediatric MRI Data Repository is https://nda.nih.gov/edit_collection.html?id=1151, accessed December 9, 2019.

[d] The website for the Montreal Neurological Institute is https://www.mcgill.ca/neuro/, accessed December 12, 2019.

[e] Some of these resources are listed here: https://scholar.google.com/scholar?as_ylo=2019&q=NIHPD&hl=en&as_sdt=0,5, accessed October 16, 2019.

offline data beyond basic storage costs, regardless of who is managing the offline storage.[8] For example, several commercial cloud platforms have lower-cost archival storage services or tiers that assume only a small fraction will be accessed during any period. Access in excess of that fraction incurs additional cost.

Rehydration costs can appropriately be ascribed to future users of data that are in deep storage.

Example decision points related to offline and deep storage:

1. Can the resource take advantage of offline storage for data that are not heavily used?
2. Does the resource have a plan for moving unused data to deep storage (i.e., State 3)?

F. Contributors and Users

This section covers aspects of a biomedical information resource associated with user characteristics and numbers that might influence costs.

F.1 Contributor Base

The number of individuals or sources that generate information to be hosted can affect development and operating costs for a biomedical information resource. If data originate from the same source (e.g., a single instrument, as with sky surveys and particle-physics experiments or a single organization or community), then less effort is required to coordinate with data providers than in situations where data originate from many communities and organizations. In addition, if all contributors are internal to the same organization that hosts the information

[8] Note that cloud providers of archival storage are known to use tape to support that storage (e.g., Lantz, 2018).

resource, good compliance with data formats and standards may be more likely and costs for review and curation might be less. Alternatively, data originating from a wide range of individual autonomous investigators spread across multiple disciplines may require more interactions between the resource managers and those investigators to collect them, and there will likely be more variation in the data to address.

How data are transferred to the resource will also affect costs. They may arrive periodically in large batches or incrementally in smaller amounts. In the extreme case, data might stream in continuously—for example, directly from wearable devices. Data transfer can be initiated by the data contributor (data push) or by those managing the resource (data pull). To the extent that data ingestion by the information resource has a manual component, more frequent arrivals means the more often a person has to manage that task. In the continuous case, automated processing will probably be a necessity, which will entail development costs.

The time and network resources required to transfer a data set from a provider to a data resource scales (although not necessarily linearly) with the size of a data set. However, in some cases a data set may be so large that network transfer is not feasible (or will not be complete in the required time frame). In such cases, physical transfer of storage media may be needed, which entails costs for purchasing the media, loading them, shipping them, and extracting the data at the resource.

Also affecting the cost of integrating contributed data into a resource are, for example, whether there are direct charges (e.g., purchasing costs, licensing fees) or indirect charges (e.g., membership fees to access) associated with acquiring the data and whether the data contributor is willing to be responsible for the data and serve as their steward. The steward is the point of contact regarding the data who responds to questions about them, addresses errors or other problems associated with them, and tracks their current location(s). Data can be effectively "orphaned" if the data collector is no longer affiliated with the organization where the data were collected. Trying to locate, obtain, and understand data with no identifiable steward can require significant effort.

Example decision points related to the contributor base:

1. Is the number of contributors known? If not, can it be estimated?
2. Are all the data originating from the same source (e.g., a single instrument or a single organization)?
3. How will data be transferred into the data resource (e.g., periodic large batches, more frequent smaller data sets, constantly streamed, by physical transfer)?
4. Will the data be pushed by the contributor or pulled by the resource?
5. Are there direct or indirect fees associated with acquiring the data from a source?
6. Will a data steward be available from among the contributors to assist with any data integration into the data resource?

F.2 User Base and Usage Scenarios

The number of people accessing an information resource and the frequency and kinds of access can all influence costs for a biomedical information resource. A resource that serves an entire research community will likely see much more use than, say, an internal project repository for a single research group. While actual storage costs will probably not depend on the number of users (unless data must be replicated to serve high access rates), computation and network costs will rise with increased use. The kinds of access can also affect those costs. For example, retrieving single items will require less network bandwidth than a bulk download of a whole data set.

Example decision points related to the user base and usage scenarios:

1. How many users will likely access the data?
2. What will be the frequency of access?
3. How will users access the data?
4. Will the resource be building analysis tools?
5. Will the resource support individual file download or bulk download?

6. Will there be any fees for downloading/accessing the data?
7. How many different types of users must be supported?

F.3 Training and Support Requirements

There may be expectations, or it may be found beneficial, that operators of a biomedical information resource provide training for resource users. That training could be more or less labor intensive and involve conducting tutorials, preparing training materials, or maintaining help pages on a website. A "help desk" function might be required that provides either live consultation or message-based responses, both of which require training and staffing. On the other hand, investing in training of and consulting with users may result in easier data integration and lower future data-collection and curation costs.

Example decision points related to training and support requirements:

1. Will training for resource use be offered?
2. What form will the training take?
3. Will a "help desk" be provided?
4. When does live help need to be available?
5. What is the expected skill level of the user base?

F.4 Outreach

If the existence and features of a biomedical information resource need to be publicized, then there may be associated labor, travel, and media costs for preparing articles, giving conference presentations, producing newsletters, conducting print or e-mail campaigns, reaching out on social media, and so forth. In some communities, there may be reticence or even resistance toward using shared information resources, which might require extensive outreach efforts to overcome.

Example decision points related to outreach:

1. Does the existence of the repository need to be advertised?
2. How many conferences per year should resource representatives attend?
3. Will the resource have a booth at the conference for live demos or conduct hands-on tutorials?
4. Are users required by funders or journals to deposit data in the repository?

G. Availability

The aspects in this section relate to expectations about how available the data in a biomedical information resource will be. Data availability encompasses the reliability of the resource hosting the data, how quickly new data appear, how fast requests for data are serviced, and from where the data can be accessed.

G.1 Tolerance for Outages

Different biomedical information resources have different tolerances for system outages. While an outage of hours or days might be tolerable for a resource that supports, for example, retrospective analysis, a similar loss of availability might be highly undesirable for a resource that is used continuously every day in support of clinical decision making. A low tolerance for outages often entails data replication, which will incur storage and other costs, possibly including the cost of network bandwidth for transferring data to a backup site. Guarantees of high availability (for example, 99 percent up-time) require support staff to be on call around the clock, which is a large labor expense.

Example decision points related to tolerance for outages:

1. What is the tolerance for outages of the resource?
2. What measures will be taken to avoid and mitigate outages?
3. How quickly and completely does the resource need to recover from an outage?

G.2 Currency

Data submitted to a biomedical information resource may need to be available to users within a fixed time frame. It might be acceptable that new items appear in data resources in monthly or quarterly releases, but other types of data resources may need to be updated daily (e.g., outputs for flu forecasting models). This requirement might affect labor costs, since in the latter case there is no opportunity to amortize effort over all items in an update batch.

Example decision points related to currency:

1. How often will the data be released?
2. How soon do data need to be made available after they are received?

G.3 Response Time

A biomedical information resource may have a target or requirement for how quickly requests are serviced, either by a computer system or by a human agent. Interactive response times (a few seconds) might require replicating the data and additional computing services, so that multiple requests for popular data can be handled at the same time. Interactive response times also limit what data can be held in lower-cost near-line or offline storage. There can also be a human element in response time, such as approvals for access or review of submitted data sets. In general, lower response times correlate with higher labor costs.

Example decision points related to response time:

1. Are there requirements for response time for service?
2. Are there requirements for responses from humans?

G.4 Local Versus Remote Access

While most biomedical information resources of which the committee is aware support remote access over the Internet, there are examples in other domains (e.g., film archives, defense-personnel information) where users must physically come to the resource to access it. Such a scenario generates space, staffing, and equipment costs for hosting users. Some resources may be accessed remotely over a network, but large data transfers may still entail shipping physical media (e.g., tapes, disks), which incurs preparation and shipping costs.

Example decision points related to remote access:

1. Does the resource require that any data be shipped via physical media?
2. Will the resource be built using commercial clouds?
3. Do users have to travel to the resource to use the data?

H. Confidentiality, Ownership, and Security

This section covers aspects of a biomedical information resource related to protecting the data and the rights of those associated with the data. These issues are complex subjects and warrant more attention than can be given in this report, but the questions provided here will allow the cost forecaster to identify the relevant cost drivers.

H.1 Confidentiality

A biomedical information resource may need to protect the confidentiality of the data it holds, because those data contain either personally identifiable information or sensitive intellectual property. In the case of personally identifiable information, there may be a need to deidentify information or restrict access. There may be additional requirements to track and audit use. If so, credentials and permissions will need to be assigned to users; systems, analytical output, and space to maintain use records will be required; and added system complexity for tracking access to and use of items will be necessary. All of these items entail added costs. Inclusion of machine-actionable metadata that capture restrictions on use could reduce cost.

Example decision points related to confidentiality:

1. Will any of the data require special protections?
2. Will any of the data have embargo periods or embargo-related limitations that may entail costs?
3. Are there any audit requirements for who has accessed or downloaded the data?

H.2 Ownership

If the data have been managed by a variety of entities (e.g., companies, laboratories, public repositories, or individual investigators and their staff), different custodians may have spent more or less time to locate and appropriately format data for forwarding to the resource, even given data-sharing requirements. Their data release processes might also be cumbersome for those wanting to use their data. In contrast, some data are maintained on behalf of patient collectives or disease organizations that actively promote and facilitate their use, possibly making such data easier and less expensive to use.

If a biomedical information resource contains proprietary information, then there may be requirements to track ownership of particular data sets and to ensure that data use conforms with any licensing conditions and to the preferences of the participants from which it came. Support for tracking and conforming use may have costs beyond those paid to license the data sets.

Hospitals or clinics can have specific release forms for allowing the transfer of patient data. Collecting patient data from a large number of such establishments might mean obtaining and executing a different release form for each patient—a time-consuming and labor-intensive process. If these release forms only consent to certain uses, then use must be audited and tracked, which may incur additional costs. Inclusion of machine-actionable metadata that capture any ownership characteristics could reduce cost.

Example decision points related to ownership:

1. If data are contributed from multiple sources, will there be a need to process multiple kinds of release forms?
2. Will all the data be released by the data resource under the same license, or will different permissions be assigned to different data sets?
3. Will data submission agreements be necessary?

H.3 Security

Similar to confidentiality, security for a biomedical information resource implies preventing unauthorized access, but it also implies protection against loss or corruption of its data, intentional or otherwise. Measures taken by a biomedical information resource likely include internal or external security audits, special operator and user training, active monitoring of the resource, applying security patches expediently, or using specially protected computing, storage, and networking platforms, all of which incur costs.

Security might also encompass offering services such as ensuring a resource complies with Health Insurance Portability and Accountability Act and Health Information Technology for Economic and Clinical Health Act requirements. Sensitive data produced by or under the auspices of federal agencies have distinct security requirements. For example, if a Federal Information Security Act (FISMA)-certified environment is necessary to comply with the National Institute of Standards and Technology regulation associated with protecting controlled unclassified information in nonfederal systems and organizations (Ross et al., 2020), additional costs will be entailed. At a minimum there are costs to documenting FISMA compliance. Those costs increase if it is determined that a higher level of FISMA certification is required. If the data are kept in a cloud environment, certification associated with the Federal Risk and Authorization Management Plan (FedRAMP)[9] may also be necessary, entailing additional costs.

Example decision points related to security:

1. What measures need to be taken to ensure the integrity and availability of the data?
2. Do these measures require using protected computing, storage, or networking platforms?

I. Maintenance and Operations

This section covers aspects of a biomedical information resource related to obligations for maintenance and operation of the resource.

I.1 Periodic Integrity Checking

As part of ongoing maintenance, operators of a biomedical information resource will need to assess the integrity of its hardware, software, and data. The frequency and detail of such assessments will affect operating costs. Processes put in place will flow from an understanding of error and failure rates and the tolerance for data corruption and loss.

Example decision points related to periodic integrity checking:

1. What processes will be put in place for checking the integrity of the hardware, software, and data?
2. How frequently will these checks be performed?

I.2 Data-Transfer Capacity

Insufficient data-transfer capacity of the facility that hosts a biomedical resource can place constraints on the operation of the resource. For example, limited connectivity can constrain the amount of data that can be downloaded from a resource, the ability to replicate the contents, or the ability to perform off-site backups.

Example decision point related to data-transfer capacity:

- Will the bandwidth available to the resource be sufficient for the data sizes and rates required?

[9] The website for FedRAMP is https://www.fedramp.gov/, accessed April 6, 2020.

I.3 Risk Management

With any biomedical information resource, there is a risk of corruption or loss of content. Who assumes that risk (and hence must take steps to ameliorate it) will influence where certain costs fall. If a resource is a data-sharing portal but not the repository of record for the data it holds, then the risk may fall largely on the contributors, who will bear the cost of maintaining backup copies of their data elsewhere. If, on the other hand, the resource is the "official" repository for the data it holds, the operators of the resource will be responsible for risk-mitigation measures in line with the perceived value of the data and hence bear the concomitant costs. For sensitive information, there is also risk of leakage (i.e., unauthorized export of data to external recipients), either through unintentional or malicious action. Even if an information resource is not the repository of record for its data, it must bear the costs of mitigating this type of risk. In addition, a response plan might be necessary to address circumstances (e.g., unexpected loss of funding or dissolution of the organization hosting the resource) that force the early termination of the information resource.

Example decision points related to risk management:

1. Will the repository be solely responsible for risk mitigation?
2. Is a response plan for unexpected termination required?

I.4 System-Reporting Requirements

The overseers and operators of a biomedical information resource may require regular reports on the status of the system, in terms of both content (e.g., number of items, storage space used) and computer-resource usage (e.g., central processing unit hours, network usage). Setting up such reports will likely be a one-time cost, with perhaps a small amount of recurring labor cost if the reports have to be invoked manually.

Example decision point related to system-reporting requirements:

- What types of system reporting will the resource be required to do?

I.5 Billing and Collections

If the biomedical information resource charges for upload, access, and download of data, then there will need to be an operational function responsible for billing for and collection of those charges.

Example decision point related to billing and collections:

- Will there be charges for use of the resource?

J. Standards, Regulatory, and Governance Concerns

Standards for the interchange of biomedical data, for the description of various kinds of biomedical data objects, and for other data practices are important enablers for the research data ecosystem. This section considers community conventions, rules, policies, laws, and stakeholder concerns with which the operators of a biomedical information resource may have or want to comply.

J.1 Applicable Standards

A biomedical information resource might have to conform to one or more standards for the content and format of the data hosted. Some standards are created and maintained by formal national or international standards-development organizations; in other cases, they are developed and managed by ad hoc community mechanisms, particularly when the scale of the community and the scope of the standard's application are limited. Where well-established and domain-based standards exist, and especially where tools also exist that automate the use of those standards at the time the data are generated, conformance to them may significantly lower the cost of later data ingest, curation, dissemination, and preservation. If tools do not yet exist, software routines to parse and extract needed data from the proprietary structures will likely need to be developed, possibly at significant expense. Even if the data as a whole do not match a standard, particular fields might be standardized—say, taken from a controlled vocabulary or a reference list of codes—in a way that matches the assumptions of the resource, thus avoiding the overhead associated with converting those fields. Highly structured data may also be more or less difficult to process, depending on whether they conform to a widely used standard (e.g., FASTQ file format for genomic-sequence data [see, e.g., Cock et al., 2010] or HDF for array data[10]) or are in some system-specific format (e.g., for a particular manufacturer's microscope or as used by a particular medical-records systems).

If no standards exist, then they may need to be developed to increase the quality of the data sets and the efficiency of data ingest and use. This process can be costly up front, as it involves bringing together groups of experts, often repeatedly over time, to achieve and document consensus (see Box 4.4). Formal national and international standards development is slow, expensive, and highly structured: there are complex bureaucratic and procedural issues and elaborate governance formalities. Typically, most of the work is volunteer labor by experts, perhaps facilitated by a paid editor. Funding of formal standards bodies is beyond the scope of this report, but it is worth noting that research grants have been used to good effect to accelerate creating new less-formal community standards.

A key aspect is whether the standard will evolve during the lifetime of the resource and whether the resource must conform to updated versions of the standard. If so, there will be associated costs for modifying the resource and possibly for restructuring or augmenting existing data holdings. A particularly challenging case is one in which a resource developed prior to the development of standards must be "retrofitted" to accommodate a standard that later emerges. The development of different national standards (e.g., the European Union's General Data Protection Regulation)[11] is another version of this challenge. Likewise, there could be standards-related costs associated with transforming data and metadata from one data state to another.

Example decision points related to applicable standards:

1. How many different standards will the resource have to support?
2. Do these standards exist?
   a. If not, is the resource expected to lead their development?
   b. What is the plan for accepting data while standards are in development?
   c. If so, are the standards mature (i.e., how much are they expected to evolve)?
3. Are the data validators and converters available for the standards, or do they have to be developed?
4. What is the plan for "retrofitting" data that have been uploaded without the standards in place?
5. How frequently will the standards update?
6. Do the standards require spatial transformations (e.g., will they need to be aligned to a common coordinate system)?
7. How many file formats will be supported?
8. Is there an open file format available?

J.2 Regulatory and Legislative Environment

A biomedical information resource may be bound by laws and government regulations, particularly if it maintains information on individuals. Those requirements may entail additional record keeping or notification of

[10] See the HDF Group at https://portal.hdfgroup.org/display/HDF5/Introduction+to+HDF5, accessed on May 12, 2020.

[11] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119), 1-88.

BOX 4.4
The Development and Evolution of Standards

The development and acceptance of a standard can take a significant amount of time and resources. As an illustration, Neurodata Without Borders (NWB)[a] is a file format developed for describing neurophysiological data (Teeters et al., 2015). NWB was initiated in 2014, but the first version was deemed insufficiently developed to be deployed widely (Rübel et al., 2019). In response, the standard was extensively revised and released in January 2019 (Rübel et al., 2019). Rübel and others (2019) explained, "It became clear that in order to achieve [the goals of the project] we needed an advanced software architecture, a well-articulated data standards ecosystem, an open community software strategy, and advancements to the NWB:N data standard itself." In response, a working group to develop NWB 2.0 was launched in 2017, and NWB 2.0 was released in January 2019. However, it has not yet been stress-tested through implementations in the community, which may lead to another round of feedback. The challenge for information resources is to understand when a standard is sufficiently stable that it is "safe" to invest in it as the basis of a software ecosystem. Standards organizations that undertake formal review of standards against a set of criteria and can help with the process of evolution of standards can be helpful here. Biomedicine is seeing some of these types of organizations emerge. For example, the International Neuroinformatics Coordinating Facility (INCF),[b] an international organization devoted to issues surrounding data sharing in neuroscience, has recently launched a standards review-and-endorsement process to help ensure that neuroscience is served by a set of harmonized standards; the Computational Modeling in Biology Network (COMBINE) initiative[c] is providing a similar platform for computational biology.

[a] The website for NWB is https://www.nwb.org/, accessed December 4, 2019.
[b] The website for the INCF is https://www.incf.org/, accessed December 4, 2019.
[c] The website for the COMBINE initiative is http://co.mbine.org/, accessed December 4, 2019.

those about whom information is maintained. The resource might be covered by an open-records act if it is maintained by a government agency. Obtaining compliance certification could involve costs associated with precisely documenting policies and procedures.

Example decision points related to the regulatory and legislative environment:

1. What laws and regulations cover the data and operation of the resource?
2. Is the resource covered by an open-records act?

J.3 Governance

A biomedical information resource may have a policy-setting body for itself or as part of a larger organization. Policies may be set either initially or on an ongoing basis. Having such a governing body incurs some personnel costs to engage with it and possibly for convening it. Following the guidance and directives of the governing body may entail changes or extensions to the resource, which will likely come with costs.

Example decision points related to governance:

1. Does the resource need to maintain an external advisory board?
2. Does the resource set policy for itself, or is it part of a larger organization?

J.4 External Consultation

The developers of a biomedical information resource may need or want to consult with external stakeholders—data contributors, potential users, and funding agencies—about the initial resource design and about later updates. Such consultation can extend the timeline for resource development and maintenance, which might increase some costs, such as labor.

Example decision points related to external consultation:

1. Will external stakeholders be consulted for initial design?
2. Will external stakeholders be consulted on an ongoing basis?

ATTACHING DOLLARS TO THE COST FORECAST

The committee applied its cost-forecasting framework to scenarios that exemplify numerous decisions about the treatment of data at different points in the data life cycle and in their various data states (see Box 2.1). The committee did not attempt to quantify the costs of data as an investigator or a data platform manager would do because there are too many variables related to the myriad of factors unique to the data, the institutional platform host, and the data contributors and user communities to be able to come up with meaningful numbers. However, now that the investigator or data platform manager has considered where the data are coming from, and how they will be used, it will be necessary to begin to quantify the costs.

Forecasting for a State 1 (Primary Research) Resource

A State 1 (i.e., the primary research environment) researcher creating a cost forecast will necessarily focus on costs that must be budgeted. Additional costs may be reported in tabulations for the public record (e.g., the "cost" of a discovery), sunk costs (e.g., equipment from the researcher's prior research project), and costs borne by the host institution that are not reimbursed in paid overhead rates (e.g., that underwrite subsidies for IT services). The researcher will use the prices paid for services in the cost forecast, whether those service costs are less than what is being paid (e.g., to support institutional "taxes"), or if they actually cost more (e.g., those subsidized by her institution).

The State 1 researcher may face many choices throughout the course of the research, with alternatives that imply different time profiles for costs. One course of action may entail low up-front investment but higher long-term operating costs, another the reverse. While the researcher may be bound by rules of the financing entity in selecting the preferred alternative, good decision making argues that the present discounted value of the alternatives be calculated, with the economically preferred alternative having the lowest present discounted value. However, if the funding entity pays expenses for only an arbitrary fixed number of years, it may be reluctant to pay the immediate cost of an up-front investment that cuts long-term operating costs, even if a present discounted value calculation would argue it should do so. This sort of shortsightedness is not confined to physical investments. It might involve a situation in which the individual researcher may not value what the larger community needs, such as the additional costs of preparing data to meet a long-term repository standard. In this case, both costs and benefits differ, requiring that both be weighed in making a decision. Although the funding entity may not sympathize with taking the long view, the researcher needs to understand what might be the better course of action in designing a long-run strategy.

The State 1 researcher can use the data set characteristics, activities, and cost drivers described in this chapter and in the template in Appendix E. Many of the activities and cost drivers in the template in Appendix E may not be directly applicable to a State 1 information resource, but the forecaster needs to remain aware of potential future cost drivers so that decisions might be made that could keep life-cycle costs low. In most circumstances, labor costs will be the largest single element of her cost forecast. The activity list for State 1 (Table 2.1) can serve as a guide to the data steps that need to be considered. In the best of circumstances, the data set characteristics provide a way to estimate the amounts of each labor type that will be required, based on past experiences of the researcher or of those at her institution.

Since labor drives much of the cost, a rough first estimate might be informed by using data from any pilots of the proposed or current project and regressing the labor hours that were required for the development, population, and support of the pilot resource to the characteristics of the current or proposed resource. Labor costs will not necessarily increase linearly as the size and complexity of the resource increases given that some amount of efficiency of labor will be gained (i.e., the "learning curve"). If data from a pilot are not available, data from a similar research project might inform the estimate.

In many cases, that tabulated experience may not be available. The researcher could then resort to estimating the relative amount of each labor type, both within and potentially across activities (i.e., a set of ratios, based on experience and judgment). If one type of labor for one activity (perhaps based on a pilot) can be reasonably estimated, the ratios implied by such estimates will provide a way to forecast the quantity of the others. The institution may specify the rates that should be used for labor prices, but the tables in this report give her a backup resource. Since the State 1 researcher is typically looking at a relatively short time horizon, disruptors of the sort discussed in this report (see Chapter 7) are unlikely to play an important role, although to the extent they create trends that affect the near future she may want to modify the estimated prices involved to reflect that reality.

Forecasting for a State 2 (Active Repository) Resource

Cost forecasts for a State 2 (active) repository may have to address many of the same issues as discussed for a State 1 resource but with additional complexities. To the extent that the resource requires external funding, the first cost estimate will be one that focuses on the proposed budget. Sunk costs will be omitted from consideration (unless the funding entity allows for their cost recovery), and, as for the case of the State 1 researcher, costs subsumed in the overhead rate will be omitted. In preparing a proposed budget, the host institution will use the prices it faces. But for its strategic planning, especially for periods beyond that covered by any near-term financing, the State 2 repository host would be well advised to prepare an estimate that includes all resources necessary to sustain the repository over the long run, even if some of those resources are currently "free" or significantly subsidized. The State 2 resource host will be in the repository business for many years, and any subsidy structure from which it now benefits could change. It should understand not only the costs for which it must budget today, but also the total cost of its repository responsibilities should it eventually have to cover that entire forecast cost.

Since this forecast will likewise be forward looking, sunk costs would still be omitted, but refreshment or replacement of investments must be included. For its "all-resource" forecast, the State 2 resource host should only use prices in those cases where they actually reflect what is required to produce the necessary input. Otherwise, the actual resources consumed should be the basis for the cost forecast. In situations where the social cost differs from the market price (e.g., environmental effects of generating power), the State 2 resource host will at least want to understand the approximate magnitude of that difference, given the interest of stakeholder communities.

Considering the cost implications of alternative courses of action will be especially important for State 2 resource hosts. Again, calculating the present discounted values of various options and courses of action will give the State 2 resource host a method to weigh the costs of one course of action against another. The present discounted value calculation will be particularly helpful given that a State 2 resource host must necessarily look a long way into the future, providing a way meaningfully to sum up the long stream of operating costs that will be encountered, as well as required periodic reinvestments.

While labor costs may generally be the largest single budget item, other costs are likely to be more significant for the State 2 resource than for the State 1 resource. Physical facilities may be important, and licenses may constitute a significant element of expense. Software costs may be significant if proprietary software is used and licenses required, if existing software needs to be customized, or if entirely new code needs to be created. Purchased services are likely to be important, raising critical "build" versus "buy" choices for the State 2 resource host. An obvious such choice is the use of an in-house data environment versus that of a cloud provider (essentially, outsourcing). Note that cloud services may be provided by a commercial vendor, or through a research community cloud built on open-source software, such as that offered by the European Organization for Nuclear Research

(CERN).[12] Service providers of any type offer no guarantee of price stability, making it particularly challenging to forecast costs, especially given the substantial expense entailed in transferring data from one provider to another.

One of the several variables in selecting a service provider is whether the degree of security chosen will prove adequate over the long run and what the cost of upgrading security might be. Using a cloud provider does not relieve the State 2 resource host of security responsibilities, although it does change and, perhaps, reduce them. The cloud provider may have advantages related to, for example, economies of scale and ability to attract top-tier experts, but it may also represent a more attractive target for attackers. The cost of security across solution needs to be compared.

The State 2 resource host will likely have more experience than the State 1 researcher on which to base its estimates of the amount of labor required for each State 2 activity (see Table 2.2), and it may also be able to solicit advice from sister institutions. It may have the ability to pilot many of the repository activities for which it will be responsible. But to the extent it cannot construct labor forecasts based on experience or pilots, it can fall back on the technique sketched above for the State 1 researcher. Unlike the State 1 researcher, however, the State 2 institution may be free to set wage rates—and because it will be operating the repository for many years, it will need to think about how real wage rates (i.e., adjusted for inflation) will change in the future. For professional activities, real wage rates have been increasing steadily for many years, and the State 2 institution will need to take into consideration that likely trajectory. (Taking into account fringe benefits, especially health care, real costs for other classes of labor have also been rising, although much more gently.) For those classes of labor it currently employs, the State 2 resource host could start with its current rates, modified to reflect its contemporary recruiting and retention experience, and use its own recent trends as the basis for at least the intermediate-term trajectory of what will be necessary to attract and keep the labor force it will need.

Changes in the labor market are only one of the several disruptors the State 2 institution must consider. The full range of the disruptors raised by this report could affect its cost forecast, creating a situation where complexity could overwhelm its understanding of what the long-term commitment it is undertaking might require. For that reason, it may be useful to first forecast based on the (unrealistic) assumption of no change, then discuss which disruptors might significantly affect the forecast, versus those whose effects might be more modest.

Forecasting for a State 3 (Long-Term Preservation) Resource

However daunting it might be to forecast costs for State 1 and 2 resources, it could be more difficult to forecast costs for a State 3 (long-term preservation) resource. The forecaster may be making decisions about the format in which the data should be preserved and the nature of access to be supported years in advance of the actual transfer of data to a State 3 environment. Community guidance and standards may be helpful in making such decisions, with due allowances for how such guidance, standards, and access might evolve. Above all, decisions should be documented since they constitute the assumptions on which the forecast rests. If they are clearly stated, it will be easier to adjust for changed circumstances as the actual transition to a State 3 environment becomes likely. Once again, the characteristics of the data sets will probably be important predictors of storage costs and IT services; these will likely dominate the State 3 forecast. Labor costs may not be especially important once the data set is formatted for long-term retention, and facilities costs may be negligible. It is probably wise to use estimates of underlying costs for storage and IT services rather than current prices—planning for a State 3 environment should assume that the State 3 resource managers will bear the actual costs (e.g., it will not enjoy subsidies). This approach also facilitates embedding appropriate trends in the forecast. Because the State 3 environment will extend over many years of costs, it is essential to calculate present discounted values when comparing alternative courses of action.

In a sense, the State 3 resource investment could be viewed as an option on the future availability of the data set. While there is no market for such options, that intellectual construct could help guide decisions regarding the State 3 resource. If preservation options make a data set more discoverable or more easily reconstructed and used,

[12] The website for CERN is https://home.cern/, accessed March 27, 2020.

the potentially more valuable is that option. Conversely, decisions that make data harder to discover, reconstruct, and use, then the less valuable is that preservation option.

Reliability of cost forecasts is a critical issue, especially for State 3 environments with their high degrees of uncertainty. While distributions for cost parameters may not be available, the forecaster should nonetheless attempt to establish ranges for parameter values that capture central tendencies. These can be used to estimate how much "reserve" for various contingencies should be established or at least guide managers regarding the "what ifs" to which they should pay attention.

Data for Forecasting

As this discussion implies, the life-cycle forecasts of dollar values for each activity in the data states outlined in this report will depend on the specifics of the research project in the State 1 environment, the nature of the State 2 (active) repository, and the data preservation and access ambitions of the State 3 (long-term preservation) resource. Those differing dependencies make life-cycle forecast a unique undertaking. Is it possible to move beyond the qualitative observations on relative cost magnitudes of this report, perhaps based on a few top-level parametric forecasting equations, using just a handful of the key activities and data characteristics listed in Tables 2.1-2.3? In other cost-forecasting domains, the outcome has been the result of a multiyear sustained effort by dedicated professional staffs (e.g., for military weapon systems), capitalizing on detailed—and often proprietary—cost data. No such cadre now exists for the biomedical data challenge. Equally important, the committee could not discover any organized data-collection effort that such a cadre would need to create top-level forecasting tools. With the explosion of life science research and clinical data, and the hunger for good cost forecasts, establishing such a data-collection effort would be the first step to a better understanding of what will be needed, whether it is for the State 1 researcher, the State 2 active repository, or for State 3 long-term preservation.

INFRASTRUCTURAL ELEMENTS NOT CONSIDERED IN THE COST MODEL

There are many infrastructural or data environment systems, standards, services, and activities that are essential to data preservation and access broadly, and to biomedical data in particular, but where it does not make sense to try to allocate costs to specific sources or collections of data. Much of this is general infrastructure that supports many other activities of the university or other data platform host institution. Other costs are more specific to the work in the biomedical sciences and the communication of scholarship in those disciplines. Here, components that are particularly important to preservation and access for biomedical data are addressed.

The organizations, governance, standards, systems, and common knowledge structures are viewed as "community" problems rather than research; thus, funders do not want to support their solutions as part of an individual research project. It is worth, at the funding program manager level, considering investments in reflecting these standards and knowledge structures in common tools that can help the relevant research community. At the level of funding bodies and stewardship institutions, consideration needs to be given on how to support all parts of this infrastructure, particularly operations and maintenance.

Object Identifier Standards, Systems, and Governance

Identifiers are mechanisms for unambiguously referencing people, organizations, data objects, and things (e.g., genes, molecules, proteins, species). Typically, a sustainable organization needs to be emplaced to oversee and govern the assignment and use of identifiers, but funding is often not available for such oversight and governance. There will also be systems that look up information associated with an identifier. Perhaps the most critical identifier operationally is the DOI, most usually assigned through DataCite[13] (see Box 4.5). It is important to recognize, however, that many large, important State 2 (active repository) data aggregations also assign identifiers outside

[13] The website for DataCite is https://datacite.org/, accessed December 5, 2019.

BOX 4.5
DataCite

DataCite is a nonprofit organization that manages the DOI-assignment process for data sets and operates a registry that can be used to discover which repositories host a data set with a given DOI. Research community organizations can be members of DataCite so that their products can be assigned identifiers. DataCite is funded through membership and service fees received from membership organizations and through project grants from program sponsors, including the National Science Foundation, the European Commission, the Horizon 2020 program, the French National Research Agency, the Sloan Foundation, and the National Institutes of Health. According to DataCite's 2018 annual review, revenues from membership and service fees that year were 601,167€ (roughly $690,000 in December 2018; DataCite, 2018).

of the DOI system; an example of this would be Genbank[14] sequence identifiers, which are widely used in the scientific literature.

Personal Identifier Systems, Standards, and Governance

The cost of supporting the many identifiers is important for the production of good metadata and for accurate discovery by searching those metadata. For example, productive reuse of data, once identified, is often dependent on being able to identify a point of contact associated with those data if the data are not compliant with current standards, if the metadata are not complete, or if there is some other query about the data. Assigning identifiers to researchers and including that information in data sets becomes important. An example of an organization that governs and assigns such PIDs is the Open Researcher and Contributor Identifier (ORCID).[15] Repositories will need to understand the costs of using PIDs to identify contributors and contact people, and researchers will need to be trained on proper maintenance of their PIDs so that they may continue to be tracked if, for example, they change institutions. Using unambiguous PIDs, rather than normalizing personal names and dealing with variant name forms, will more efficiently provide better results when describing and searching.

Discovery Systems

It is important to recognize that, just as the formulation of preserving and storing data sets and related metadata is often oversimplified, oversimplification pervades the discussion of data discovery. The problem of aggregating and searching metadata for data sets held in a collection of repositories (e.g., the National Science Foundation Data Observation Network for Earth project for ecological and environmental data)[16] is complex. As the number of information resources multiplies, discovery systems will be needed (and will need to be supported) that allow people to find relevant resources. It is unclear who will support, build, and operate the key discovery systems. As indicated, DataCite operates a registry, but its searching capabilities are somewhat limited. Google has built Google Cloud Public Datasets[17] and Amazon Web services has built the Registry of Open Data,[18] although these are still best viewed as experimental. Many State 2 systems offer some kinds of searching over information that they host, but those capabilities will not extend to other repositories. The literature offers an important pathway to resource discovery, and the National Center for Biotechnology Information has invested heavily over the years in

[14] The website for Genbank is https://www.ncbi.nlm.nih.gov/genbank/, accessed December 5, 2019.

[15] The website for ORCID is https://orcid.org/, accessed December 5, 2019.

[16] The website for the Data Observation Network for Earth project is https://www.dataone.org/, accessed December 5, 2019.

[17] The website for Google Cloud Public Datasets is https://cloud.google.com/public-datasets/, accessed December 5, 2019.

[18] The website for the Registry of Open Data is https://registry.opendata.aws/, accessed February 12, 2020.

interconnecting PubMed[19] with some major State 2 platforms. As platforms for State 2 data aggregations multiply, there is also going to be a growing need for discovery tools, training, and outreach related to these resources.

Knowledge Structures

Standards and best practices for description of biomedical data objects rely not only on the use of identifiers as previously discussed but also on tools such as managed vocabularies and ontologies (i.e., knowledge structures). Many of these are highly specific to particular biomedical applications, and they often need regular maintenance to reflect new scientific developments. For example, NLM designed and maintains the MeSH thesaurus,[20] which serves to index the data in multiple NLM databases, including PubMed. NLM also provides a resource, through its Unified Medical Language System (UMLS),[21] that interlinks more than 200 terminologies in the biomedical domain. Some example terminologies include MeSH for the literature, SNOMED International[22] for clinical applications, and The Gene Ontology Resource[23] for genetic data.

REFERENCES

Cock, P.J.A., C.J. Fields, N. Goto, M.L. Heuer, and P.M. Rice. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* 38(6):1767-1771.

DataCite. 2018. DataCite Annual Review. https://datacite.org/documents/datacite_annual_review_2018_final.pdf.

Lantz, M. 2018. Why the Future of Data Storage is (Still) Magnetic Tape. *IEEE Spectrum*. https://spectrum.ieee.org/computing/hardware/why-the-future-of-data-storage-is-still-magnetic-tape.

Ross, R., V. Pillitteri, K. Dempsey, M. Riddle, and G. Guissanie. 2020. Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations. NIST Special Publication 800-171.

Rübel, O., A. Tritt, B. Dichter, T. Braun, N. Cain, N. Clack, T.J. Davidson, et al. 2019. NWB:N 2.0: An Accessible Data Standard for Neurophysiology. *BioRxiv: The Preprint Server for Biology*. Cold Spring Harbor Laboratory.

Teeters, J.L., K. Godfrey, R. Young, C. Dang, C. Friedsam, B. Wark, H. Asar, et al. 2015. Neurodata without borders: Creating a common data format for neurophysiology. *Neuroview* 88(4):629-634.

[19] The website for PubMed is https://pubmed.ncbi.nlm.nih.gov/, accessed December 5, 2019.

[20] The website for the MeSH thesaurus is https://www.nlm.nih.gov/mesh/meshhome.html, accessed December 5, 2019.

[21] The website for the UMLS is https://www.nlm.nih.gov/research/umls/index.html, accessed December 5, 2019.

[22] The website for SNOMED International is http://www.snomed.org/, accessed December 5, 2019.

[23] The website for the Gene Ontology Resource is http://geneontology.org/, accessed December 5, 2019.

5

Applying the Framework to a New State 2 Data Resource

In its statement of task (see Box 1.1), the committee was asked to apply the cost-forecasting framework to two case studies relevant to the National Library of Medicine's (NLM's) data resources. The case studies presented are based on hypothetical examples provided by NLM to the committee (personal communication, E. Kittrie, January 4, 2019). This chapter presents the first use case which describes decisions by a policy maker, program officer, or research group to estimate costs for a new repository hosting a large amount of data. Although the scenarios presented in this chapter and Chapter 6 represent traditional research environments, the framework could be applied to different research scenarios.

These case studies are not quantitative cost analyses. A quantitative forecast of a real-life scenario would require greater resources and time than the committee had to accomplish the task, and values obtained for the hypothetical cases would be meaningless given the number of variables presented by the data, the institutions involved, the resources available, the requirements of the funding entity, and so on. Instead, the committee provides high-level examples of how an investigator, a data-resource developer, or a resource manager could use the framework to systematically identify all the cost components which they then can use to develop their own meaningful forecast. In a true quantitative forecast, many more details—as dictated by circumstance—would need to be considered.

Cost forecasters creating or managing a State 1 (primary research) or State 2 (active) platform will unlikely be able to quantify the costs of data beyond the performance of their respective research projects. However, as stated in Chapter 4, understanding the potential future value of data and how decisions in early data states are made may affect the effectiveness and efficiency of future data preservation, curation, and use of the data. Considering the life cycle of data beyond the current data state and the resources necessary to transition between states will be increasingly important as data sets become bigger and more complex. Future cost ramifications could inform near-term decisions.

The cost forecasts take advantage of the cost-driver template in Appendix E. This template, based on the cost drivers in each state outlined in Table 4.2, assists in developing a narrative regarding the data life cycle. The approach to forecasting costs for the use cases follows the basic steps described in Table 4.1.

USE CASE 1: ESTIMATING COSTS ASSOCIATED WITH SETTING UP A NEW DATA REPOSITORY FOR THE U.S. BRAIN INITIATIVE

The cost-forecasting framework is applied to a new State 2 (active repository) platform. The study committee applied the framework as would a likely cost forecaster, in this case, a neuroscientist (see Box 5.1), and provides some information about existing platforms for context (see Box 5.2).

BOX 5.1
The Use Case 1 Cost Forecaster

The Use Case 1 persona is neuroscientist at University X. He is part of a large imaging center, and he works with computer scientists to develop infrastructure to manage and analyze large microscopic imaging data sets. He and computer science colleagues respond to a request for application (RFA) announced by the National Institutes for Health (NIH) Brain Research Through Advancing Innovative Neurotechnologies (the BRAIN Initiative)[a] for investments in new technologies, many of which involve imaging. The researcher applies the cost-forecasting framework to estimate a budget for a 5-year project for a new platform.

[a] The website for the BRAIN Initiative is https://braininitiative.nih.gov/, accessed December 12, 2019.

BOX 5.2
Existing BRAIN Repositories and Annual Budgets

The BRAIN Initiative goal is to increase the pace of innovative technology development and applications to allow researchers to produce a dynamic picture of the brain. To establish the necessary neuroinformatics infrastructure for data generated through funding, the BRAIN Initiative issued RFAs to (1) develop the necessary standards to support the data types generated by BRAIN investigators; (2) establish a set of specialized archives around a particular data type or technology to house the data; and (3) promote analysis and processing of data. Seven repositories have been funded in 2017-2019. The range of repositories and annual budgets are listed in Table 5.2.1.

Researchers supported by the BRAIN Initiative[a] must submit their data to BRAIN-sanctioned repositories, ideally on a continual basis (i.e., every 6 months). Proposals must include data management plans with descriptions of data to be shared, standard(s) to be used to describe the data set, the data archive(s) that will house the data, and the proposed timeline for submission to the archive and sharing data with the research community. Costs of complying with this mandate may be included in the budgets.

continued

BOX 5.2 Continued

TABLE 5.2.1 BRAIN Initiative-Funded Repositories (2017-2019) and Annual Budgets

Repository	Data Type	Location	Current Size	Annual Award (Total costs/Indirect costs/Direct costs)
BRAIN Image Library	Microscopic images	University of Pittsburgh	2.3 million files; 120 terabytes (TB)[a]	$1,054,315 $830,434 $223,881
Neuroscience Multi 'Omic Archive (NEMO)	'Omics data	University of Maryland	112.4 TB[b]	$1,247,018 $860,044 $386,974[c]
Data Archive for the BRAIN Initiative	Human neurophysiology	University of California, San Diego	New	$1,245,149 $806,242 $438,907[d]
OpenNeuro	Neuroimaging	Stanford University	5 TB[e]	$908,521 $578,676 $329,845
Block and Object Storage Service	Electron microscopy and x-ray microtomography	Johns Hopkins University	New	$593,507 $438,034 $155,473[f]
Distributed Archives for Neurophysiology Data Integration	Neurophysiology	Massachusetts Institute of Technology	New	$1,349,674 $1,131,040 $218,634
Brain Cell Data Center	Integrated data set for Brain Cell Census based on multiple data types	Allen Brain Institute	New	$2,965,990 $2,233,081 $732,909[g]

[a] 120 TB related to BRAIN project; 1.1 petabytes (PB) total as of July 2, 2019. Personal communication, Greg Hood, July 2, 2019.

[b] Posted on NEMO website, https://nemoarchive.org/, accessed January 11, 2019.

[c] Funding for 2017 from NIH Research Portfolio Online Reporting Tools (RePORTER).

[d] Funding for 2018 from NIH RePORTER.

[e] Personal communication, R. Poldrack, Stanford University, April 15, 2019.

[f] Funding from 2018 from NIH RePORTER.

[g] Funding for 2017 from NIH RePORTER.

[a] See National Institutes of Health, "Notice of Data Sharing Policy for the BRAIN Initiative," NOT-MH-19-010, Release Date January 22, 2019, https://grants.nih.gov/grants/guide/notice-files/NOT-MH-19-010.html.

Applying the Framework to Use Case 1

Using the forecasting steps provided in Table 4.1, the researcher begins to construct the cost forecast.

Step 1. Determine the type of data resource environment, its data state(s), and how data might transition between those states during the data life cycle.

The first two columns in Table 5.1 list the data archive requirements as specified in the RFA. After considering the activities associated with each of the data states described in Tables 2.1, 2.2, and 2.3, the researcher concludes

TABLE 5.1 Specific Services Specified in the Request for Application Mapped to Data States, Activities, and Subactivities

Research Objective Number	Archive Requirements as Specified in the BRAIN Initiative RFA-MH-17-255	States, Activities, and Subactivities[a]
1	The data archive is expected to use relevant standards that describe BRAIN Initiative experiments. Such standards may be developed under RFA-MH-17-256 or may already exist.	II.A.1, II.B.1
2	A data archive will develop a data submission pipeline ensuring appropriate quality control standards for laboratories that are trying to upload data. For example, if an experimental standard defines an allowable range of values for a particular data element, the submission pipeline should make sure that uploaded data respect the current data standard.	II.B.1, II.B.7, II.E.2
3	Ideally, the data archive will create both a submission pipeline and a related validation tool to allow researchers to check the quality of their data even if they are not trying to upload data. . . . Data submission pipelines that originate with the data-collection instrument in the depositor's laboratory and require minimal manual intervention would be ideal but are not required.	II.B.1, II.B.7, II.C
4	A data archive will work closely with BRAIN Initiative awardees and others to collect and archive relevant data sets.	II.A.3, II.D.1, II.D.2
5	Each data archive should plan for a help desk to work with those who are trying to upload data.	II.I.2
6	Each data archive must develop plans to make the data readily available to the broad research community and to citizen scientists, as appropriate.	II.B, II.I
7	Depending on the type of data, data submission agreements and data access agreements may be necessary.	II.D.2, II.D.3
8	In many cases, processed data may be as useful to the research community as the raw data produced in the laboratory. Each data archive should consider storing and curating the appropriate data (either raw or processed) and make them available to the community.	II.E, II.H
9	A data archive may propose evaluating deposited data and scoring them to allow the research community to have some guidance about data quality.	II.E.2, II.E.3
10	Each data archive should plan to assign persistent identifiers to deposited data and to processed data to allow the research community a very easy way to cite the data sets that are being used.	II.E.5
11	A data archive should allow researchers to have a space where they can share data privately to facilitate collaboration prior to publication. Such private enclaves must last for only a defined period of time before that data set is shared with the rest of the research community.	II.B.9
12	A data archive may help users deposit data into other sustainable databases, such as those supported by the National Center for Biotechnology Information, but this is not a requirement.	II.I.2, II.L.3
13	There may be cases where data are stored in more than one data archive. In those cases, a data archive funded under this funding opportunity announcement will ensure that the user community can find all relevant data using appropriate linkages or database federation strategies no matter where the data are actually stored.	II.F.2
14	Furthermore, each data archive will provide an interface that is accessible to anyone with a web browser.	II.B.7, II.H.3, II.H.4
15	A data archive will make appropriate query tools and summary data easily available to allow the research community to check whether data of interest are held in the archive.	II.B.3, II.B.4, II.B.8, II.B.10, II.H
16	The user interface should make the maximum amount of information available to the research community while considering user friendliness and ease of interpretation.	II.B.7, II.H

continued

TABLE 5.1 Continued

Research Objective Number	Archive Requirements as Specified in the BRAIN Initiative RFA-MH-17-255	States, Activities, and Subactivities[a]
17	The website is expected to have a broad user base that will include both naïve users and experienced bioinformaticians, and should provide an interface that will accommodate both types of users.	II.B.7, II.H
18	In many cases, users will want to analyze or use visualization tools to interact with the data without downloading any data. Those interactions should be anticipated by the data archive.	II.B.3, II.B.5
19	Expensive computations could result from some analysis activities, and the data archive should explain plans to deal with such eventualities.	II.B.3, II.B.4, II.B.5, II.I
20	A data archive may, but is not required to, use cloud storage and computing capabilities to enable the research community to analyze data without downloading them. A data archive should (but is not required to) allow users to bring their own analysis tools to the data.	II.A.3, II.B.1, II.K.1
21	Each data archive will be expected to have staff who are knowledgeable about informatics and the experimental data being collected. The informaticists will be responsible for coordination with other relevant informatics efforts.	II.B
22	In particular, a data archive will be expected to identify and federate the archive with other data repositories and knowledge bases, as appropriate.	II.F.2, II.F.3
23	This data archive integration should create ways for users to query all relevant data repositories for relevant information. Funded data archives will be members of a larger BRAIN Initiative Data Network that will work across BRAIN Initiative activities to promote integration of a variety of data types.	II.F.2, II.F.3
24	In addition, the data archive will interact, as appropriate, with informatics activities outside the BRAIN Initiative such as the NIH Big Data to Knowledge effort and the work of the International Neuroinformatics Coordinating Facility (INCF).	II.A
25	When possible, a data archive is expected to use existing infrastructures and standards. These could include persistent identifiers such as Digital Object Identifiers (DOIs) or Resource Identifiers.	II.B.10, II.E.4, II.E.5

[a] Activities are defined in Tables 2.1, 2.2, and 2.3 of this report. The Roman numeral refers to the data state, the capital letter refers to the major activity, and the Arabic numeral refers to the subactivity. The cost forecaster can use this information to consult Table 4.2 to identify likely cost drivers for each activity.

that the proposed data resource will be an active repository and platform and therefore a State 2 resource. The researcher begins to match activities associated with a State 2 (Table 2.2) resource with the specific research objectives described in the RFA and lists which State 2 activities found in Table 2.2 would be necessary to accomplish each of the research objectives (the third column in Table 5.1). The costs and cost drivers associated with the activities will be revisited later in the cost forecast by consulting Table 4.2.

Because the researcher is interested in preserving the long-term value of the data and increasing the efficiency and effectiveness of their long-term curation and use, the researcher also considers activities related to eventual transfer of data to another State 2 resource or long-term State 3 archive. These latter considerations were not activities specified in the RFA.

Step 2. Identify the characteristics of the data (Chapter 4), data contributors, and users.

The next sections summarize a high-level consideration of this step, although, in reality, this step would be revisited several times as resources are characterized; choices about the repository are refined; and characteristics of the data, data platform, and contributors and users are better defined through use of the template in Appendix E.

Data Characteristics (Sections A and E of the Cost-Driver Template in Appendix E)

- The files are large: TB per individual data set.
- There are many files and large size of individual files. Raw data may be contained in thousands of individual files. For example, a single serial section electron microscopy data set covering less than 0.5 mm^3 of cortex by Bock et al. (2011) comprised 36 TB of raw data and 10 TB after processing to stitch the individual tiles together and reconstruct the volume.
- Sizes are likely to increase over the life span of the resource.
- There are multiple modalities.
- The data are complex: two-, three-, and four-dimensional images.
- There are significant metadata requirements.

Because of the rapid development in algorithms for processing and reconstructing the data, both raw and processed data will likely need to be stored, and compression algorithms for high-resolution scientific imaging data are likely to interfere with the reuse of the data for many applications. Imaging will likely be from animal subjects, minimizing costs associated with security and confidentiality (Section H in Table 2.2). The repository has decided that all data will be offered under the same license, minimizing any costs associated with enforcing multiple permissions.

Contributors/User Community (Section F of the Cost-Driver Template in Appendix E)

Assuming 200 BRAIN-funded users submit data twice a year, 400 independent submissions per year could be expected. Contributor support needs will likely be high, given data complexities and size, particularly in the early years when data validation and upload pipelines may not be fully mature. Data contributors will likely have a sense of urgency to upload backlogs of data before their grant funding runs out. If standards and best practices are not fully in place when the resource begins to acquire data, then a backlog of data will need curation. The resource has to decide whether to devote extra staff and funding to re-curating those data when standards and tools are in place.

The user community is also expected to be diverse, with scientists working in different environments. The RFA requires the resource to work closely with contributors (Table 5.1, research objective 4), maintain a help desk (research objective 5), and make data available to the broad research community, including citizen scientists as appropriate (research objective 6). Given the range of user skills to be accommodated—and the cost to develop intuitive user interfaces to do so—general help, training, and outreach materials are likely to be increased. Frequent updating of help materials may be necessary during early phases, when the technology is changing on a regular basis.

Step 3. Identify the current and potential value of the data and how the data value might be maintained or increased with time.

The perceived and long-term value of data can be informed by answers to Sections A, D, and E of the cost-driver template in Appendix E and through consultation with experts and colleagues. The perceived long-term value of the data in the proposed resource will depend on hard-to-estimate factors. Some data will derive from new and rapidly developing techniques. Colleagues and experts think that some data may be superseded as technologies improve. On the other hand, data that are the result of a complex experimental paradigm—for example, the carefully correlated light and electron microscopy work of Bock and others (2011)—may be quite valuable even if of lower quality. Well-annotated imaging data tend to be interpretable and usable for a long time in different contexts. The long-term value cannot be estimated at the time of proposal preparation, making decisions about the level of replication and access (e.g., transfer to a less expensive form of storage) difficult in the early stages.

Step 4. Identify the personnel and infrastructure likely necessary in the short and long terms.

The forecaster considers which of the activities described in Table 2.2 are relevant to the RFA requirements (Table 5.1) and the proposed database more deeply. This resource will not handle sensitive information; thus, some activities will not be necessary. The forecaster next considers how these activities might be accomplished (and by whom), again referring to the list of expertise included for each activity in Table 2.2. To estimate personnel costs the forecaster needs to consider how long each task will take, the skill levels necessary, the availability of people with the skills, and the tool support or training necessary for the people to perform the tasks. In reality, many of the positions listed in Table 2.2 are likely not to be included or consulted when setting up a typical researcher-led scientific infrastructure, but it is worth considering the value of including them and how their involvement influences overall cost. For example, many new resources struggle with metadata. Consulting a data librarian or records specialist early may help to reduce cost, improve quality, and increase FAIRness by providing advice on community standards for high-level metadata and specialized metadata schemas.

Because the repository infrastructure already exists in this fictional use case, many setup costs might be reduced, but significant customization will likely be necessary. If the infrastructure had to be developed from scratch, the forecasters might consider whether instances of existing infrastructure could be set up or whether they could partner with an existing repository to provide the back-end infrastructure.

Step 5. Identify the major cost drivers associated with each activity based on the steps above, including how decisions might affect future data use and its cost.

Table 4.2 is consulted to identify the cost drivers often associated with a State 2 resource, and the cost-driver template found in Appendix E is completed (the template is based on the cost-driver questions found in Chapter 4). The completed template is presented as Table 5.2, following the discussion of Use Case 1, below. The completed template will help the forecaster determine which decision points will likely control costs now and in the future, and it will help the forecaster understand when specific costs will be borne and by whom.

The responses to the cost driver questions shown in Table 5.2 allowed the forecaster to create a narrative to help him identify exactly what will be involved in establishing the State 2 (active repository) resource. From that narrative, the forecaster could determine how influential each of the respective costs is likely to be in the overall costs (listed below). In a quantitative cost forecast, the costs for the activities could then be quantified, and each of the major cost components (e.g., Box 3.2) worked out.

- A: Content → Likely **high**
- B: Capabilities → Likely **medium-high**
- C: Control → Likely **medium**
- D: External Context → Likely **low**
- E: Data Life Cycle → Likely **high**
- F: Contributors and Users → Likely **high**
- G: Availability → Likely **medium-high**
- H: Confidentiality, etc. → Likely **low**
- I: Maintenance and Operations → Likely **low**
- J: Standards, etc. → Likely **medium**

Step 6. Estimate the costs for relevant cost components based on the characteristics of the data and information resource.

As noted elsewhere in the report, the ability to estimate actual costs is dependent on so many factors that the committee elected not to attempt this exercise. How data size can influence costs can be exemplified by using cost estimators provided by commercial cloud services (in this case, the Amazon Simple Storage Service[1] cost tools). The absolute size beyond a certain threshold may not impose many additional costs for cloud storage. For example, as of this writing, the cost to store up to 50 TB is $0.023 per gigabyte (GB)/month, whereas over 500 TB for a month brings the cost down to $0.021 per GB, according to that cost estimator. However, given the anticipated growth of the data, the storage cost is not insignificant in absolute terms. For one PB of data, the cost, absent any institutional discounts, would be $21,000 per month depending on level of access. Cost over time will need to be considered. Cloud service prices may change, or circumstances may warrant a change to a different provider with different cost structures, services, and data formatting requirements. The size of the data may also impose costs for functions such as external backup, replication, and data transfers (G.4), depending on what infrastructure is available to the resource. The forecaster will want to compare full costs of storage from multiple service providers, including the fully loaded costs of local computer resources.

The complexity of the data can also impose significant costs. The capabilities related to functional specification and implementation (Activity II.B) will need to be developed or modified and maintained and the standards for multiple data types and paradigms developed or implemented. These functions may need to be multiplied by the number of data types and modalities to be supported, depending on how well the tool set generalizes.

Once the resource is mature and data access and use patterns emerge, some significant cost savings may be realized by moving unused or obsolete data to cold storage. Again, using commercial cloud provider cost tools illustrates how storage costs are affected by access and responsiveness requirements. For the Amazon Web Services S3 Intelligent Tier pricing model, designed for data where access is infrequent or unknown, the cost for storage that is accessed at high frequency is $0.021-$0.023 per 50-500 TB but only $0.0125 if infrequently accessed. If it is known that the data are infrequently accessed and users can tolerate slow retrieval times (minutes to hours), then the cost of access will drop to $0.004 per GB. For a 500-TB data set, the cost of storage would drop from $10,500 per month to $2,000. Cold-storage options are best considered during the first funding period and in consultation with the community served so that expectations are clear.

The data and the user community characteristics will also be a major determinant of decisions about infrastructure for hosting and accessing the data, as well as the necessary user support levels. The RFA does not require that the resource utilize the cloud; the large size of the data and the unknown growth characteristics of both the data and the user community make the cloud attractive, as it can scale with increasing demand. Costs associated with data transfers, search, computational services, and downloads will need to be carefully monitored. Costs might be driven by unexpected demand surges (e.g., a data set is posted on social media and is heavily accessed). This fictional use case will protect itself from unexpected and uncontrollable charges by passing the cost for download to the end user. Many cloud providers now provide tools and safeguards for monitoring and limiting costs. Taking advantage of local or government programs (e.g., Cloudbank and the Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability) to make informed decisions and gain access to expertise about building platforms in the cloud could also lower costs.

Last, although the RFA does not specify that the resource should develop an exit strategy, thinking about the long-term data and resource viability is good stewardship. Bilder and others (2015), in their principles of open scholarly infrastructures, recommend that every resource have a “living will” that describes how a resource would wind down. In the proposed large (multiple PB) hypothetical BRAIN repository, the costs of transferring data to another State 2 active repository or to a long-term archive could be significant.

REFERENCES

Bilder, G., J. Lin, and C. Neylon. 2015. “Principles for Open Scholarly Infrastructure-v1.” https://doi.org/10.6084/m9.figshare.1314859.

Bock, D.D., W.-C. Allen Lee, A.M. Kerlin, M.L. Andermann, G. Hood, A.W. Wetzel, S. Yurgenson, et al. 2011. Network anatomy and in vivo physiology of visual cortical neurons. *Nature* 471(7337):177-182.

[1] See Amazon Web Services, “Amazon S3 Pricing,” https://aws.amazon.com/s3/pricing, accessed December 16, 2019.

TABLE 5.2 Completed Cost-Driver Template for Use Case 1: Setting up a BRAIN Archive

Category	**Cost Driver**	**Decision Points/Issues**	**Relative Cost Potential (Low, Medium, High)**
A. **Content**			
A.1	**Size** (volume and number of items) > size = higher costs	1. **How many files will be in a single data submission?** Varies, likely from 10 to 10,000. 2. **How large is an average data submission in total?** Multiple TB. 3. **Are the data sizes likely to stay stable over the life of the resource?** No, file sizes will likely increase as technologies are developed. 4. **What is the total amount of data expected?** PBs. 5. **In what kind of medium will data be captured in the short and long terms?** Data upload into the cloud for short and long term will be captured.	H
A.2	**Complexity and Diversity of Data Types** **>** complexity + diversity = higher cost	1. **How complex is the underlying structure of the data?** Complex-image data. 2. **How are the included data to be organized?** To be determined after interviewing funded investigators. Likely individual data sets that include the raw and processed data, but need to determine whether the data should be organized according to studies or projects. 3. **How complex is the experimental paradigm that produced the data?** Varies—some simple acquisitions; some associated with complex behavioral paradigms. 4. **What sort of additional files might be necessary to upload with the data to properly understand them?** Experimental protocols, fiducial maps. 5. **How many different data types are being produced?** Multiple types of imaging data (multimodal data—light and electron microscopy, multiple microscopy types within each; correlated physiology and genomics. 6. **What are the relationships among these data types (e.g., are the data correlated)?** Some correlated data sets; related data sets will be deposited in the appropriate repository.	H
A.3	**Metadata Requirements** > metadata amounts + type = higher cost	1. **How much metadata must be stored with each data object to make them FAIR?** Basic descriptive metadata, imaging parameters, experimental metadata, processing metadata, anatomical metadata. 2. **Will metadata be entered manually by the submitter/curator?** Yes, by both submitters and curators. 3. **Will the data to be deposited include a data schema, or will one be generated?** Eventually, a common schema will be created for all data based on standards such as Open Microscopy Environment and those deriving from BRAIN. 4. **Is the provenance of a data set sufficiently described, or will it need to be?** It will need to be described, as the processing pipelines are not standardized. 5. **How much metadata can be extracted computationally?** Imaging parameters may be able to be extracted from image headers.	M
A.4	**Depth Versus Breadth** > breadth = higher cost	**Will the repository be restricted to certain data classes or types that the repository must support?** It will primarily focus on imaging data, but there will be multiple types of imaging data from multiple domains.	M

TABLE 5.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
A.5	**Processing Level and Fidelity** > compression = lower cost	1. **Do the raw data need to be stored?** Yes, as the reconstruction algorithms are rapidly developing. 2. **Do processed data need to be stored?** Yes, it would be computationally intractable to reconstruct two-, three-, and four-dimensional data sets each time a user accesses them. Some data may be mapped to a common coordinate framework, and both the raw and aligned data will likely be stored. 3. **Are there compression algorithms that can reduce the file size without compromising fidelity?** Some, but generally they result in signal loss, so it is not advised. 4. **What kind of data structure requirements will the resource have?** The goal is a common data structure to organize the large numbers of raw files and any derived data, e.g., reconstructions. 5. **Is the data contributor or the repository responsible for any restructuring necessary?** Data contributor. 6. **How is the data structure verified?** A validator will be developed as per the RFA. It may not be ready in year 1 because of testing of data set structures based on data likely to be received.	H
A.6	**Replaceability of Data** > replaceability = lower cost	1. **Is the archive the primary steward of the data, or do copies exist elsewhere?** The resource is expected to assume stewardship of the data. 2. **Can the data be easily recreated?** Not currently, but possibly in the future.	H
B. **Capabilities**			
B.1	**User Annotation** > user annotation functions = higher cost	1. **Will the repository have to provide user annotation capabilities?** Given the size of the data, it would be ideal to have annotation capability available to the user base. 2. **What is the nature of these annotations?** Anatomical delineations, molecular distributions. 3. **Are they provided by humans or machines, and how will they be authenticated?** Mainly through machine-based segmentation. Users will have an account to annotate; annotations will be tied to their Open Researcher and Contributor Identifiers. 4. **Are permissions required to annotate the data?** No, the data will be in the public domain and they are free to annotate, although stored annotations will have to be attributed to the individual researcher.	H
B.2	**Persistent Identifiers** type of identifier = potential costs	1. **What persistent identifier (PID) scheme will be used by the archive?** DOIs for data sets and for reconstructed images/volumes. 2. **Is there a cost associated with issuing the PID?** Yes, but covered by institutional membership to DataCite. 3. **How many objects need to be identified?** DOIs to data sets and reconstructions will be issued, but not individual files; therefore, two to five identifiers per data set. 4. **Who will be responsible for keeping the PIDs resolvable?** The database administrator will be responsible for notifying DataCite of any changes in data object location.	L

continued

TABLE 5.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
B.3	**Citation** > citation functions = increased cost	1. **Will users be able to create arbitrary subsets of data files and mint a PID for citation?** Yes. 2. **Will the repository provide machine-readable metadata for supporting data citation?** No. 3. **Will the repository provide export of data citations for use in reference managers?** No.	L
B.4	**Search Capabilities** > search capabilities = increased cost	1. **Will the repository provide a search capability for data sets?** Yes. The repository is also required to provide means to search other BRAIN repositories through the BRAIN Initiative Data Network. 2. **How much of the metadata will be included in search?** Initially, the repository will support search via database-level metadata. More detailed fields may be added in response to user requirements. 3. **How complex are the queries that will be supported?** Keyword search, structured search on basic metadata fields. 4. **What type of features for search will be provided?** Synonym expansion using the Neuroscience Information Framework's vocabulary services. 5. **Will the repository deploy services to search the data directly?** Data-feature search capability is planned.	H
B.5	**Data Linking and Merging** > linking and merging = increased cost	1. **Will the data require/benefit from linkages to other related items?** Required to link to data in other BRAIN repositories. 2. **Will the resource provide the ability to combine data across records based on common entities/standards?** Will build a knowledge graph on top of our data records so that we can combine across data records.	H
B.6	**Use Tracking** > tracking = increased cost	1. **Will the resource provide the ability to track uploads, views, and downloads?** Yes. 2. **If so, and if made available to users, how will this information be made available?** Tracked and displayed per data set. 3. **Will the resource track data citations to its data?** No.	L
B.7	**Data Analysis and Visualization** > services = higher cost	1. **What type of data visualization will the repository support?** Interactive viewing of images and three-dimensional volumes using image services. 2. **What types of other data operations will the repository support (e.g., file conversions, sequence comparison)?** Will develop an image-feature search capability. 3. **Do these services require significant computational resources?** Yes. 4. **Who will pay for these computations?** We will assume the costs of the search algorithms.	H

TABLE 5.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
C. **Control**			
C.1	**Content Control** > review processes = increased cost	1. **Will all appropriate data be accepted, or will there be a review process?** All relevant data from BRAIN investigators will be accepted; Data from outside BRAIN will be accepted after the infrastructure is built. 2. **Will the review process be automated, or will it require human oversight?** Data must pass our automated validation checks, but human curators will oversee the project and provide additional curation of metadata.	H
C.2	**Quality Control** > quality control = increased cost	1. **What quality control processes will the repository support?** Format and metadata review quality control on quality of data and reconstructions will be up to the submitter. 2. **Will these be automated or require human oversight?** See C.1. 3. **What level of data correctness will be required, and how will it be validated?** Data are expected to pass our validation checks with no errors. 4. **What gaps in the data at the record or field level will be tolerable?** Difficult to estimate at this time. Most likely applicable at raw-data level—missing or corrupted files may impact the quality of the reconstruction. 5. **Will any of the data be time sensitive, and how will data currency be ensured?** Not beyond ensuring that data referred to in publications are released after the agreed-upon embargo period. 6. **How will duplication within or between data sets be addressed?** Given the size of the data, if any data are cross-referenced across data sets or resources, it will be in the form of a link and not a duplication of the data. 7. **Will prevalidation guidelines or routines be distributed by the resource to the data contributors?** Yes, as per the RFA. Researchers should be able to validate their data as they are acquired. 8. **Will human curation be necessary?** Data submissions will be monitored in the early phase to determine whether human curators will be necessary to improve the quality of the data. While the hope is that automated tools may be sufficient, some human curation likely will be necessary.	L
C.3	**Access Control** > controls = increased cost	1. **What types of access control are required for the repository (e.g., will there be an embargo period)?** Data are public; embargo period will be provided. 2. **At what level are they instituted (e.g., individual users, individual data sets)?** Embargo periods will be instituted for individual data sets where only specific users, including reviewers if required, will have access to them. After the embargo period, all data are public. 3. **Does use of the data require approval by a data access committee?** No.	M

continued

TABLE 5.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
C.4	**Platform Control** > platform restrictions = increased cost	**Are there restrictions on the type of platform that may or must be used?** No, free to use the cloud if desired, and there are no requirements to use a specific cloud provider. Data will not be mirrored overseas.	L
D. **External Context**			
D.1	**Resource Replication** > replication = increased cost	**Is there a requirement to replicate the information resource at multiple sites (i.e., mirroring)?** No.	L
D.2	**External Information Dependencies** > external dependencies may or may not = increased cost	**Will the resource be dependent on information maintained by an outside source?** Will use community ontologies for certain metadata.	L
D.3	**Distinctiveness** > distinctiveness = increased cost	**Are there existing resources available that provide similar types of data and services?** Yes, the Cell Image Library. The EU also has a Bioimaging Database.[a]	L
E. **Data Life Cycle**			
E.1	**Anticipated Growth** > growth = increased costs	1. **Is the repository expected to continuously grow over its lifetime?** Yes. 2. **Is the likely rate of growth in data and services known?** Not entirely. 3. **Is the use of the repository likely to grow over time?** Yes. 4. **Is the likely growth of the user base known?** No.	H
E.2	**Update and Versions** > updates + multiple versions = increased cost	1. **Will the deposited data require updates (e.g., in response to new data or error corrctions)?** Yes, some data will be submitted in batch mode, as the data need to be deposited at regular intervals. A policy on error correction will be developed (i.e., related to when corrections trigger a new DOI). 2. **Will prior versions of the data need to be retained and be made available locally or in a different resource?** Yes, if the data are in the public domain; no, if they are in the embargo phase. 3. **How frequently will individual data sets be updated?** Unknown.	H
E.3	**Useful Lifetime** limited lifetime = decreased cost	1. **Are the data to be housed likely to have a limited period of usefulness?** Hard to predict; later acquisitions likely to have longer periods of usefulness. 2. **Does the resource have a defined period of time for which it will operate?** No. 3. **Does the resource have to provide a guarantee that the data will be available for a finite period of time (e.g., 10 years)?** No, there is no set period specified.	L

TABLE 5.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
E.4	**Offline and Deep Storage** > offline/ deep storage = decreased costs > transfers = increased cost	1. **Can the resource take advantage of offline storage for data that are not heavily used?** As the resource grows, access to data sets will be monitored. Those not accessed heavily will be moved to less expensive storage. 2. **Does the resource have a plan for moving unused data to deep storage (i.e., State 3)?** At the end of the life span of the project, data will be moved to a suitable State 3 archive; however, the specific archive has not yet been identified.	H
F. **Contributors and Users**			
F.1	**Contributor Base** > number and diversity of contributors = increased cost	1. **Is the number of contributors known? If not, can it be estimated?** The precise number is unknown, but it is assumed that one-third to one-half of the 700 BRAIN grant awardees generate imaging data that would be appropriate for this resource. Assuming ~200-250 contributors. 2. **Are all data originating from the same source (e.g., a single instrument or a single organization)?** No, the data will be coming from all different laboratories and therefore different environments and instruments. 3. **How will data be transferred into the data resource (e.g., periodic large batches, more frequent smaller data sets, constantly streamed, by physical transfer)?** Periodic large batches as specified by the data-sharing policy. 4. **Will the data be pushed by the contributor or pulled by the resource?** Pushed by the contributor. 5. **Are there direct or indirect fees associated with acquiring the data from a source?** No. 6. **Will a data steward be available from among the contributors to assist with any data integration into the data resource?** Unknown, but unlikely.	H

continued

TABLE 5.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
F.2	**User Base and Usage Scenarios** > access and diversity of users = increased cost	1. **How many users will likely access the data?** Unknown, but assuming that it will be around 5,000 to 10,000 per month based on analytic data from similar resources. 2. **What will be the frequency of access?** Difficult to estimate and likely depends on whether we get the image-analysis services running. 3. **How will users access the data?** As these are large image data sets, most researchers will likely interact with the data using our image services and computational platform rather than downloading it. For some operations and tool sets, e.g., manual segmentation, it may be necessary to download the data or transfer them to another cloud. 4. **Will the resource be building analysis tools?** No, beyond basic functions, the RFA states that the resource does not have to build analysis pipelines or tools. 5. **Will the resource support individual file download or bulk download?** Download will be at the level of data sets (all relevant files) and individual files. No bulk download will be provided. However, it is anticipated that researchers will not download data but bring their compute needs to the data. 6. **Will there be any fees for downloading/accessing the data?** If cloud provider used, users will pay for downloads and for the deployment of their algorithms. 7. **How many different types of users must be supported?** Three different types of users are anticipated: (1) neuroscience researchers with domain expertise, (2) computational researchers with little domain expertise, and (3) citizen scientists, as per the RFA.	H
F.3	**Training and Support Requirements** > training + services = increased cost	1. **Will training for resource use be offered?** Yes. 2. **What form will the training take?** Online tutorials, hackathons, webinars, live demos at conferences. 3. **Will a "help desk" be provided?** Yes, as per the RFA. 4. **When does live help need to be available?** During normal business hours. 5. **What is the expected skill level of the user base?** As indicated in F.2., the resource will need to support a broad user base with a range of skills.	H
F.4	**Outreach** > outreach = increased costs	1. **Does the existence of the repository need to be advertised?** Yes, to attract outside users. It is assumed that BRAIN awardees will know of our existence. 2. **How many conferences per year should resource representatives attend?** At least two. 3. **Will the resource have a booth at the conference for live demos or to conduct hands-on tutorials?** Booths at at least one conference per year for live demos. 4. **Are users required by funders or journals to deposit data in the repository?** Yes.	M

TABLE 5.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
G. **Availability**			
G.1	**Tolerance for Outages** < tolerance for outages = increased costs	1. **What is the tolerance for outages for the resource?** As users from around the world are expected, the resource will be up 24/7, except for scheduled maintenance. However, it is difficult to predict the amount of usage for the resource at this time. If the resource is heavily used, the tolerance for outages will be less and we will aim for > 99 percent availability. 2. **What measures will be taken to avoid and mitigate outages?** A fully redundant system with high fault tolerance will be implemented as the resource scales up. Such redundancy would roughly double the cost of maintaining the system. However, as this level of tolerance will not be needed in the first few years, databases and system architecture will be designed such that it will be easy to fully replicate the system in the future, if level of usage demands it. 3. **How quickly and completely does the resource need to recover from an outage?** As researchers are required to submit their data to the archive to meet their requirements, the data will likely be used by third parties. Outages not due to regular maintenance will be kept to under 30 minutes. However, even the major commercial cloud providers can experience outages that last for several hours. Given the size of the data, replicating the full resource at multiple sites, including our local site, would be cost prohibitive. So although rare, outages that last longer may occur.	M
G.2	**Currency** > currency = increased cost	1. **How often will the data be released?** Data sets will be released to embargo as soon as they are uploaded; when the embargo period passes, they will be automatically released to the public with a DOI and the appropriate license. 2. **How soon do data need to be made available after they are received?** Will have to be negotiated with individual users, but we expect within 1 week on average.	M
G.3	**Response Time** > responsiveness = increased cost	1. **Are there requirements for response time for service?** For interactive browsing of the two- and three-dimensional image data, very responsive image services needed. 2. **Are there requirements for responses from humans?** Users should receive an automated response for any help request immediately and a human follow-up within 1 business day.	M
G.4	**Local Versus Remote Access** > cloud could lead to increased costs	1. **Does the resource require that any data be shipped via physical media?** Yes. Depending on the size of the data set and the bandwidth available, data may need to be shipped back and forth via physical media. 2. **Will the resource be built using commercial clouds?** Yes, commercial clouds for storage and computation will be used. 3. **Do users have to travel to the resource to use the data?** No. All access is through the web.	H

continued

TABLE 5.2 Continued

Category	**Cost Driver**	**Decision Points/Issues**	**Relative Cost Potential (Low, Medium, High)**
H. **Confidentiality, Ownership, and Security**			
H.1	**Confidentiality** > confidentiality = increased cost	1. **Will any of the data require special protections?** No, identified human data will not be hosted. 2. **Will any of the data have embargo periods or embargo-related limitations that may entail costs?** There are initial costs for implementing embargo features; ongoing costs will be minor. 3. **Are there any audit requirements for who has accessed or downloaded the data?** No.	L
H.2	**Ownership** > ownership = increased costs	1. **If data are contributed from multiple sources, will there be a need to process multiple kinds of release forms?** No, all data will be released under the same license and it will be an open license per the requirements of the BRAIN Initiative for public data. 2. **Will all the data be released under the same license, or will different permissions be assigned to different data sets?** All data will be released under the same license, CC-4.0-BY, as per requirements by the funder. 3. **Will data submission agreements be necessary?** Not anticipated. Data acquired under BRAIN are required to be submitted to an archive and made available.	L
H.3	**Security** > security = increased cost	1. **What measures need to be taken to ensure the integrity and availability of the data?** Standard practices will be used. 2. **Do these measures require using protected computing, storage, or networking platforms?** No.	L
I. **Maintenance and Operations**			
I.1	**Periodic Integrity Checking** > integrity checking = increased cost	1. **What processes will be put in place for checking the integrity of the hardware, software, and data?** A hashing function will be implemented to ensure data integrity. Checksums will be used for each data upload and download. 2. **How frequently will these checks be performed?** Every 3 to 6 months for system checks. At every upload and download for data use.	M
I.2	**Data-Transfer Capacity** > data-transfer upgrades = increased cost	**Will the bandwidth available to the resource be sufficient for the data sizes and rates required?** Campus connectivity was recently upgraded, so no internal problems anticipated, but there is no control over our submitters and users. See G.4.	L
I.3	**Risk Management** > risk mitigation = increased cost	1. **Will the repository be solely responsible for risk mitigation?** As the repository of record, responsibility for the data assumed and therefore appropriate backup strategies will be implemented. 2. **Is a response plan for unexpected termination required?** No requirements were given to have an exit plan and we would assume that funding would be given by NIH to terminate our resource and transfer the data to an archive of their choosing.	H

TABLE 5.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
I.4	**System-Reporting Requirements** > system-reporting requirements = increased costs	**What types of system reporting will the resource be required to do?** No specific information has been requested in the RFA; monthly reports on acquisitions, total size, and amount of use will be generated for internal purposes.	L
I.5	**Billing and Collections**	**Will there be charges for use of the resource?** There will not be a charge for accessing data within our resource, nor for invoking services we provide. However, users will be required to bear costs associated with download and any custom computations they want to perform.	
J. **Standards, Regulatory, and Governance Concerns**			
J.1	**Applicable Standards** > mature standards = decreased costs	1. **How many different standards will the resource have to support?** Descriptive metadata standards, data standards for different types of light microscopy and electron microscopy data, ontologies or controlled vocabularies for anatomy, imaging, cellular components, gene/protein names. 2. **Do these standards exist?** Some do. a. **If not, is the resource expected to lead their development?** The RFA specifies that the resource is to use relevant standards, but that it is not responsible for the creation of the standards. b. **What is the plan for accepting data while standards are in development?** Data will be accepted as soon as the infrastructure is ready, regardless of the state of the standards. Human curators will review all metadata to avoid common problems like cryptic abbreviations and nonstandard usage of terms. c. **If so, are the standards mature?** With the exception of descriptive metadata. 3. **Are the data validators and converters available for the standards, or do they have to be developed?** No, they have to be developed as per the RFA. 4. **What is the plan for "retrofitting" data that have been uploaded without the standards in place?** Data sets will be tagged accordingly, but unless specifically requested to do so, data will not be re-curated absent automated tools to do so. 5. **How frequently will the standards update?** The first release of a standard will be subject to extensive revision and so the standards will not be implemented until vetted by the community. The INCF standards review and endorsement process will be helpful here. 6. **Do the standards require spatial transformations?** Some data may be aligned to a common coordinate system to spatially align it with other data. Transformation coordinates and perhaps aligned files (depending on the volume of this type of data) will be stored. 7. **How many file formats will be supported?** Resource will be built around open file formats for large images using the Bioformats recommendation. The user will ensure that their data are in the required format. 8. **Is there an open file format available?** Yes. See previous item.	H

continued

TABLE 5.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
J.2	**Regulatory and Legislative Environment** > regulation = increased cost	1. **What laws and regulations cover the data and operation of the resource?** The resource is expected to have a large user base in Europe, so the resource will be General Data Privacy Regulation compliant. The website will meet accessibility requirements of the Americans with Disabilities Act and the institution. The resource will not include human-subjects data. 2. **Is the resource covered by an open-records act?** No.	L
J.3	**Governance** > outside governance = increased costs	1. **Does the resource need to maintain an external advisory board (EAB)?** There is no requirement for an EAB. 2. **Does the resource set policy for itself, or is it part of a larger organization?** Subject to BRAIN Initiative polices; otherwise, policy set by the resource.	L
J.4	**External Consultation** > consultations = increased time = increased costs	1. **Will external stakeholders be consulted for initial design?** Yes, outreach ahead of designing our website and services will be conducted. 2. **Will external stakeholders be consulted on an ongoing basis?** Yes, agile user testing for all new features will be employed.	M

[a] The website for Euro Bioimaging is https://www.eurobioimaging.eu/, accessed January 11, 2020.

6

Applying the Framework to a New Data Set

Per the statement of task, the cost-forecasting framework was applied to a second scenario, in this case, to the development of a new data set in a State 1 (primary research) platform.

USE CASE 2: ESTIMATING COSTS ASSOCIATED WITH A PRIMARY RESEARCH DATA SET

The cost-forecasting framework is applied to a proposed State 1 (primary research) data platform. The study committee applied the framework as might a young investigator (see Box 6.1). Box 6.2 demonstrates the logic introduced by the forecaster who, although enthusiastic, might be less experienced and unaware of available resources.

Applying the Framework to Use Case 2

Using the forecasting steps in provided in Table 4.1, the forecaster (in this case, the researcher) begins to construct the cost forecast.

Step 1. Determine the type of data resource environment, its data state(s), and how data might transition between those states during the data life cycle.

The forecaster examines the request for application (RFA) for requirements related to data management. Comparing the RFA requirements with the descriptions of the data states in Chapter 2, the forecaster determines this will be a State 1 (primary research) platform for her laboratory's use. However, the forecaster also plans to transfer the data to a State 2 active repository. Funding for transfer activities between platforms will also be considered.

Step 2. Identify the characteristics of the data (Chapter 4), data contributors, and users.

In light of needs, goals, and RFA requirements, the following preliminary assumptions about the data are made that will be refined throughout the conduct of the cost forecast.

BOX 6.1
The Use Case 2 Forecaster

A young investigator is exploring functional magnetic resonance imaging (fMRI) as a measuring technique for determining anatomical correlates for cognitive decline in patients diagnosed with Alzheimer's disease. The investigator is deciding whether to use conventional methods or to try new multiband imaging techniques in development. The researcher/forecaster explores several funding options, finding that the National Institute on Aging (NIA) has announced funding. NIA requires that all genomic data on Alzheimer's disease will be deposited in one of their databases, but she is not sure whether the policies cover her neuroimaging data. She is aware, however, that changes to the National Institutes of Health (NIH) data management and sharing policies are coming. If implemented, details regarding data stewardship and how and when data will be made available will need to be provided.

BOX 6.2
A Demonstration of Information Gathering to Inform Use Case 2

A cost forecaster, in this case a researcher in a primary research environment, gathers information prior to preparing a proposal (see Box 6.1). She first tries to identify the state of the art in fMRI imaging in Alzheimer's disease through the literature but wonders if there are data sets in the public domain that might be used for exploratory work. Not knowing where to look, she conducts an online search using the keywords "Alzheimer's disease + data + fMRI", which returns mostly articles. She has heard of and searches the Alzheimer's Disease Neuroimaging Initiative (ADNI),[a] but it does not have fMRI data. OpenNeuro has no publicly available fMRI data. She is not aware of the NeuroImaging Tools and Resources Collaboratory (NITRC)[b] or the Neuroscience Information Framework (NIF),[c] two NIH Blueprint for Neuroscience Research[d] initiatives that could point toward potential resources (e.g., NeuroVault,[e] the Open Access Series of Imaging Studies (OASIS),[f] or the 1000 Functional Connectomes Project).[g] A Google data set search lists only epidemiology studies. The Human Connectome Project[h] has a resting state fMRI longitudinal study under way, the Alzheimer's Disease Connectome Project, but the data will not be available for several months.[i] She reads Alzheimer's studies in the literature and hopes that a few made data available or referenced public data sets. She concludes that no suitable public data are available. Had the researcher/forecaster some way to know about available resources, she might have found data sets to inform her research direction, strategies, and her data management plan. This knowledge might have yielded cost savings and informed her, for example, of existing data sets that might have been aggregated or even that her study might not need to be conducted at all.

When considering how to share her neuroimaging data when the study is complete, the forecaster/researcher wonders if the NIH pilot program with figshare[j] is an option. She consults a data librarian at her institution for help designing her data management plan. The librarian searches the list of repositories on the National Library of Medicine's website[k] and sees that the OpenfMRI database (now OpenNeuro) takes fMRI data. Figshare has no specific requirements for formats or metadata, while OpenNeuro requires her data to be in Brain Imaging Data Structure (BIDS) format.[l] In either case, she will have to deidentify her data to publish them. The data librarian notes that while it may be easier and perhaps less costly in the short term to publish in figshare, data shared through a more specific repository such as OpenNeuro are aggregated with similar data and generally formatted to a common standard, likely giving the data greater value and visibility. Further, domain-specific repositories tend to have supporting software for upload and analysis. In fact, the researcher finds that the BIDS community is building a series of applications that significantly lower the cost of use.

BOX 6.2 Continued

The researcher/forecaster decides to deposit data in OpenNeuro and prepares her data management accordingly. OpenNeuro runs a validator that ensures format compliance, so it will be the researcher's responsibility to ensure data conform to the BIDS format. She may or may not be aware that there is a metadata specification designed for neuroimaging (Maumet et al., 2016) that will help guide the description of data and that harmonize specific variables with other data. Although not required by OpenNeuro, the researcher/forecaster is aware that rich metadata are critical for data reuse, not only by others but in her own laboratory as well.

By submitting data to a State 2 repository, data stewardship is transferred to the repository. The researcher can continue to benefit from the data, although they have been processed to comply with the Health Insurance Portability and Accountability Act (HIPAA) requirements. The researcher must decide whether to preserve the original data for the long term and about the disposition of other unpublished data related to the study. Decisions must also be made regarding how preserved data will be stored long term (i.e., by herself, or through institutional or commercial cloud resources). After consulting with a data librarian and her information technology (IT) department, the researcher decides against storing preserved data long term herself because institutional or cloud services can ensure the data are backed up appropriately and migrated to new platforms as necessary. Her institution has contracts with a cloud provider that provides generous storage allowances, but the cloud is not HIPAA-compliant. She therefore contracts with institutional IT services for long-term data management. The data librarian provides the researcher with a metadata template to ensure that the data can be retrieved reliably and that critical information about privacy and data ownership are documented.

[a] The website for the ADNI is http://adni.loni.usc.edu/, accessed April 15, 2020.

[b] The website for NITRC is https://www.nitrc.org/, accessed April 15, 2020.

[c] The website for NIF is https://neuinfo.org/, accessed April 15, 2020.

[d] The website for the NIH Blueprint for Neuroscience Research is https://neuroscienceblueprint.nih.gov/, accessed April 15, 2020.

[e] The website for NeuroVault is https://neurovault.org/, accessed April 15, 2020.

[f] The website for OASIS is https://www.oasis-brains.org/, accessed April 15, 2020.

[g] The website for the 1000 Functional Connectomes Project is https://www.nitrc.org/projects/fcon_1000/, accessed April 15, 2020.

[h] The website for The Human Connectome Project is http://www.humanconnectomeproject.org/, accessed April 15, 2020.

[i] The website for The Alzheimer's Disease Connectome Project is https://humanconnectome.org/study/alzheimers-disease-connectome-project, accessed April 15, 2020.

[j] The website for Figshare is https://datascience.nih.gov/news/nih-funded-researchers-invited-use-nih-figshare#:~:targetText=NIH%20Figshare%20is%20a%20one,.com%2Ff%2Ffaq%20, accessed April 15, 2020.

[k] The website for the National Library of Medicine is https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html, accessed April 15, 2020.

[l] The website for the BIDS is https://bids.neuroimaging.io/, accessed April 15, 2020.

Data Characteristics (Section A, Appendix E)

- The data are moderate in size: gigabytes (GB) per individual data set (several mature packages currently support fMRI).
- There are a moderate number of files and moderate size of individual files.
- Sizes of data sets will be stable over the life of the project.
- There are multiple neuroimaging modalities.
- The data are complex.
- There are significant metadata requirements.
- Data will come from a single contributor.

Data acquisition costs can be estimated because the number of subjects will be known ahead of time through institutional approvals. If the researcher decides to use a newer technology (e.g., multiband imaging), data sizes will increase fourfold to fivefold and the computational methods for processing and analyzing the data are less well known. In that case, the raw k-space data[1] will be kept available for reprocessing as new algorithms and approaches emerge.

As the forecaster, at this point, is only estimating the costs for her own use of the data, she skips the questions regarding the user community (Section F, Appendix E) but does keep in mind that the data may be of value to others in the future.

Step 3. Identify the current and potential value of the data and how the data value might be maintained or increased with time.

Perceived value is difficult to predict. However, all data sets underlying the results of a study will be made public so that the data can be inspected and reanalyzed. The availability of public data sets may also encourage technology development if she chooses to use more advanced techniques. As outlined in Box 6.2, if the data are well annotated and prepared according to community standards, they might be an important source of information and data for designing future studies.

Step 4. Identify the personnel and infrastructure likely necessary in the short and long terms.

Based on consideration of State 1 (primary research) and activities necessary to prepare data for State 2 (active) as described in Tables 2.1 and 2.2, respectively, the forecaster identifies the relevant major activities. The project objectives, informed by the RFA, the relevant activities, and personnel necessary (based on Table 2.1) are listed in Table 6.1.

Step 5. Identify the major cost drivers associated with each activity based on the steps above, including how decisions might affect future data use and its cost.

Table 4.2 is consulted to understand the likely important cost drivers for a State 1 resource, and the cost-driver template in Appendix E is filled in too (see Table 6.2 shown after the discussion of the use case). In this application of the framework, the guiding questions in Chapter 4 and the template about cost drivers are not all applicable, and so the forecaster revises the template to help delineate costs and decision points in as complete a manner as possible.

The relative costs related to data acquisition for this use case are straightforward to predict using the cost-forecasting framework. Relative costs associated with cost drivers identified in Table 4.2 are provided below based on the assessment made while filling out Table 6.2. In a real-world application of the cost-forecasting framework, these costs would be quantified with the help of State 2 (active) repository resources.

[1] K-space data are arrays of numbers that represent different spatial frequencies of the image.

- A: Content → Likely **low-medium**
- B: Capabilities → Likely **low**
- C: Control → Likely **medium**
- D: External Context → Likely **low**
- E: Data Life Cycle → Likely **low-medium**
- F: Contributers and Users → Likely **low-medium**
- G: Availability → Likely **low-medium**
- H: Confidentiality, etc. → Likely **medium**
- I: Maintenance and Operations → Likely **low**
- J: Standards, etc. → Likely **medium-high**

TABLE 6.1 Map of the Use Case 2 Scenario to Data States, Activities, and Subactivities

Project Objectives and Tasks	States, Activities, and Subactivities[a]	Personnel
1. Review of the literature and publicly available resources leads to a proposal to assess the feasibility of fMRI measurement techniques for this purpose.	I.B.1	Researcher, data scientist, software engineer, research domain project manager, policy specialist, administrative staff
2. Consider various funding sources and determine that potential funders expect collected data to be publicly shared.	I.A.1, I.B.2	Researcher, records management specialist, data scientist, data librarian, education specialist, policy specialist, software engineer, research domain project manager, administration staff
3. Assess suitability of existing repositories for the ultimate data deposit. Outline in data management plan the management and sharing approaches and costs estimates while data are under her stewardship. Consent methods for sharing data described.	I.B.3., I.B.4, I.B.5	Researcher, data scientist, software engineer, research domain project manager, policy specialist, administrative staff
4. Consider available tools for collecting, processing, and validating data using community-accepted standards. Considers documentation and curation levels required.	I.A.2, I.A.3, I.C	Researcher, records management specialist, data scientist, data librarian, metadata librarian, education specialist, policy specialist, research domain project manager, research domain curator, software engineer
5. Data management processes are in place that maintain primary and derived data (given evolving technologies). Derived data may include data in deidentified form.	I.C.3	Researcher, metadata librarian, data scientist, research domain project manager, research domain curator, software engineer
6. Deposit data in chosen repository on a regular schedule or when all data collection and analysis are complete.	I.D	Researcher, research domain project manager, IT project manager, software engineer, data wrangler

[a] The activity numerals correspond with labels in columns of Table 2.1.

TABLE 6.2 Decision Points for Use Case 2

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
A. **Content**			
A.1	**Size (volume and number of items)** > size = higher costs	1. **What is the order of magnitude of data that will be produced?** GB. 2. **How large is an average data set?** Per subject ~ 10 GB (multiple scans over time). 3. **Are the data sizes likely to stay stable over the life of the project?** Yes. 4. **What is the total amount of data expected?** ~400 GB. 5. **How many individual files in a typical data set?** Hundreds. 6. **If the data are to be transferred to a repository for long-term management, is there a cost depending on size?** No. Data will be submitted to OpenNeuro, which currently does not have costs associated with these data. 7. **Are there publicly available data that can be used to augment these data or perform preliminary analyses?** No relevant data were found.	L-M
A.2	**Complexity and Diversity of Data Types** **>** complexity + diversity = higher cost	1. **How complex is the underlying structure of the data?** Complex-image data. 2. **How complex is the experimental paradigm that produced the data?** Standard fMRI block design. 3. **What sort of additional data are acquired along with the primary data?** Cognitive assessments, statistical maps, demographic data. 4. **How many different data types are being produced?** Multiple modalities. 5. **What are the relationships among these data types—for example, are the data correlated?** Not applicable.	M
A.3	**Metadata Requirements** > metadata amounts + type = higher cost	1. **How much metadata must be stored with the data to make them findable, accessible, interoperable, and reusable?** Basic descriptive metadata, imaging parameters, experimental metadata, processing metadata, anatomical metadata. 2. **How are metadata recorded?** In data file headers, in Neuro Imaging Data Model (NIDM), in laboratory notebooks, in BIDS manifests.	M
A.4	**Depth Versus Breadth** > breadth = higher cost	1. **Is this study part of a multicenter study?** No. 2. **How many institutions/collaborators are involved?** Not applicable.	L

TABLE 6.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
A.5	**Processing Level and Fidelity** > compression = lower cost	1. **Do the raw data need to be stored?** K-space data not stored for standard fMRI. Will likely store k-space data if multiband imaging used. 2. **Do processed data need to be stored?** Yes. Analyses are performed on the reconstructed data. 3. **Are there compression algorithms that can reduce the file size without compromising fidelity?** Data files are not that large, so compression not typically used. 4. **What kind of data structure requirements will the resource have?** No particular structure enforced by imaging center. Data submitted to OpenNeuro must be organized according to the BIDS standard. 5. **Is the data contributor or the repository responsible for any restructuring necessary?** Researcher is responsible for restructuring data transferred to OpenNeuro. 6. **How is the data structure verified?** BIDS validator will likely implement it within our imaging pipeline.	H
A.6	**Replaceability of Data** > replaceability = lower cost	1. **Are there existing data sets that might be used instead of gathering primary data?** Not to our knowledge. 2. **Are the data managed by an institutional repository?** Our imaging center provides primary storage. 3. **Are there copies of the data elsewhere?** Local copy of data kept on a workstation in laboratory. 4. **Can the data be easily recreated?** No. It would be expensive to retest subjects. Disease progression information would be lost.	L
B. **Capabilities**			
B.1	**User Annotation** > user annotation functions = higher cost	1. **How long does it take to annotate/segment a data set?** Processing does not take very long. 2. **Is the process largely manual or automated?** Analysis data annotation is fully automated; experimental and descriptive metadata is added manually. 3. **Are these annotations stored with the data?** They are in a separate file. 4. **Is the relationship (provenance) between the data file and the annotations recorded in the metadata?** No, the association is captured through file-naming conventions.	L
B.2	**Persistent Identifiers** type of identifier = potential costs	1. **What persistent identifiers are used when annotating these data (e.g., Open Researcher and Contributor Identifiers, Ontology IDs)?** None. 2. **How are these persistent identifiers accessed?** Not applicable.	L
B.3	**Citation** > citation functions = increased cost	1. **Are the contributors to the production of a data set recorded in the metadata?** No. 2. **Is there a plan to submit the data to a repository that supports data citation?** Yes.	L

continued

TABLE 6.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
B.4	**Search Capabilities** > advanced search may lead to decreased cost	1. **Does the platform where the data are stored provide any search functions?** Just the native functions of the storage system (search on file name, creation date, owner, etc.). 2. **Was a search performed to locate data sets that might be relevant to this study?** Yes. 3. **What tools were used?** OpenNeuro; PubMed.	L
B.7	**Data Analysis and Visualization** > services = higher cost	1. **What type of data visualization tools are required?** Interactive viewing of images and 3D volumes; visualization of statistical maps. Freely available open-source tools used. 2. **What types of other data operations need to be supported?** Processing pipelines for the data; signal-extraction tools. 3. **Do these services require significant computational resources?** Moderate. 4. **Is there an explicit cost associated with compute resources?** Basic compute time is included with the fee paid to imaging center; many operations run locally on workstation.	L
C. **Control**			
C.2	**Quality Control** > quality control = increased cost	1. **What quality control processes are used?** Some automated and manual inspection of the data for issues such as motion artifacts. 2. **Does the public data repository have any quality control requirements?** OpenNeuro requires the data to be in BIDS format, so BIDS validator run.	L
C.3	**Access Control** > controls = increased cost	1. **What types of access control are required for the data?** Human-subjects data—institutional requirements for handling human-subjects data followed. Only qualified laboratory personnel can access the data. 2. **How is access to data managed, e.g., data access committees?** The principal investigator is responsible for managing access to the data.	L
C.4	**Platform Control** > platform restrictions = increased cost	**Are there restrictions on the type of platform that must be used for storing or analyzing the data?** Yes. Data infrastructure must adhere to our institution's security requirements for storing human-subjects data.	M
D. **External Context**			
D.1	**Resource Replication** > replication = increased cost	**Is there a requirement to replicate the information resource at multiple sites (i.e., mirroring)?** The imaging center backs up primary data to a local private cloud. Costs associated with replication are included in our fee to the imaging center.	L
D.2	**External Information Dependencie**s > external dependencies may or may not = increased cost	**Will the resource be dependent on information maintained by an outside source?** No.	L

TABLE 6.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
E. **Data Life Cycle**			
E.1	**Anticipated Growth** > growth = increased costs	1. **Is the total amount of data to be generated over the course of the project known?** Yes. 2. **Are there any factors that might affect the amount of data?** Not likely. The possibility that techniques used could increase data sizes has been accounted for, but approval gained to obtain data from a specified number of subjects and the processing pipelines, and so on, are well established.	L
E.2	**Update and Versions** > updates + multiple versions = increased cost	1. **Are multiple versions of the data created?** Yes, sometimes we have to reprocess individual subjects. 2. **If so, how are they managed locally?** Through the file names.	M
E.3	**Useful Lifetime** limited lifetime = decreased cost	1. **Are the data likely to have a limited period of usefulness?** Hard to predict; it will depend on the rate at which imaging technology evolves and whether new processing approaches are developed to compare our data to data collected by new instruments. 2. **Are there specific data retention institutional or regulatory requirements for these data?** Copies of all study data generally kept for at least 5 years after the study is completed.	L
E.4	**Offline and Deep Storage** > offline/ deep storage = decreased costs > transfers = increased cost	1. **For long-term storage of laboratory data, are there offline/deep storage resources available?** Yes, the institution runs a data archive for faculty research. 2. **Is there a plan for migrating laboratory data to a State 3 archive for long-term preservation?** Yes, data will be placed in the institutional archive after the study is completed.	M
F. **Contributors and Users**			
F.1	**Contributor Base** > number and diversity of contributors = increased cost	1. **Is the number of contributors known? If not, can it be estimated?** Just our laboratory members. 2. **Are all the data originating from the same source (e.g., a single instrument or a single organization)?** Yes.	L
F.2	**User Base and Usage Scenarios** > access and diversity of users = increased cost	1. **How many users will likely access the data?** Laboratory members (currently six). 2. **What will be the frequency of access?** Data accessed daily during the study and processing phase. 3. **How will users access the data?** Necessary compute infrastructure is available—the data will be on local machines. 4. **Will the resource be building analysis tools?** Yes, customized pipelines for processing our data, based on open-source toolkits, are built. 5. **How many different types of users must be supported?** Not applicable.	L

continued

TABLE 6.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
F.3	**Training and Support Requirements** > training + services = increased cost	1. **Is special training required for data upload to the repository?** Yes. 2. **What form will the training take?** Online tutorials and workshops. 3. **How long will this training take?** We will attend a training workshop on BIDS. 4. **What is the skill level required for data wrangling?** Moderate knowledge of neuroimaging and computer skills.	M
G. **Availability**			
G.1	**Tolerance for Outages** > reliability = increased costs	**What is the tolerance for outages of the resource?** Access to the data reliably is necessary. Will maintain adequate backups and system performance; scheduled outages for system patches and upgrades are tolerable.	M
G.4	**Local Versus Remote Access** > cloud could lead to increased costs	1. **Does the resource require that any data be shipped via physical media?** No, that is not likely. We have adequate bandwidth to transmit our data where required. 2. **Will commercial clouds be used?** No, not for primary storage.	L
H. **Confidentiality, Ownership, and Security**			
H.1	**Confidentiality** > confidentiality = increased cost	1. **Will any of the data require special protections?** Yes, they are human-subjects data. 2. **Are there any audit requirements for those who have accessed or downloaded the data?** No, we expect no users outside of laboratory staff.	M
H.2	**Ownership** > ownership = increased costs	1. **Do rights to use the data have to be negotiated with collaborators, institutions, commercial entities, or funders?** No. 2. **Will all data be released under the same license, or will different permissions be assigned to different data sets?** Data will be released under the license used by OpenNeuro. 3. **Will data submission agreements be necessary?** No.	L
H.3	**Security** > security = increased cost	1. **What types of security measures must be taken to protect against loss or corruption of data?** Standard practices will be used. 2. **Do these measures require using protected computing, storage, or networking platforms?** Yes.	L
I. **Maintenance and Operations**			
I.1	**Periodic Integrity Checking** > integrity checking = increased cost	1. **What processes will be put in place for checking the integrity of the hardware, software, and data?** We do not have any specific processes for this. 2. **How frequently will these checks be performed?** Not applicable.	L

TABLE 6.2 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
I.2	**Data-Transfer Capacity** > data-transfer upgrades = increased cost	**Will the bandwidth available be sufficient for the data sizes and rates required for transfer/access?** Yes. Campus connectivity recently upgraded. No problems anticipated.	L
I.3	**Risk Management** > risk mitigation = increased cost	1. **Will the researcher be solely responsible for risk mitigation?** Yes 2. **Is a response plan for unexpected termination required?** No	H
I.4	**System-Reporting Requirement** > system reporting-requirements = increased costs	**What types of system reporting will the resource be required to do?** None.	L
I.5	**Billing and Collections**	**Will there be charges for use of the resource?** No. All laboratory members have free access.	
J. **Standards, Regulatory, and Governance Concerns**			
J.1	**Applicable Standards** > mature standards = decreased costs	1. **How many different standards will be needed for the data?** Will use BIDS and NIDM along with standard registration tools to a common coordinate space. 2. **Do these standards exist?** Yes. 3. **Has the researcher worked with the standards before?** Yes. 4. **Are the standards mature?** Yes. 5. **Are tools (e.g., data validators and converters) available for the standards, or do they have to be developed?** Yes. 6. **How frequently will the standards update?** BIDS is a fairly mature standard. It is currently on version 1.2.1. 7. **Do the standards require spatial transformations?** Yes. 8. **How many file formats will be supported?** Digital Imaging and Communications in Medicine used. 9. **Is there an open file format available?** Yes. Neuroimaging Informatics Technology Initiative.	H
J.2	**Regulatory and Legislative Environment** > regulation = increased cost	1. **What laws and regulations cover the data and operation of the resource?** HIPAA. 2. **Is the resource covered by an open-records act?** Not applicable.	L
J.3	**Governance** > outside governance = increased costs	1. **How are decisions regarding data use managed?** Not applicable, no use outside the laboratory (i.e., no collaborators). 2. **Is a formal data-sharing agreement in place among the collaborators?** Not applicable.	L

Decisions made in the project planning stage, and the information resources available to the researcher during that planning, can influence the overall project costs, the study outcomes, and future data curation and preservation. For example, given that data might be transferred to a repository that has submission requirements, additional data preparation costs may be incurred. If the forecaster/researcher uses no formal data management software in the laboratory, a decision can be made to include additional costs in the budget to account for the effort. Funds could be requested for a data manager or wrangler to manage the data and set up the necessary infrastructure to adhere to data formatting standards. Automated pipelines could also assist transfer to a State 2 active repository on a regular basis. Cost to implement those pipelines may be greater up front but could also save many human hours over the duration of the project.

Because an individual forecaster, in this case a primary research environment researcher, cannot be responsible for estimating all costs for data management in perpetuity, the goal in applying the forecasting framework should be to estimate costs incurred during data acquisition and stewardship while they are in the researcher's control (i.e., the costs incurred while data are in State 1). However, the forecaster needs to be aware of requirements for long-term stewardship and be ready with the resources required (e.g., time, money, personnel) to prepare data for transfer to a State 2 (active) repository if to be shared or, if not, to a State 3 repository for long-term preservation.

Step 6. Estimate the costs for relevant cost components based on the characteristics of the data and information resource.

In a quantitative cost forecast, the costs for the activities in the previous section would be quantified for each of the major cost components (e.g., Box 3.2). As noted previously in the report, quantifying costs is dependent on numerous case-specific factors such as the objectives for the information resource, the personnel and infrastructural resources available to the forecaster, and host institution requirements. In a real cost forecast, all of these would be considered to arrive at monetary values.

REFERENCE

Maumet, C., T. Auer, A. Bowring, G. Chen, S. Das, G. Flandin, S. Ghosh, et al. 2016. Sharing brain mapping statistical results with the neuroimaging data model. *Scientific Data* 3:160102. https://doi.org/10.1038/sdata.2016.102.

7

Potential Disruptors to Forecasting Costs

In this report, a disruptor is anything that may cause radical changes to the ways research is conducted and data are collected, used, archived, or preserved. Disruptors may be positive or negative and may raise or lower the cost of data management and preservation. This chapter considers some of the future developments and disruptors in data technologies and data science that may reduce or increase data costs in the next 5 to 10 years. Recent examples of disruptors affecting biomedical research are the widespread use of high-resolution imaging instruments (e.g., electron microscopes [Courtland, 2018; Guzzinati et al., 2018]); the decreasing cost of sequencing, the rate of which even surpasses Moore's law (Wetterstran, 2019); and the advent and accessibility of cloud storage and computing. The use of high-resolution imaging instruments alongside the decreasing cost of sequencing has resulted in the ability to collect huge volumes of data, while cloud computing has resulted in new ways to store, aggregate, and analyze data. Cloud computing, however, has also resulted in new or different costs that many researchers do not yet fully understand. Those costs must be considered in the context of potential gains in capacity or functionality.

There is no way to fully anticipate factors that might radically affect the costs of future data preservation, archiving, and use. This chapter focuses on certain emerging challenges spanning different dimensions, including

- biomedical data volume and variety,
- advances in machine learning and artificial intelligence (AI),
- changes in storage technologies and practices,
- future computing technologies,
- workforce-development challenges,
- legal and policy disruptors, and
- human-subjects research.

This illustrative list gives examples of disruptors likely to affect costs of data management and use in the next 5 to 10 years. Although quantifying the contributions of these disruptors to long-term data-preservation costs is beyond the scope of the study committee's charge, these issues warrant attention so that associated cost changes can be anticipated and minimized or exploited to some extent.

BIOMEDICAL DATA VOLUME AND VARIETY

The biomedical sciences have generated steady streams of data for decades, but there have been sudden orders-of-magnitude increases in data collection in particular domains. Over the past decade, emerging and evolving data from sources such as next-generation sequencing, correlated light and electron microscopic imaging, and multiscale high-performance computing simulations have led to large increases in the volumes of data that can be collected and have pushed biomedical research into the realm of "big data." There are many centralized core facilities serving research communities, such as the National Center for Microscopy and Imaging Research,[1] that can produce extremely large live data feeds. Many researchers and laboratories have already acquired or will acquire volumes of data that cannot, at present, be completely analyzed.

Imaging tools are undergoing a revolution, and new microscopy technologies produce ever more detailed images, leading to a data-size explosion. Both large centers and conventional research laboratories are exploring imaging regimes that cross fundamental length scales: from tens of centimeters to angstroms. These image data sets are on the order of tens of terabytes per project and accumulate petabytes of data per year per instrument. The scales of such data are critical to advance further understanding of key biological processes. Other areas of biomedical research are experiencing similar growths in data volume, including genomics, where next-generation sequence data are introducing unprecedented challenges in data management, organization, and analysis. Electronic medical records and small data collected at individual laboratories that must be aggregated with existing data sets also present challenges to efficient data analysis in the quest for actionable knowledge.

In the foreseeable future, the biomedical research community will experience spurts in data growth that will tend to either (1) add a dimension to the data space or (2) extend a dimension by an order of magnitude. This growth may be related to, for example, the following:

- *Gene sequencing*. Moving from sample-level sequencing, to cell-level sequencing (representing a new dimension), to "cell in context" sequencing (representing yet another new dimension—the cell location in terms of both its position in the body and surrounding cell types and other structures). The shift from per-sample sequencing to per-cell sequencing results in 1,800 times as much data per subject (Ameur et al., 2011).
- *Population size (extending a dimension)*. Moving from a sample size of 100 or 1,000 to 1,000,000.
- *Time dimension*. Images of the same cell or piece of tissue over time, or gene sequencing the same individual cell or tissue at multiple points in time (e.g., to establish a "healthy" baseline and then watch precursors of disease develop).
- *Sequencing depth*. Going from a coverage depth (i.e., how many times, on average, each location in a sequence of interest is sequenced) of 30 times or so to 100 times and more to find rare transcripts or mutants (such as in RNA-Seq).
- *Reanalysis of existing images or reimaging samples*. New techniques or methods allowing greater resolution or precision.

The ever-expanding data-collection capability continues to impose challenges to biomedical science applications owing to its volume, velocity, variety, veracity, and variability (e.g., Ristevski and Chen, 2018) but promises transformative advances. However, the size and complexity of those data sets are overwhelming existing repository structures and are pushing the boundaries of the current capabilities of technologies to access, manage, integrate, and analyze them at scale. Increasingly, biomedical data are too voluminous for a single platform, too unstructured for a traditional database system, or too continuous to store for analysis at a later time. More than ever, such challenges or possible cost increases associated with big data must be considered in the context of additional value and novel opportunities for scientific understanding at different scales.

[1] The website for the National Center for Microscopy and Imaging Research is https://ncmir.ucsd.edu/, accessed December 13, 2019.

ADVANCES IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

The above-mentioned volumes of data shifted the bottleneck in biomedical sciences from data availability to the generation of insight from data. This shift has resulted in increased use of the newest advances in machine learning and AI in biomedical sciences. A simple search for the term "machine learning" in medical literature on Semantic Scholar[2] reflects this increasing use. Continuous automatic annotation of data and metadata generation is one growing use of machine learning. Any biomedical researcher using big data needs to be able to reduce the size and complexity of the data while adding more meaning and value to them with the addition of richer metadata. This trend is already evident in many parts of the field (Shah et al., 2019; Zhu et al., 2019). Automated metadata generation using techniques that allow regular updates to volumes of data increases the need for active and more costly storage approaches. Automated data analysis requires programmatic access to data. Increased use of services that enable data search and access as well as findable, accessible, interoperable, reusable (FAIR), and responsible use will likely result in the need for additional or new human resources and talent development.

AI offers the potential to lower costs by automating ethical and regulatory processes. There is growing use of deep learning as an approach to AI in biomedical science, although many challenges, including interpretability (Xu and Jackson, 2019), remain. This creates the need for AI-ready solutions and systems in which data curation and storage are no longer independent of their analyses. For instance, the Broad Institute has developed and tested a Data Use Oversight System (DUOS)[3] meant to reduce the person-hour costs of data access committees. These committees are staffed by highly trained professionals who consider requests to access data for secondary research purposes in light of the restrictions on data use that are built into consent forms or other governance commitments. DUOS semi-automates this process by building ontologies into both consent forms and secondary data access request forms. Other costs associated with responsible data use and sharing might be lowered through use of automated processes that have the additional benefit of placing more control over data in the hands of participants—for example, an automated process by which participants with revised preferences for secondary use or sharing of their data could unilaterally change the access policies that apply to their data and could lower costs associated with manual tracking and updating of participant preferences.

On the other hand, AI has the potential to be a negative disruptor that drives up costs associated with ethics and regulation by upending assumptions about which data are nonidentifiable[4] or deidentifiable.[5] For instance, AI makes it easier to re-identify facial and cranial images. There are also concerns about AI-based hacking, upending assumptions about the security of data.

Changes in Storage Technologies and Practices

Biomedical big data from many sources today place different constraints on data management, from their acquisition and movement, to storage and access. Greater constraints on physical data storage are posed by the need for bandwidth and computing to move and analyze data. The doubling of storage capacity represented by Kryder's law (Walter, 2005), which has slowed down over the past decade owing to different approaches to storage and cloud computing, is likely to increase storage costs over time. These changes could easily affect free and infinite storage by adding charges to computing and networking around the data, most of which are not built for multicloud solution scenarios.

As discussed in the previous section, there will likely be a shift away from merely storing data toward approaches that allow continuous extraction of value from data using machine learning and AI techniques. This shift will affect raw data coming off the sources in active and passive archives and include models and knowledge

[2] The website for Semantic Scholar is https://www.semanticscholar.org/search?q=%22machine%20learning%22&sort=relevance&fos=medicine, accessed December 13, 2019.

[3] The website for the DUOS is https://duos.broadinstitute.org/, accessed December 13, 2019.

[4] See 45 C.F.R. § 46.102(f)(2).

[5] See the Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 100 Stat. 2548 (1996).

generated from data. Connections to open knowledge networks that enable the search and access of data generate a new set of requirements around how data are stored and used.

These factors have influenced a major shift in how data storage has been managed in the past decade, and new technologies will likely continue to stress the capabilities of data storage systems. Today's storage systems are multifaceted and software driven in a way that optimizes storage performance for various uses. Next-generation storage systems are being built around AI-integrated approaches that enable users to monitor storage performance for its use rather than bitbucket costs. Although the cost implications of these approaches and their use are not trivial and are difficult to estimate for scientific users, the storage systems have already led to major cost-cutting efficiencies for enterprise use. Software-defined and hybrid storage approaches (Donald, 2019) are also potential areas that will disrupt how scientific data are stored and managed as AI-ready entities.

Future Computing Technologies

In addition to the disruption in data generation and intelligent value-driven storage, the analysis and computing costs of data will have a large role in the cost structure over the next 5 to 10 years. Further advances in moving computation to data are likely to continue to shift the cost of data, from download and storage (e.g., from the cloud) and toward computing. New cloud cost models will be the determinant factor for the effects of this shift on the overall cost of data. This concept might be encapsulated as a shift away from data sets to data streams. In addition, emerging edge computing[6] architectures are expected to disrupt central repository-driven computation and privacy strategies to a more distributed mode for the generation of insight from data, especially on biomedical data sensed via "Internet of Things" devices that capture data in real time for health monitoring and alert-generation scenarios. Last, the increasing number of non–von Neuman architectures and machine learning accelerators will require careful consideration regarding co-locating data and computing based on the models that need to make use of co-dependent composable services at the digital continuum.

DEVELOPMENTS WITH POTENTIAL COST SAVINGS

Of all existing technologies, a few may reduce costs within shorter time frames. These range from new approaches to managing cloud use to analyzing data through service and managing the integrity of data. A few examples include the following:

- Scalable search approaches and libraries that can combine many types of searches on heterogeneous data systems across distributed storage platforms. Use of such searches could have implications for costs associated with metadata. Elasticsearch[7] is an example of a service offering this type of capability.
- Blockchain is a chain of timestamped hashes of information that enable a number of applications to preserve data integrity and ownership as well as lead to a new credit-economy discussion around data (e.g., exemplified by LunaDNA[8]). The potential uses for this technology in the biomedical sciences range from patient-controlled data access to researchers managing who has access to their data and for how long.
- Open Knowledge Networks are community efforts to develop national-scale data infrastructure (see, e.g., OSTP, 2018). Universities, funders, and companies are working on knowledge networks that would provide greater and richer access (e.g., semantic) to data through more accessible interfaces (e.g., natural language). Arguably, such networks have not yet achieved widespread adoption, but they merit some examination for potential impact on costs.
- The National Science Foundation (NSF) CloudBank is a collaborative NSF award (with a $5 million grant)[9] to make accessing the cloud easier and less costly. It aims to be helpful for many stakeholders, including

[6] In edge computing, data are first processed at a center geographically closer to the data sources. The resulting smaller or compressed information is then sent to the cloud for computing. This process reduces latency periods.

[7] The website for Elasticsearch is https://www.elastic.co/, accessed December 13, 2019.

[8] The website for LunaDNA is https://www.lunadna.com/, accessed December 13, 2019.

[9] The website for the CloudBank award is https://www.nsf.gov/awardsearch/showAward?AWD_ID=1925001, accessed December 13, 2019.

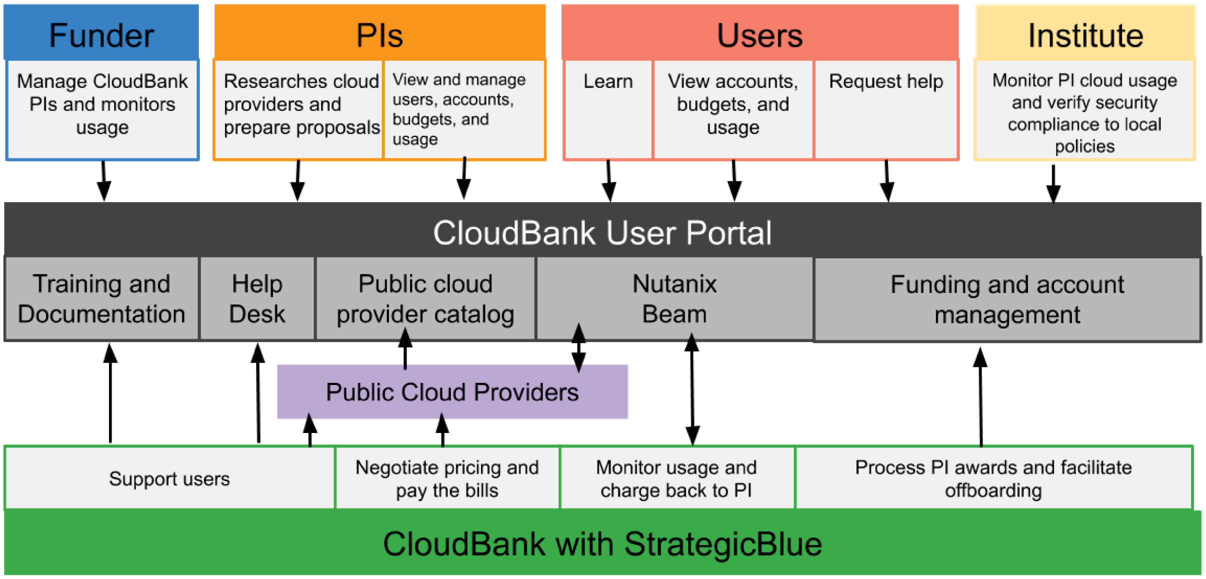

Funder
Manage CloudBank PIs and monitors usage
PIs
Researches cloud providers and prepare proposals
View and manage users, accounts, budgets, and usage
Users
Learn
View accounts, budgets, and usage
Request help
Institute
Monitor PI cloud usage and verify security compliance to local policies
CloudBank User Portal
Training and Documentation
Help Desk
Public cloud provider catalog
Nutanix Beam
Funding and account management
Public Cloud Providers
Support users
Negotiate pricing and pay the bills
Monitor usage and charge back to PI
Process PI awards and facilitate offboarding
CloudBank with StrategicBlue

FIGURE 7.1 An overview of CloudBank. SOURCE: Michael Norman, San Diego Supercomputer Center, presentation to the committee, September 12, 2019.

NSF program officers, researchers, students, and cloud providers. CloudBank will provide oversight of the cloud ecosystem. Because many challenges associated with cloud computing are related to account management and account monitoring, CloudBank is actively building methods to enable diverse users to manage their cloud credits through business operations functions and services (Norman, 2019; San Diego Supercomputer Center, 2019). This type of resource could greatly assist researchers in understanding the consequences of their choices using cloud-based storage or computing (see Figure 7.1).

WORKFORCE-DEVELOPMENT CHALLENGES

As has been described in Chapter 2, of all the resources required to make biomedical data useful for science, perhaps the most cost intensive is the human one. A major challenge to the biomedical research community, especially influenced by the above-mentioned disruptors, is the training and education of the current and next generation of biomedical scientists and a workforce that can effectively process and manage data in their different states. For this reason, workforce development is a disruptor to the cost structure in biomedical data archival, preservation, and access in the long term. The biomedical data community must work collectively to address the need for well-trained human talent.

The number of tools and techniques for working with biomedical data is increasing, and open access and cloud computing place those tools, as well as the data and infrastructures, within reach. But human talent, particularly in data science, is difficult to find. So far, biomedical scientists and researchers who are developing applications and models for big data have filled this role, but such individuals represent a small percentage of analytical talent. Data scientists are in high demand by those who can offer higher wages than offered through the public sector and academia. This disparity makes it difficult and expensive to attract and sustain an adequate workforce. Training biomedical data scientists who are well versed to take advantage of the emerging disruptive technologies in scientific applications is of critical importance to the future of biomedical data-driven research and knowledge advancement. However, no single group or strategy will be able to cover the full spectrum of educational requirements to comprehensively train biomedical data researchers. Approaches that take advantage of open, online teaching modules to train in-house experts on emerging data and technology trends may be a means to reduce the cost of talent generation.

LEGAL AND POLICY DISRUPTORS

The legal and policy environments—and the evolution of those environments—are another source of potential disruption that may affect costs. Some challenges to forecasting the costs of data curation, dissemination, and preservation arise as unintended consequences of U.S. science policy. For example, the NSF policy against cost sharing, intended to provide a level playing field for researchers from differently resourced institutions, may obscure cost information about data preservation and access that occurs after or as an adjunct activity to the central research activity. Federal Statistical Research Data Centers (FSRDCs) are an important dissemination mechanism for a set of confidential, high-value, population-level biomedical data produced by the National Center for Health Statistics and other statistical agencies. FSRDCs are supported by universities ("institutional partners") and member federal statistical agencies. There are currently 30 secure enclaves across the United States, through which thousands of researchers access data. It is reasonable for science and health agencies to ask about costs to provide access via such data centers. NSF has provided funding for most of these enclaves, but that funding is explicitly intended to cover start-up costs and requires ongoing support from institutional partners or individual, externally funded research projects. However, NSF, in pursuit of its goal of fair access to external funding, has decided to not require applicants to demonstrate their plans for financial sustainability (and in fact prohibits prospective research data centers from including those plans in funding proposals). Thus, neither NSF nor the federal statistical agencies has any information about the ongoing costs to the institutional partners of maintaining this data-dissemination infrastructure—no one knows how much it costs to disseminate these data. This situation could become an additional disruptor to sustainability, especially as data and data infrastructure become increasingly large and complex.

Other disruptors generated as a result of changing legislation and policy are related to data privacy. Illustrative examples are developed in the next sections concerning, respectively, when data are considered to be "identifiable" and when and under what circumstances it is permitted or seen as appropriate to collect, store, or share data.

Data Identifiability

Current regulatory definitions of data and tissue "identifiability" are volatile. The Common Rule is a set of regulations in place since 1991 that applies, directly, to HSR conducted or funded by most federal departments and agencies and, indirectly, to virtually all academic (and some industry) HSR by institutional policy. It defines data as "identifiable" when "the identity of the subject is or may readily be ascertained by the investigator or associated with the information" (45 C.F.R. § 46.102(f)(2)). This definition of (non-)"identifiable" under the Common Rule is critical to how data (and tissue) are collected, preserved, and accessed. (The distinct but related concept of "deidentified" under the Health Insurance Portability and Accountability Act [HIPAA] carries similar consequences.) Research with existing data and tissue (whether originally collected for research, clinical, administrative, or other purposes) that meet the Common Rule's definition of "nonidentifiable" does not involve "human subjects" as the Common Rule defines that term, and therefore such research falls outside of the Common Rule, including its default rules requiring Institutional Review Board (IRB) review and informed consent. The rationale behind this policy was that the main risk of research that involves neither intervention nor interaction but only analysis of existing data is informational privacy; analysis of data that cannot be linked to an individual's identity does not pose such a risk. Historically, IRBs and other governance and compliance actors have considered genomic data not to constitute "identifiers" in and of themselves, without being linked to additional information.

From 2011 to 2017, federal regulators engaged in public notice-and-comment rulemaking to revise the Common Rule, whose substance had not been significantly changed since 1991. Among the most controversial proposals was altering the definition of "human subject" to include both identifiable and nonidentifiable biospecimens. That change would have defined research using existing tissue samples that were stripped of identifiers as HSR and therefore subject to IRB review and consent.

The rationale behind the proposal was twofold. First, a series of academic reidentification "attacks" demonstrated the possibility, under certain circumstances, of reidentifying genomic and a wide variety of other data (e.g., consumer and geolocation data) that were considered to be nonidentifiable (Narayanan and Shmatikov, 2008; El Emam et al., 2011; Gymrek et al., 2013; De Montjoye et al., 2013; Gambs et al., 2014; De Montjoye et al., 2015). These attacks cast doubt on the assumption that research with data considered to be nonidentifiable under the Common Rule does not implicate participant privacy. Second, some commentators were of the opinion that people have autonomy interests in controlling the use of their data for various projects, even if those data are never associated with them as individuals (Javitt, 2010; Mello and Wolf, 2010; Tarini and Lantos, 2013).

The proposed rule would apply only prospectively, so that researchers would not be required to recontact and consent those whose tissue comprises existing biobanks. Nevertheless, the research community was largely opposed to the proposal; in fact, every stakeholder category failed to receive public comments to support the proposal. As a result, the revised Common Rule does not include the proposed redefinition of identifiability.

The current Common Rule does, however, require federal departments and agencies to reconsider, within the first year of the new rule going into effect, and at least once every 4 years thereafter, both the definition of "identifiable" data and biospecimens and whether any "technologies or techniques" applied to biospecimens, such as whole-genome sequencing, should be considered to generate data that are necessarily identifiable. If regulators determine there is a need to alter the definition of "identifiable," then agencies are to develop interpretive guidance to achieve this goal. Similarly, if regulators determine that technologies determined to necessarily generate identifiable data shall be placed on a public list following a public notice-and-comment period. Under such circumstances, agencies could then issue guidance recommending that limited IRB review and broad consent be required for research involving those technologies without public notice or comment (Lynch and Meyer, 2017).

Agency guidance is not legally binding and, presumably, as with the proposal to change to the Common Rule itself, it would only apply prospectively. Still, processes developed or modified, implemented, and relied on for the collection, archiving, and secondary use of nonidentifiable data would likely be deemed no longer fit for purpose in a new regime under which best practice is to consider those data identifiable. New processes would need to be developed and implemented at both the State 1 (researcher) and State 2 (active repository) levels, which are costly endeavors. (State 3 [long term] is not included because the Common Rule applies only to research use of data to contribute to generalizable knowledge, not to the mere act of storing it. To the extent that State 3 does not include access as a key component, changes to the Common Rule's understanding of identifiability would not apply directly.) And because agency guidance development is not subject to the time schedules of public notice-and-comment rulemaking, regulated entities might have relatively little notice.

Permissible Data Collection, Storage, and Sharing

In reaction to events such as the Cambridge Analytica scandal (see Davies, 2015), regulators sometimes enact data protection laws that impose substantial burdens on regulated entities and do not always consider the impact on research. For instance, the EU General Data Privacy Regulation (GDPR),[10] which went into effect on May 25, 2018, enacted sweeping changes in how nonanonymous personal data, including data that might immediately or eventually be used in research, may be collected, stored, and disseminated. The GDPR has a global impact, applying to data collected from individuals residing in the EU at the time of data processing.

To date, the United States has had no such comprehensive privacy law. Instead, it has a patchwork of federal and state laws, such as HIPAA, the Common Rule, and the Family Educational Rights and Privacy Act.[11] The 2018 California Consumer Privacy Act (CCPA)[12] covers individual, identifiable data and went into effect on January 1, 2020. Although the CCPA (unlike the GDPR) obligates only for-profit businesses, and specifically excludes data that are already subject to federal privacy laws (e.g., HIPAA) and information collected for a clinical trial subject to the Common Rule, biomedical researchers increasingly collaborate with for-profit businesses around non-HIPAA data, including consumer wearables, mobile health apps, and genetic data from direct-to-consumer testing companies. The anticipated impact on research of the GDPR and the CCPA are approximately the same: both contain various exceptions for research (e.g., the right to erasure, or so-called right to be forgotten, has only

[10] The website for GDPR is https://gdpr-info.eu/, accessed December 13, 2019.

[11] See 20 U.S.C. § 1232g; 34 C.F.R. Part 99.

[12] See Assembly Bill No. 375, an act to add Title 1.81.5 (commencing with Section 1798.100) to Part 4 of Division 3 of the Civil Code, relating to privacy.

limited applicability to research data) while still regulating (some) research activity, and in both cases the actual impact on research remains unknown. In April 2019, U.S. Senator Edward J. Markey introduced into Congress the much more sweeping Privacy Bill of Rights Act[13] that closely resembles the GDPR and applies to "any person that collects or otherwise obtains personal information."

CHANGING UNDERSTANDING OF HUMAN-SUBJECTS POLICY

When research data are about human beings, a variety of laws, policies, and norms are likely to apply throughout the data life cycle. Compliance with each of these laws, policies, and norms has associated costs, often in the form of administration, tracking, and training. In State 1 (the primary research environment described in Chapter 2), data are captured initially through interaction or intervention with humans for research use, existing data may be collected for new research or nonresearch use (e.g., clinical or administrative purposes), or a research project might include both kinds of data capture.

Researchers may already be familiar with HSR activities and their legal requirements, as researchers bear at least some of the associated burdens. These activities include HSR ethics training for everyone engaged in the research and a variety of prospective reviews of the research protocol. IRBs review a research proposal or determine that it is either exempt HSR or not HSR (e.g., because it involves analysis only of existing data that include no personal information that could lead to the identification of an individual—nonidentifiable data). Other reviews might also apply. Data subject to HIPAA[14] might require a review by a Privacy Board (if the IRB is not also acting as the covered entity's Privacy Board) and a review by an institution's information security office. If research involves consent from or notice by the human subject, those materials and processes must be developed and, often, pretested for participant comprehension or operational feasibility. In some cases, consent processes will require creating new institutional infrastructure that has its own costs. For example, a biobank (i.e., a type of repository for biological samples) that uses a "front door" opt-in consent will have to train the relevant staff to solicit such consent. That consent process will then have to be integrated into the clinical workflow, and patients' enrollment status will need to be recorded and perhaps incorporated into the electronic medical record. Successful large-scale research projects such as biobanks often involve potentially extensive and costly participant incentives or engagement activities (e.g., "results" might be provided to participants for engagement or other purposes, or something else of value might be provided). Innovative consent methodologies may also carry costs. For instance, some large research projects—especially those such as biobanks, where data are collected under broad, rather than study-specific, consent—take a more or less "dynamic" approach to consent, in which participants are invited to change the way their data are used in response to changing circumstances. The ongoing communication of the project(s) status to participants and inviting and implementing their evolving consent preferences requires a significant investment of time and, to varying degrees, material costs. Another kind of consent—tiered consent—enables different participants to choose the degree of data collection, use, or sharing they authorize. Tiered consent may involve costly tracking to ensure adherence to those heterogeneous preferences. State 1 HSR activity costs are direct costs charged to the funder, while activity costs associated with accessing data in States 2 and 3 are indirect costs that are typically at least offset by grant funding.

The costs of HSR activities associated with State 2 (i.e., active repository) acquisition, aggregation, and support for access are less visible to researchers because they often are externalized onto repositories. As a result, those costs are more difficult for researchers to anticipate. When data are transferred to a repository, they may need to be deidentified (i.e., personally identifying information removed), consistent with HIPAA requirements, or otherwise anonymized or pseudonymized. Participants generally have a right to withdraw their data from a data set, which requires the repository to remove those data and update related resources as necessary. If secondary data use is restricted, for example, by the terms of consent, a data use agreement, prospective review (e.g., by a data access committee), or auditing might be required to enforce those restrictions. Often, some data in a data set are more sensitive than others, such that tiers of data access among users is necessary. The least sensitive data may be

[13] U.S. Congress, Senate, Privacy Bill of Rights Act, S.1214, 116th Congress, introduced April 11, 2019.

[14] Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 100 Stat. 2548 (1996).

openly accessible to any users, while other tiers of data are available only to certain people, under certain circumstances, or for certain purposes. Determining criteria for each tier of access and sorting the data accordingly could be laborious and therefore incur cost. The more sensitive the data, the more repositories might want to develop sandboxes[15] or enclaves where researchers can access and analyze those data but not remove them. The development of sandboxes and enclaves, and their periodic update in light of new research purposes or new technological requirements, imposes costs. A good example of many of these mechanisms is the National Institutes of Health (NIH) All of Us Research Program,[16] which plans to use three tiers of access, identity verification, a researcher code of conduct, voluntary prospective review of sensitive projects by a Resource Access Board, a "data passport" (to allow access to registered or controlled-access data sets) and sandbox, and retrospective data use audits, to preserve data privacy and security and participant trust. Merely developing these governance mechanisms required significant person-hours before actual data access began.

In State 3 (long-term preservation), the primary HSR activity is ensuring that data placed in a long-term archive continue to adhere to current legal and governance requirements. For instance, data considered or perceived as nonidentifiable when generated may later become identifiable (e.g., because additional information about the data sources becomes available or reidentification techniques are developed) or are redefined as identifiable under new laws or policies (see the section "Legal and Policy Disruptors" in this chapter).

OTHER POTENTIAL DISRUPTORS

There are many other potential disruptors that are not discussed in this report but that could affect long-term costs within the next 5 to 10 years or beyond. Examples include

- open data practices;
- long-term resilience of technology production;
- evolving requirements for cybersecurity (e.g., surreptitious cyberattacks to corrupt data; data misuse and theft that undercut support of repositories);
- influences of the FAIR data principles, open science, and Responsible Data movements, particularly of increasing acceptance and adoption of standards;
- transfer learning;
- investigating connected data that cross spatial and temporal scales and modalities;
- transitioning from needing specialized expertise to providing self-contained tools and resources;
- risks associated with third-party vendors (particularly if they capture a large share of the biomedical data market); and
- natural disasters that disrupt technology production in the long term.

Although the committee did not deliberate on the effects of those disruptors, they may warrant further attention by the biomedical research community.

REFERENCES

Ameur, A., J.B. Stewart, C. Freyer, E. Hagström, M. Ingman, N.-G. Larsson, and U. Gyllensten. 2011. Ultra-deep sequencing of mouse mitochondrial DNA: Mutational patterns and their origins. *PLoS Genetics* 7(3):e1002028.

Courtland, R. 2018. The microscope revolution that's sweeping through materials science. *Nature* 563:462-464.

Davies, H. 2015. Ted Cruz campaign using firm that harvested data on millions of unwitting Facebook users. *Guardian*, December 11.

De Montjoye, Y.-A., C.A. Hidalgo, M. Verleysen, and V.D. Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports* 3:1376. https://doi.org/10.1038/srep01376.

[15] A sandbox is a separate platform on which researchers can use tools to explore and experiment with data.

[16] The website for NIH's All of Us Research Program is https://allofus.nih.gov/, accessed December 5, 2019.

De Montjoye, Y.-A., L. Radaelli, V.K. Singh, and A.S. Pentland. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347(6221):536-539.

Donald, D. 2019. Major disruptions in data storage technology: What this shake-up means for the Enterprise. *insideBIGDATA*, April 5. https://insidebigdata.com/2019/04/05/major-disruptions-in-data-storage-technology-what-this-shake-up-means-for-the-enterprise/.

El Emam, K., E. Jonker, L. Arbuckle, and B. Malin. 2011. A systematic review of re-identification attacks on health data. *PLoS One* 6(12):e28071.

Gambs, S., M.-O. Killijian, and M.N. del Prado Cortez. 2014. De-anonymization attack on geolocated data. *Journal of Computer and System Sciences* 80(8):1597-1614.

Guzzinati, G., T. Altantzis, M. Batuk, A. De Backer, G. Lumbeeck, V. Samaee, D. Batuk, et al. 2018. Recent advances in transmission electron microscopy for materials science at the EMAT Lab of the University of Antwerp. *Materials (Basel)* 11(8):1304.

Gymrek, M., A.L. McGuire, D. Golan, E. Halperin, and Y. Erlich. 2013. Identifying personal genomes by surname inference. *Science* 339(6117):321-324.

Javitt, G. 2010. Why not take all of me: Reflections on *The Immortal Life of Henrietta Lacks* and the status of participants in research using human specimens. *Minnesota Journal of Law, Science, and Technology* 11(2):713-755.

Lynch, H.F., and M.N. Meyer. 2017. Regulating research with biospecimens under the revised Common Rule. *Hastings Center Report* 3(May-June):3-4.

Mello, M.M., and L.E. Wolf. 2010. The Havasupai Indian Tribe case—lessons for research involving stored biologic samples. *New England Journal of Medicine* 363(3):204-207.

Narayanan, A., and V. Shmatikov. 2008. Robust de-anonymization of large sparse datasets. *29th IEEE Symposium on Security and Privacy* 111-125.

Norman, M. 2019. Presentation to the Committee on Forecasting Costs for Preserving and Promoting Access to Biomedical Data, September 12.

OSTP (Office of Science and Technology Policy). 2018. *Open Knowledge Network: Summary of the Big Data IWG Workshop*. https://www.nitrd.gov/pubs/Open-Knowledge-Network-Workshop-Report-2018.pdf.

Ristevski, B., and M. Chen. 2018. Big data analytics in medicine and healthcare. *Journal of Integrative Bioinformatics* 15(3):20170030.

San Diego Supercomputer Center. 2019. UC San Diego, UC Berkeley, U Washington Announce ‘CloudBank’ Award. Press Release, August 8. https://www.sdsc.edu/News%20Items/PR20190808_CloudBank.html.

Shah, P., F. Kendall, S. Khozin, R. Goosen, J. Hu, J. Laramie, M. Ringel, and N. Schork. 2019. Artificial intelligence and machine learning in clinical development: A translational perspective. *npj Digital Medicine* 2:69.

Tarini, B.A., and J.D. Lantos. 2013. Lessons that newborn screening in the USA can teach us about biobanking and large-scale genetic studies. *Personalized Medicine* 10(1):81-87.

Walter, C. 2005. Kryder’s law. *Scientific American*, August 1. https://www.scientificamerican.com/article/kryders-law/.

Wetterstran, K.A. 2019. DNA sequencing costs: Data from the NHGRI Genome Sequencing Program. www.genome.gov/sequencingcostsdata.

Xu, C., and S.A. Jackson. 2019. Machine learning and complex biological data. *Genome Biology* 20:76.

Zhu, G., B. Jiang, L. Tong, Y. Xie, G. Zaharchuk, and M. Wintermark. 2019. Applications of deep learning to neuro-imaging techniques. *Frontiers in Neurology* 10:869.

8

Fostering the Data Management Environment

The cost-forecasting framework presented in this report directs the forecaster through a series of questions related to the cost drivers identified in Chapter 4 and summarized in a template in Appendix E. The framework, if used properly, could drive an analysis of the infrastructure and management activities needed at various points in the data life cycle, and the expertise that will need to be engaged. Understanding personnel needs is at least as important as understanding infrastructure costs because personnel costs associated with data and data platform management are likely to dominate total costs.

It is not part of common practice to think about data management budgets beyond the current funding period; however, creating a research and data environment that allows long-term, efficient, and cost-effective data discovery and data reuse requires long-term planning. That planning, in turn, requires that all involved in the scientific endeavor—researchers, research institutions, data curators and managers, data resource hosts, and funding institutions—embrace long-term planning approaches, regardless of the state of the data platform (i.e., primary research, active repository, or long-term preservation) being managed. This chapter presents strategies, actions, and advances that could be applied by members of the biomedical research community to create an environment conducive to long-term cost forecasts. The reader will need to determine how best to apply these based on his or her role in the scientific endeavor and on the data environment in which he or she works.

STRATEGIES

Efficient long-term data management is more likely if data resource managers, cost forecasters, and institutions that support them apply the strategies presented below (in italics).

- **Create data environments that foster discoverability and interpretability through long-term planning and investment throughout the data life cycle.** Data sharing is not equivalent to data reuse, and developing processes that allow efficient data preservation, archiving, and access to facilitate data reuse could benefit scientific discovery.

Advances in biomedical and information sciences result in larger and more complex data sets. The growing volumes of complex data exacerbate the challenges already faced by those who generate, use, or manage data. Members of the U.S. biomedical research community understand that scientific discovery is a key benefit of data preservation, aggregation, and access. That community increasingly advocates for the sharing of data to advance the scientific enterprise (e.g., NASEM, 2018). However, it is data reuse—not just data sharing—that is the objective. Making data discoverable and interpretable, and therefore reusable, requires forethought and sustained long-term investment.

- **Incorporate data management activities throughout the data life cycle to strengthen data curation and preservation.** Up-front costs may be increased, but data value may also increase, and the overall cost of research may be reduced.

There is a need for a cultural change within the biomedical research community related to data management. Curation activities are often left to the end of the funding period when few resources (or interest, time, or energy) remain. Instead, long-term data curation and data management needs ought to be considered throughout the course of research and the management of information resources. Data management and curation needs are vital in all data states, including during primary research.

- **Incorporate the expertise and resources needed to create and curate metadata throughout the data life cycle, and in the transition between data states into the cost forecast.** Data discoverability and reusability depend on adherence to community-accepted data and metadata standards.

The potential value of data will not be realized without strategic curatorial decisions by knowledgeable experts resulting in metadata that facilitate data discoverability and interpretability. That expertise needs to be anticipated and included in project budgets. Understanding the expertise and resources needed to create and curate metadata during the different data states and in the transition between states is vital and needs to be supported and encouraged by all within the biomedical research community, including institutions and funding agencies. At the individual-institution level, data-preservation efforts more likely will succeed if, for example, researchers are involved in decision-making and preservation efforts. Closer interactions between data librarians and researchers would result in a more efficient enterprise.

- **Weigh the benefits, risks (e.g., data loss), and costs (both up-front and anticipated) of data storage and computation options before selecting among options.** A service may look attractive from an immediate-financing perspective, but service-provider strategies require vetting and verification, including examination of exit or transition strategies and costs. Long-term costs need to be informed by a provider's risk-management strategies.

Substantial attention to confidentiality, ownership, and security; to standards, regulatory, and governance concerns; to access control; and to the various disruptors described in Chapter 7 will always be required, regardless of storage and computational options chosen. The risk-management strategy of service providers, and of any evolution that strategy undergoes with time, needs to be understood and addressed. The institution managing an information resource is not absolved from information technology responsibilities if commercial vendors are chosen to provide services.

ACTIONS

Individuals and select institutions within specific biomedical sectors may collaborate to increase the efficiency of data management efforts, but there is little guidance available from funding agencies and the institutions that support biomedical data resources on practices for long-term management and cost forecasting for the biomedical research community. The actions described below, especially if taken by funding agencies and institutions that support data resources, could expand the capacity of data producers and managers to make sound management decisions and cost forecasts.

- **Explicitly recognize the value of State 2 data resources (i.e., active repositories) to the enhanced curation, discoverability, and use of data.** This recognition is absent among the funding entities, researchers, and institutions supporting research, most of which apply the more traditional data management approach of transitioning data directly from the primary research environment (i.e., State 1) to long-term archiving (i.e., State 3).

Many recent advances in biomedical research have been possible because of new technologies that allow efficient aggregation, search, and compute of data in ways previously not possible. It is the State 2 environment in which analytic tools and data can be brought together to allow the sophisticated data manipulation necessary to produce those advances. However, creating State 2 platforms implies investments in ingesting and validating data, and the number of cost drivers affecting State 2 platforms is greater than for State 3 platforms. Developing and operating a sophisticated State 2 aggregating platform requires an organization, developers, user-interface designers, training and documentation, help desks, and community building. Current storage costs (even total storage costs) are only one—and, in many cases, probably not the dominant—factor in total system costs. Further, many researchers consider preparing data for public sharing (e.g., moving data from the State 1 primary research environment to a State 2 active repository and platform) to be burdensome and a task that provides little personal benefit. The biomedical research community needs to recognize that the long-term benefits of properly supporting State 2 data resources outweigh the costs and short-term burdens of establishing the resources and preparing data for them.

- **Structure cost forecasts for State 2 resources around communities and research programs rather than individual research efforts.** Because State 2 resources serve communities of researchers, it may not be appropriate to allocate the costs of managing data in a State 2 resource back to the individual data contributor.

State 2 platform costs generally do not track data from an individual researcher or research project, and the present study committee is not able to identify a good analysis of fixed versus incremental costs associated with individual streams of data contributed to active repositories. The committee was unable to find good examples of how State 2 data management costs—such as those incurred to bring data into compliance with community-developed standards—might be allocated back to the individual researchers who contributed the data. Because communities of researchers are involved, cost forecasting in this setting is better structured around communities and research programs rather than individual research efforts.

- **Support standardization efforts, including developing tools and methodologies to estimate the cost of standards development, encouraging the use of those tools and standards as part of the funding programs where appropriate, and explicitly supporting metadata preparation.** Support could take the form of funding and the provision of tools. Issuing clarifying language about the use of federal funds for data preservation beyond the performance period of the project that collected them would also help assist in the development and promotion of the use of community standards and metadata preparation.

As has been stated throughout this report, data that do not comply with standards or that have not been documented with appropriate metadata are of lesser value because they cannot be easily aggregated with other data or, more simply, may not be able to be found or understood. Existing incentives for researchers to deposit data in useful formats when standards exist are weak, and requirements to do so lack enforcement. Where no standards exist, data may be collected but then must be retrofit to comply with standards that are established. This process may occur years after the data were collected and possibly long after the supporting research grant has run out, and the last expert has moved on to other efforts. Few mechanisms exist to pay for retrofitting data, and perhaps little interest or incentive to do so exists on the part of the researcher, as she may have moved on to other projects or have little training in data management. Even if the researcher could anticipate and accurately forecast the cost of compliance, grants are not structured to allow money to be "held aside" until standards are established. Funding agencies can assist by contributing to tools for estimating the cost of standards development and metadata preparation, by explicitly funding metadata preparation, and by issuing clarifying language about the use of federal funds to preserve data beyond the end of the grant.

- **Identify incentives, tools, and training for adopting good data management practices, including cost-forecasting practices, which facilitate sustainable long-term data preservation, curation, and access.** Such activities would benefit the entire biomedical research community, including the institutions and funding entities that support research. To support these endeavors, funding entities need to better understand research-community needs, help the community to define desired outcomes, support training, develop realistic and actionable metrics for success, and provide near-term incentives for success.

The biomedical research community, including the institutions and those that fund research, needs to provide incentives for adopting good data management practices, including good cost-forecasting practices, that facilitate sustainable long-term data preservation, curation, and access. Researchers often lack the skills needed for efficient and effective data management, which translates to a lack of meaningful management and good data stewardship, and little understanding of the real costs of effective management or of how to forecast them. Based on interactions with various stakeholders during the conduct of this study, data management training for researchers is needed and desired. Training could help change the biomedical research culture so that good data management and cost forecasting become the norm in responsible research.

An incentive for researchers to more accurately account for the uncertainties associated with sharing data and future reuse might be for funders to place greater emphasis on such accounting in data management plans (DMPs) in grant proposals (discussed later in this chapter). Researchers would see an immediate benefit (i.e., research funding is contingent on action), and the prompt to take action is coming when the researchers are establishing their processes for research conduct (thus providing a timely prompt). But clear guidance for the researcher is also necessary for DMPs to be meaningful. For example, requests for application for funding sometimes seem to require new data resources to be all things for all stakeholders and even include potentially contradictory requirements. Incorporating better-directed guidance and training of individuals in data management would increase the likelihood of the desired outcomes.

Publishers and journals could also provide incentives, for example, by requiring data citations. Efforts to implement formal data citation across publishers (Cousijn et al., 2018; Fenner et al., 2019) are gaining traction, and most publishers at least informally accept data citations, although fully machine-readable data citations are still rare. Fully actionable data citations, however, require the infrastructure of a State 2 active repository or State 3 long-term preservation archive to ensure compliance with “findable, accessible, interoperable, and reusable” data principles (Wilkinson et al., 2016). Thus, by requiring data citations, publishers and journals can motivate researchers to use such infrastructure more consistently and possibly earlier.

Data management capacity might be increased by incorporating greater detail in, for example, training offered through the Collaborative Institutional Training Initiative (CITI) Program’s Responsible Conduct for Research modules.[1] Requiring independent proof of training as a requirement of receiving awards might improve capacity, as might encouraging multidisciplinary training much like that offered through the Integrative Graduate Education and Research Traineeship (IGERT)[2] program at scale, perhaps through multiagency support. Research on the normative outcomes of any increase in benefits resulting from improved data management skills could inform future training efforts.

Another incentive for researchers to participate constructively in data management, and especially State 2 resource planning, is providing them the opportunity to influence a superior computational environment. A data science platform could support complex research environments that free the researcher to focus on the science rather than on data collection and management. This capability could effectively reduce a State 1 environment to data

[1] The website for the CITI Program’s Responsible Conduct for Research modules is https://about.citiprogram.org/en/series/responsible-conduct-of-research-rcr/, accessed December 19, 2019.

[2] The website for the IGERT award is https://nsf.gov/awardsearch/showAward?AWD_ID=0903629&HistoricalAwards=false, accessed December 19, 2019.

(i.e., signal) capture. This idea underscores the benefit of a closer interaction between the data curators of State 2 platforms and individual researchers, recognizing that there are a variety of approaches to building and managing archives. This approach would co-locate the costs of supporting computing and analytics with an active repository.

- **Understand the charges associated with storage and computation in a data resource, regardless of who "pays the bill," when making decisions about data and workflows.** Institutions supporting research might develop mechanisms to inform researchers of the actual costs paid for the services rendered to them and encourage them to limit those costs.

Regardless of who provides the resources, there is a lack of visibility regarding storage costs in individual laboratories, institutions, and community resources. Understanding the charges associated with storage and computation in a data resource is vital for researchers making decisions about their own data and workflows. Researchers are often unaware of costs associated with data management in part because they typically are not responsible directly for those costs. Costs may be invisible to them if borne by their institutions or by a data-resource-platform manager (see Box 3.3). Purchased services (e.g., storage and computing) may be important, although the ability of individual researchers working in a primary research environment to forecast and manage those costs depends on the transparency of the information-technology environment. Mechanisms are needed to inform researchers of the actual costs paid for the services rendered to them, even if they are not directly charged.

ADVANCES FOR PRACTICE

Successful cost forecasting and sustainable management depend largely on an environment that supports decision makers, whether they are researchers, data scientists, data resource managers, or funding agencies. Methodologies for forecasting the life-cycle costs for preserving, archiving, and accessing biomedical data are immature and few tools and resources are available for those to quantify long-term costs with confidence and aid better understanding of uncertainties that can be tracked. Addressing the necessary advances identified below could facilitate the change in culture needed among decision makers to create such an environment. The following activities, likely to be enabled at an agency or research-institution level, could advance practices and drive future improvements in the ability to forecast costs.

- **Recognize explicitly that scientific data constitute an asset and that data stewardship requires support.** Biomedical research data and data resources are vital to the delivery of good science, and, ultimately, to the public good. The universities and institutions that support or enable research and host data resources, in turn, benefit from the recognition of that support.

Measuring data value in monetary terms is difficult, and yet it is the potential value of data that warrants the financial investments associated with their preservation. Unlike physical infrastructure, biomedical research data and the resources that house them are assets that contribute to the delivery of good science and, ultimately, the public good. The institutions that host or enable that public good will likely benefit from the recognition received for supporting such assets. Even so, there is only so much that can be done on the project or platform level. Currently, it is impossible to look across all data in the distributed biomedicine data enterprise to learn what data sets exist. Persistent metadata repositories are needed that include data set and research object identifiers.

- **Systematically collect data on costs associated with the biomedical research data enterprise to allow the translation of the framework outlined in this report into resources and methodologies that would benefit individual researchers and repository institutions.** A clear locus of responsibility for compiling this information systematically is necessary.

The true costs of preserving, archiving, and accessing biomedical research data need to be investigated in a systematic way at the funding-program-manager level rather than at the individual researcher or project level. Cost information at the researcher level could be collected at the outset of many projects when funds are requested, through DMPs, and tracked when progress is evaluated. Information thus collected so far has neither been uniformly integrated into award decisions nor transmitted to other parties involved. Researchers working in a State 1 primary research environment are often required to keep data for a prescribed period of time but are typically not responsible for costs or management beyond this.

Costs in the longer run (e.g., States 2 and 3) generally become an institutional responsibility. But institutional-level planning horizons are often only 1 or 2 years ahead rather than the many years required to realize the promise of current and future repositories. Some federal agencies (e.g., Department of Defense and Department of Energy) sustain a cadre of cost analysts and consider gathering the data needed for estimating costs as an important agency responsibility. Those agencies treat cost estimation as a profession and invest in training, recognizing success, critiquing failures, and encouraging assembly of cost-related data. The biomedical research data-preservation enterprise has become an undertaking that warrants a similar cadre to augment domain expertise and expertise in data science.

- **Develop easier mechanisms for creating and maintaining DMPs, automatically incorporating data and metadata into resources, and improving citations for data to work together with other research products.** By providing these mechanisms, funders and research institutions could help improve efficiency, return value for stakeholders, and increase the likelihood that stakeholders will make sound data-related decisions.

DMPs (see Appendix B) are typically static documents prepared as a mandatory—but not necessarily influential—part of the funding process. Placing more emphasis on quantified cost forecasts during the development of the DMP and the award process may be one way to incentivize early planning and communication, even if cost forecasts are uncertain. However, placing greater emphasis on cost forecasting at that time does not mean that the forecasts will become more precise estimates, rather they could be considered accurate reflections of uncertainties. Cost forecasts and DMPs need to evolve and be updated as research progresses and as associated data and the resources and technologies available to manage those data evolve. Monitored evolution of a DMP (e.g., at mid-term evaluations or at the end of the award period when making payment on awards) might inform eligibility for future funding. Machine-actionable DMP (e.g., Simms et al., 2017; see Appendix B) technologies may address issues related to realistic and evolving data management.

FACTORS FOR SUCCESSFUL ADOPTION OF DATA-FORECASTING APPROACHES

The current system for funding research cannot accommodate data life-cycle cost forecasting. For instance, the quantity, quality, and format of data collected might be uncertain when a proposal is written. They may become increasingly less uncertain a year into the award and when a grant is only partially spent out. During an information-gathering workshop organized by the committee (NASEM, 2020; see Appendix A for agenda), participants described incentives to do cost forecasting, with much discussion of how both incentives ("carrots") and rules and enforcement ("sticks") were important. Participants described how developing those rules and educating the community about the value of implementing them was fundamental to cost forecasting becoming a part of the responsible conduct of research (rather than a bureaucratic chore).

The culture change for the biomedical research community described in this report needs to be driven by community engagement. The Behavioral Insights Team[3] in the United Kingdom developed principles for encouraging desired outcomes, which might be applicable in the development and management of community repositories (Service et al., 2010). They recommend making processes easy, attractive, social, and timely (the "EAST" principles). People are more likely to engage in desired behavior if doing so is easy. To the greatest extent possible, it should be made easy for researchers and other stakeholders to make good data-related decisions from the onset. Research funders, research institutions, and journals are in positions to offer incentives, but processes need to be

[3] The website for the Behavioral Insights Team is https://www.bi.team/, accessed December 19, 2019.

driven by researchers so as to meet their needs and so they fully understand and agree to the value returned to them for their efforts. The ultimate beneficiaries of such efforts, of course, are the scientific enterprise and our nation's citizens, whose well-being biomedical science seeks to advance.

REFERENCES

Cousijn, H., A. Kenall, E. Ganley, M. Harrison, D. Kernohan, T. Lemberger, F. Murphy, et al. 2018. A data citation roadmap for scientific publishers. *Scientific Data* 5:180259. https://doi.org/10.1038/sdata.2018.259.

Fenner, M., M. Crosas, J.S. Grethe, D. Kennedy, H. Hermjakob, P. Rocca-Serra, G. Durand, et al. 2019. A data citation roadmap for scholarly data repositories. *Scientific Data* 6(1):28. https://doi.org/10.1038/s41597-019-0031-8.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, D.C.: The National Academies Press.

NASEM. 2020. *Planning for Long-Term Use of Biomedical Data: Proceedings of a Workshop*. Washington, D.C.: The National Academies Press.

Service, O., M. Hallsworth, D. Halpern, F. Algate, R. Gallagher, S. Nguyen, S. Ruda, et al. 2010. EAST: Four simple ways to apply behavioural insights. https://www.bi.team/wp-content/uploads/2015/07/BIT-Publication-EAST_FA_WEB.pdf.

Simms, S., S. Jones, D. Mietchen, and T. Miksa. 2017. Machine-actionable data management plans (maDMPs). *Research Ideas and Outcomes* 3:e13086.

Wilkinson, M.D., M. Dumontier, I. Jan Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018.

Appendixes

A

Meetings and Presentations

FIRST COMMITTEE MEETING
Washington, D.C.
February 27-28, 2019

National Library of Medicine (NLM) Interpretation of Study Statement of Task
Patricia Brennan, Director, NLM

NLM Program Organization, Services, Products, Resources, and Current NLM Decision Processes for Preserving/Archiving/Accessing/Deaccessioning Data
Patricia Brennan, Director, NLM
Jim Ostell, Director, National Center for Biotechnology Information (NCBI) at NLM

National Institutes of Health (NIH) Strategic Plan for Data Science
Susan Gregurick, Director, Division of Biomedical Technology, Bioinformatics, and Computational Biology, National Institute of General Medical Sciences and Senior Advisor for the NIH Office of Data Science Strategy

Panel Discussion with Representatives of NLM Leadership
Ivor D'Souza, NLM Chief Information Officer
Kim Pruitt, Acting Chief, Information Engineering Branch at NCBI at NLM
Dina Paltoo, NLM Assistant Director for Policy

SECOND COMMITTEE MEETING
Washington, D.C.
March 12-13, 2019

Cost Management of Big Data: Perspectives from Outside NIH
Jeffrey Spies, Founder, 221B LLC
Anita de Waard, Vice President of Research and Collaborations, Elsevier

THIRD COMMITTEE MEETING
Washington, D.C.
May 6-7, 2019

Disruptors in Digital Archiving: Presentation from the U.S. National Archives
Leslie Johnston, Director of Digital Preservation, U.S. National Archives

Disruptors in the Cloud
Vamshidhar Kommineni, Principal Project Manager, Azure Blob Storage, Microsoft

Indicators of Data Management Costs at the European Organization for Nuclear Research
Simone Campana, Deputy Project Leader of the Worldwide Computing Grid

WORKSHOP
Washington, D.C.
July 11-12, 2019

Welcome and Introductory Remarks
David Chu, Institute for Defense Analyses
Tyler Kloefkorn, National Academies
Sammantha Magsino, National Academies

Sponsor Expectations
Patricia Flatley Brennan, NLM

The Burdens and Benefits of "Long Tail" Data Sharing
Adam Ferguson, University of California, San Francisco

Panel Discussion: Researchers' Perspectives—Managing Risks and Forecasting Costs for Long-Term Data Preservation
Moderator: Margaret Levenstein, University of Michigan
Nuno Bandeira, University of California, San Diego
Jessie Tenenbaum, Duke University and the North Carolina Department of Health and Human Services
Georgia (Gina) Tourassi, Oak Ridge National Laboratory
Robert Williams, University of Tennessee Health Science Center

Panel Discussion: Addressing Data Risks and Their Costs
Moderator: Michelle Meyer, Geisinger
Amy O'Hara, Georgetown University
Brad Malin, Vanderbilt University Medical Center
Trevor Owens, U.S. Library of Congress

Breakout Sessions—Tools and Practices That NLM Could Use to Help Researchers and Funders Better Integrate Risk Management Practices and Considerations into Data Preservation, Archiving, and Accessing Decisions

Data—What's It Going to Cost, and What's in It for Me?
Phil Bourne, University of Virginia

Precisely Practicing Medicine from 700 Trillion Points of Data
Atul Butte, University of California, San Francisco

Open Discussion—Reflections, Plans for Day 2, Coordination with Study
Alexa McCray, Harvard Medical School

Panel Discussion: Incentives, Mechanisms, and Practices for Improved Awareness of Cost Consequences in Data Decisions
Moderator: Lars Vilhuber, Cornell University
John Chodacki, University of California Curation Center, California Digital Library
Melissa Cragin, San Diego Supercomputer Center
Wendy Nilsen, National Science Foundation
Lucy Ofiesh, Center for Open Science

Breakout Sessions—Methods to Encourage NIH-Funded Researchers to Consider, Update, and Track Lifetime Data Costs

Panel Discussion: Researchers' Perspectives—Reflections and Next Steps
Moderator: Margaret Levenstein, University of Michigan
Nuno Bandeira, University of California, San Diego
Jessie Tenenbaum, Duke University and the North Carolina Department of Health and Human Services
Georgia (Gina) Tourassi, Oak Ridge National Laboratory
Robert Williams, University of Tennessee Health Science Center

Closing Remarks—Themes and Opportunities
Maryann Martone, University of California, San Diego

FOURTH COMMITTEE MEETING
Washington, D.C.
September 17-18, 2019

No open session presentations were held during this meeting.

FIFTH COMMITTEE MEETING
Washington, D.C.
October 29-30, 2019

No open session presentations were held during this meeting.

SITE VISITS

National Center for Microscopy and Imaging Research
La Jolla, California
September 11, 2019

National Center for Microscopy and Imaging Research

Mark Ellisman, Director
Steven Peltier, Deputy Director
Willy Wong, Software Technical Lead
Matthew Madany, Researcher
Sean Penticoff, Information Technology Manager
Matthias Haberl, Postdoctoral Fellow

University of California, San Diego/San Diego Supercomputer Center Advanced Cyber-Infrastructure Development Lab
La Jolla, California
September 12, 2019

UCSD/SDSC Advanced Cyber-Infrastructure Development Lab

Jim Short, San Diego Supercomputer Center
Mike Norman, San Diego Supercomputer Center
Sandeep Chandar, San Diego Supercomputer Center
Brian Balderston, San Diego Supercomputer Center
Christine Kirkpatrick, San Diego Supercomputer Center
David Minor, University of California, San Diego Library
Sibyl Schaefer, University of California, San Diego Library
Jeffrey Burke, San Diego Supercomputer Center and Retired Senior Vice President at Seagate
Scott Kahn, Chief Information Officer at LunaDNA, previously Chief Information Officer at Illumina

National Institutes of Health
Bethesda, Maryland
September 18, 2019

National Institute of Mental Health

Greg Farber, Director of the Office of Technology Development and Coordination

National Institute of Nursing Research

Jessica Gill, Lasker Clinical Research Scholar

National Institute of Allergy and Infectious Diseases

John McGowan, Deputy Director for Science Management
Jill Harper, Director for the Office of Biodefense Research and Surety

National Human Genome Research Institute

Eric Green, Director
Valentina Di Francesco, Lead Program Director, Computational Genomics and Data Science, Division of Genome Sciences

Carolyn M. Hutter, Director, Division of Genome Sciences
Ajay Pillai, Program Director, Molecular Libraries Program, Division of Genome Sciences
Shurjo K. Sen, Program Director, Division of Genome Sciences
Ken Wiley, Program Director, Division of Genomic Medicine

Dana-Farber Cancer Institute
Boston, Massachusetts
September 25, 2019

Participants

Eliezer Van Allen
Moritz Kircher
Robert Gray
Laura MacConaill
Clifford Meyer
Daphne Haas-Kogan
Hugo Aerts
Bruce Johnson

Harvard Medical School
Cambridge, Massachusetts
September 25, 2019

Participants

Hosts: *Mercè Crosas*, *David Golan*, *Caroline Shamu*

Faculty

Brent Coull
Chris Harvey
Jason Key
Steve McCarroll
Sean Megason
Jeremy Muhlich
Peter Park
Jon Seidman
Piotr Sliz
Artem Sokolov
Peter Sorger
Yaoyu Wang
Ista Zahn

Libraries

Steve Abrams
Ceilyn Boyd
Julie Goldman
Emily Gustainis
Meghan Kerr
Amber LaFountain
Scott Lapinski

Elaine Martin
Stuart Snydman
Amy Van Epps
Suzanne Wones

Research Computing and Information Technology
Paul Edmons
Mason Miranda
Deb Scott
Paul Williams

The Broad Institute of the Massachusetts Institute of Technology and Harvard
Cambridge, Massachusetts
September 26, 2019

Introduction to the Data Science Platform (DSP)
Anthony Philippakis

Patient Facing Platforms
Andrew Zimmer, Jen Lapan

Data Engineering
Kathleen Tibbetts, *Kristian Cibulskis*

Overview of Terra
Clare Bernard, *Kristian Cibulskis*

Meeting with Daniel MacArthur

Overview of the Genome Analysis Toolkit
Eric Banks

ML4CVD: Machine Learning for Cardiovascular Disease
Puneet Batra

Security at Broad
David Bernick

Tour of DSP and Agile Overview
Diolinda Vaz

Amazon Web Services
Seattle, Washington
October 23, 2019

Participants
Marcy Collinson
Aaron Friedman

Elliot Menschik
Ann Merrihew
Sanjay Padhi
Kam Syed

Institute for Systems Biology
Seattle, Washington
October 24, 2019

Participants

Jim Heath
Nathan Price
Sui Huang
Jennifer Hadlock
Andrew Magis
John Earls
Christian Diener

Allen Institute
Seattle, Washington
October 25, 2019

Brain Science

Michael Hawrylycz
Lydia Ng
Shoaib Mufti
Christof Koch
Carol Thompson
Tyler Mollenkopf
Rob Young
John Phillips

Cell Science

Basu Chaudhuri

Immunology

Paul Meijer

Fred Hutchinson Cancer Research Center
Seattle, Washington
October 25, 2019

Participants

Elizabeth Boyd
David Browdy
Marior Dorer
Rachel Galbraith

Jennifer Griffith
Brenda Kostelecky
Elizabeth Masnari
Anders McConachie
Dirk Petersen
Niki Robinson
Bonnie Schae
Matthew Trunnell

B

Active Data Management Plans as a Planning Tool

A data management plan (DMP) is a formal document that outlines data types and formats, dissemination and sharing plans, roles and responsibilities, and preservation plans for data generated by a project. Various federal research funders, including the National Science Foundation (NSF) and various institutes of the National Institutes of Health (NIH) require grantees to propose a DMP, but specific requirements are not uniform among the agencies. Quality and utility of DMPs are an issue, with the perception that DMPs are an "annoying administrative exercise" (Simms et al., 2017). The typical DMP is a text document written as a verbose narrative.

Various guidance documents and templates for DMP production evolved (e.g., the National Network of Libraries of Medicine [NNLM][1]), and a first generation of tools to facilitate the production of DMPs was created, and are widely available. Some examples of tools include DMPTool (California Digital Libraries),[2] the DMPOnline (Digital Curation Centre—UK),[3] and the Interdisciplinary Earth Data Alliance (IEDA) DMP Tool.[4] Generally, those tools guided the creation of Microsoft Word or PDF documents and incorporated templates for use with various funders. More recently, a second generation of those tools was created, with the goal of making DMPs "machine actionable" (i.e., written so that computer programs can parse the content cleanly and take action based on the information in such machine-actionable DMPs [maDMPs]). Newer DMP tools incorporate richer information, such as a list of acceptable repositories, and better guidance, with richer prefillable information.

While the creation of a DMP guides the researcher in formulating a data management process, the role of the DMP in the evaluation of grant proposals and in post-award evaluations is less clear. Most grant agencies require a DMP but do not explicitly score DMPs or integrate them formally into scoring a proposal under consideration for funding. For instance, NIH scoring guidelines[5] make no reference to DMPs, although data-sharing plans are a required element of proposals (in fact, the word "data" does not appear in the scoring guide). The NSF-wide

[1] The NNLM website providing a collection of data management guides is https://nnlm.gov/data/data-management-plan, accessed January 14, 2020.

[2] The website for DMPTool is https://dmptool.org/, accessed January 14, 2020.

[3] The website for DMPOnline is https://dmponline.dcc.ac.uk/, accessed January 14, 2020.

[4] The website for the IEDA DMP is https://www.iedadata.org/dmp/, accessed January 14, 2020.

[5] The website describing the NIH scoring guidance is https://grants.nih.gov/grants/policy/review/rev_prep/scoring.htm, accessed January 14, 2020.

Proposal and Award Policies and Procedures Guide[6] notes a requirement for (two-page) DMPs but only as a supplementary material, leaving the evaluation thereof in the realm of individual divisions and program managers. The IEDA, acting as a repository for observational geoscience data, provides researchers with the ability to generate a Data Compliance Report[7] based on their NSF award number. The resulting output can be used "to demonstrate that your data are registered with IEDA systems and you are compliant with NSF Data Policies," but it is not clear how much weight that carries in post-award evaluation or in subsequent proposals. The NSF Directorate for Biological Sciences notes a requirement for inclusion in annual and final reports, as well as part of "Results of prior NSF Support" in subsequent proposals, but, again, it is not clear what weight these are given during the evaluation process.[8] The Canadian "Tri-Agency Statement of Principles on Digital Data Management"[9] highlights researcher and institution responsibilities only with respect to the development of and compliance with DMPs, and the Canadian Institutes of Health Research's Peer Review Manual[10] does not include DMPs in its scoring criteria.

Halbert (2013) and Keralis and colleagues (2013) identify lack of consistency across funding agencies as a barrier for a consistent response by researchers and data librarians to data management challenges (see also Williams et al., 2017). The development of the second generation of DMP tools may be seen as a direct result of these findings and the attempt to construct a "meta-DMP" that provides consistent guidance regardless of the underlying agency reporting requirement.

maDMPs, also referred to as dynamic DMPs (Simms and Jones, 2017; Simms et al., 2017) or data management records (Morgan and Janke, 2017), may be useful for forecasting of costs of data preservation. They are specifically proposed as a more formal (machine-readable) document, allowing for data exchange across various entities, in particular across the entire data life cycle. Integration with funders, as well as institutional and community capacity planning, are specifically identified (Simms et al., 2017). maDMPs are evolving, and a standard has not yet emerged, although several use cases and implementations (Morgan and Janke, 2017) exist. At the time of writing, working groups at the Research Data Alliance[11] and FORCE11[12] (Chodacki et al., 2016) are working on use cases from a variety of disciplines and coordinating on standards. In particular, maDMPs target metadata such as quantity and type of data, regardless of storage location, allowing for an assessment of time-varying cost of storing such data. They strongly encourage use of persistant identifiers for people, institutions, and assets, so that maDMPs are globally intelligible (Bakos et al., 2018).

REFERENCES

Bakos, A., T. Miksa, and A. Rauber. 2018. Research data preservation using process engines and machine-actionable data management plans. *Digital Libraries for Open Knowledge*. https://doi.org/10.1007/978-3-030-00066-0_6.

Chodacki, J., M. Crosas, M. Martone, and S.-A. Sansone. 2016. FAIR DMP. *FORCE11*, May 3. https://www.force11.org/group/fairdmp.

Halbert, M. 2013. The problematic future of research data management: Challenges, opportunities and emerging patterns identified by the DataRes Project. *International Journal of Digital Curation*. 8(2):111-122. https://doi.org/10.2218/ijdc.v8i2.276.

Keralis, S.D.C., S. Stark, M. Halbert, and W.E. Moen. 2013. Research data management in policy and practice: The DataRes Project. In *Research Data Management: Principles, Practices, and Prospects,* 16-38. Council on Library and Information Resources. https://libres.uncg.edu/ir/uncg/f/M_Halbert_Research_2013.pdf.

Morgan, H., and A. Janke. 2017. DMRs, making DMPs relevant again. May 22. http://andscentral.blogspot.com/2017/05/dmrs-making-dmps-relevant-again.html.

[6] The NSF Proposal and Award Policies and Procedures Guide can be found at https://www.nsf.gov/publications/pub_summ.jsp?ods_key=pappg&WT.z_pims_id=0, accessed January 14, 2020.

[7] The IEDA data compliance reporting tool can be found at http://app.iedadata.org/dcr/report.php, accessed January 14, 2020.

[8] The website for the NSF Directorate for Biological Sciences updated information about DMP is https://www.nsf.gov/bio/pubs/BIODMP061511.pdf, accessed January 14, 2020.

[9] The Government of Canada website with the 2015 Tri-Agency Statement of Principles on Digital Data Management is http://science.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html, accessed January 14, 2020.

[10] The Canadian Institutes for Health Research Peer Review Manual can be found at https://cihr-irsc.gc.ca/e/49564.html, accessed January 14, 2020.

[11] The website for the Data Research Alliance is https://www.rd-alliance.org/, accessed January 14, 2020.

[12] The website for FORCE11 is https://www.force11.org/, accessed January 14, 2020.

Simms, S., and S. Jones. 2017. Next-generation data management plans: Global, machine-actionable, FAIR. *International Journal of Digital Curation* 12(1):36. https://doi.org/10.2218/ijdc.v12i1.513.

Simms, S., S. Jones, D. Mietchen, and T. Miksa. 2017. Machine-actionable data management plans (maDMPs). *Research Ideas and Outcomes* 3:e13086.

Williams, M., J. Bagwell, and M.N. Zozus. 2017. Data management plans: The missing perspective. *Journal of Biomedical Informatics* 71(July):130-142.

C

Identifying Salary Ranges for Jobs Relevant to the Data Life Cycle

The main text identifies certain job descriptions with associated salary ranges, from L (low) to VH (very high). This appendix identifies possible job titles and associated salary ranges observed in workplace and occupational surveys conducted by the Bureau of Labor Statistics (BLS, 2019a).

DATA USED

Occupational Employment Statistics

Collection methods, estimation methodology, and coverage are described in BLS (2019b). The committee downloaded the data from https://www.bls.gov/oes/special.requests/oesm18nat.zip on October 30, 2019. From the downloaded data, national_M2018_dl.xlsx was used.

Occupational Information Network

The Occupational Information Network (O*NET) database is a U.S. Department of Labor–sponsored database developed by the National Center for O*Net Development.[1] The database provides standardized descriptions of hundreds of occupations within the U.S. economy. The database comprises worker attributes and job characteristics. Information is collected using a two-stage design in which the following occurs:

- A statistically random sample of businesses expected to employ workers in the targeted occupations is identified.
- A random sample of workers in those occupations within those businesses is selected. Data are collected by surveying job incumbents using a randomly assigned standardized questionnaire on occupation characteristics, out of three questionnaires. Additional questions cover tasks and demographic information.
- Abilities and skills information is developed by occupational analysts using the updated information from incumbent workers (National Center for O*NET Development, 2019a).

[1] See https://www.onetonline.org/, accessed August 12, 2020.

A data dictionary (National Center for O*NET Development, 2019b) provides additional information.

Version 23_2 of the data (National Center for O*NET Development, 2019c)[2] was used for committee determination of salaries. Both Occupation Data.xlsx and Alternate Titles.xlsx were used.

METHODS

Mapping Job Titles to Standard Occupational Classifications

O*NET is structured around Standard Occupational Classification (SOC; BLS, 2019c). The committee's main text has a normative list of job descriptions based on data management practiced at university libraries. These may not match reported standard occupation titles. The O*NET data provide a long but not exhaustive list of alternate mentions of job titles for specific occupations (Alternate Titles.xlsx). Using both the standard occupation title as well as the alternative mention, the normative job title is matched via probabilistic matching, using the Jaro-Winkler distance (Winkler, 1990) as implemented in the R package fuzzyjoin (Robinson, 2019). All reasonable matches ($d < 0.05$) were kept to obtain a list of similar occupations and their SOC codes.

Mapping SOC into Salary Ranges

Occupational Employment Statistics computes for each SOC code a salary range, comprising annual salary and hourly wages, and characterized by the 25th and 75th percentile, as well as the median. The annual salary distributions were attached to each of the identified occupations (Table C.1), and then these statistics were collapsed to a triplet of information for each normative job description (Table C.2). To do so, the minimum of all observed 25th percentiles, the median of all observed medians, and the maximum of all observed 75th percentiles were chosen. No weights were applied. An alternative implementation might use the employment shares to create weighted statistics. Reliability statistics were not computed, as the resulting table is meant to be indicative, not precise.

RESULTS

Table C.1 lists the annual salaries, as of 2018, by job title (median, and the 25th and 75th percentile), for all occupations identified as having similar names as the normative description in Chapter 2. Blank salaries ("NA") indicate that no occupation code could be found on O*NET based on the normative description. Table C.2 lists the ranges, as defined above, for each of the normative descriptions (Chapter 2), based on the underlying occupations identified. Table C.3 lists the statistics associated with each of the salary categories, from low to very high. While the categories are defined based on the experience of members of the committee, ex ante, they match up well with observed median salaries in 2018.

FULL CODE AND DATA

The code and data underlying this appendix, including an exhaustive list of the committee's edits (inclusions and exclusions) to the list of occupations, are available at https://github.com/labordynamicsinstitute/job-description-and-wages.

TABLE C.1 Annual Salaries (2018) by Job Title for Occupations

Job Title	Title	SOC	Alternative Title	25th Percentile ($)	Median Income ($)	75th Percentile ($)
Researcher	Industrial Ecologists	19-2041	Researcher	53,580	71,130	94,590
Researcher	Anthropologists	19-3091	Researcher	48,020	62,410	80,230
Researcher	Historians	19-3093	Researcher	40,670	61,140	85,700
Researcher	Biofuels/Biodiesel Technology and Product Development Managers	11-9041	Scientist	112,400	140,760	173,180
Researcher	Mathematicians	15-2021	Scientist	73,490	101,900	126,070
Researcher	Chemical Engineers	17-2041	Scientist	81,900	104,910	133,320
Researcher	Nanosystems Engineers	17-2199	Scientist	69,890	96,980	126,200
Researcher	Manufacturing Engineering Technologists	17-3029	Scientist	47,500	63,200	80,670
Researcher	Biologists	19-1020	Scientist	56,730	77,550	103,540
Researcher	Biochemists and Biophysicists	19-1021	Scientist	64,230	93,280	129,950
Researcher	Bioinformatics Scientists	19-1029	Scientist	60,250	79,590	98,040
Researcher	Medical Scientists, Except Epidemiologists	19-1042	Scientist	59,580	84,810	118,040
Researcher	Chemists	19-2031	Scientist	56,290	76,890	103,820
Researcher	Hydrologists	19-2043	Scientist	61,280	79,370	100,090
Researcher	Remote Sensing Scientists and Technologists	19-2099	Scientist	75,830	107,230	136,930
Researcher	Geographers	19-3092	Scientist	63,270	80,300	96,980
Data Librarian	Librarians	25-4021	NA	46,130	59,050	74,740
Data Librarian	Library Science Teachers, Postsecondary	25-1082	Librarian	56,550	71,560	90,550
Data Librarian	Archivists	25-4011	Librarian	38,090	52,240	71,250
Metadata Librarian	Librarians	25-4021	NA	46,130	59,050	74,740
Metadata Librarian	Library Science Teachers, Postsecondary	25-1082	Librarian	56,550	71,560	90,550
Metadata Librarian	Archivists	25-4011	Librarian	38,090	52,240	71,250
Records Management Specialist	Librarians	25-4021	NA	46,130	59,050	74,740
Records Management Specialist	Library Science Teachers, Postsecondary	25-1082	Librarian	56,550	71,560	90,550
Records Management Specialist	Archivists	25-4011	Librarian	38,090	52,240	71,250
Curator	Curators	25-4012	NA	39,580	53,780	72,830
Curator	Archivists	25-4011	NA	38,090	52,240	71,250
Curator	Archeologists	19-3091	Curator	48,020	62,410	80,230

TABLE C.1 Continued

Job Title	Title	SOC	Alternative Title	25th Percentile ($)	Median Income ($)	75th Percentile ($)
Research Domain Curator	Biofuels/Biodiesel Technology and Product Development Managers	11-9041	Scientist	112,400	140,760	173,180
Research Domain Curator	Mathematicians	15-2021	Scientist	73,490	101,900	126,070
Research Domain Curator	Chemical Engineers	17-2041	Scientist	81,900	104,910	133,320
Research Domain Curator	Nanosystems Engineers	17-2199	Scientist	69,890	96,980	126,200
Research Domain Curator	Manufacturing Engineering Technologists	17-3029	Scientist	47,500	63,200	80,670
Research Domain Curator	Biologists	19-1020	Scientist	56,730	77,550	103,540
Research Domain Curator	Biochemists and Biophysicists	19-1021	Scientist	64,230	93,280	129,950
Research Domain Curator	Bioinformatics Scientists	19-1029	Scientist	60,250	79,590	98,040
Research Domain Curator	Medical Scientists, Except Epidemiologists	19-1042	Scientist	59,580	84,810	118,040
Research Domain Curator	Chemists	19-2031	Scientist	56,290	76,890	103,820
Research Domain Curator	Climate Change Analysts	19-2041	Scientist	53,580	71,130	94,590
Research Domain Curator	Hydrologists	19-2043	Scientist	61,280	79,370	100,090
Research Domain Curator	Remote Sensing Scientists and Technologists	19-2099	Scientist	75,830	107,230	136,930
Research Domain Curator	Anthropologists	19-3091	Scientist	48,020	62,410	80,230
Research Domain Curator	Geographers	19-3092	Scientist	63,270	80,300	96,980
Research Domain Project Manager	Biofuels/Biodiesel Technology and Product Development Managers	11-9041	Scientist	112,400	140,760	173,180
Research Domain Project Manager	Mathematicians	15-2021	Scientist	73,490	101,900	126,070
Research Domain Project Manager	Chemical Engineers	17-2041	Scientist	81,900	104,910	133,320
Research Domain Project Manager	Nanosystems Engineers	17-2199	Scientist	69,890	96,980	126,200
Research Domain Project Manager	Manufacturing Engineering Technologists	17-3029	Scientist	47,500	63,200	80,670
Research Domain Project Manager	Biologists	19-1020	Scientist	56,730	77,550	103,540
Research Domain Project Manager	Biochemists and Biophysicists	19-1021	Scientist	64,230	93,280	129,950

continued

TABLE C.1 Continued

Job Title	Title	SOC	Alternative Title	25th Percentile ($)	Median Income ($)	75th Percentile ($)
Research Domain Project Manager	Bioinformatics Scientists	19-1029	Scientist	60,250	79,590	98,040
Research Domain Project Manager	Medical Scientists, Except Epidemiologists	19-1042	Scientist	59,580	84,810	118,040
Research Domain Project Manager	Chemists	19-2031	Scientist	56,290	76,890	103,820
Research Domain Project Manager	Climate Change Analysts	19-2041	Scientist	53,580	71,130	94,590
Research Domain Project Manager	Hydrologists	19-2043	Scientist	61,280	79,370	100,090
Research Domain Project Manager	Remote Sensing Scientists and Technologists	19-2099	Scientist	75,830	107,230	136,930
Research Domain Project Manager	Anthropologists	19-3091	Scientist	48,020	62,410	80,230
Research Domain Project Manager	Geographers	19-3092	Scientist	63,270	80,300	96,980
Informatician	Computer Systems Analysts	15-1121	NA	68,730	887,40	113,460
Informatician	Information Technology Project Managers	15-1199	IT Specialist	66,410	90,270	117,070
Data Wrangler	Information Technology Project Managers	15-1199	IT Specialist	66,410	90,270	117,070
Education Specialist	Health Educators	21-1091	Education Specialist	39,800	54,220	74,660
Education Specialist	Special Education Teachers, Secondary School	25-2054	Education Specialist	48,630	60,600	77,820
Education Specialist	Instructional Coordinators	25-9031	Education Specialist	49,280	64,450	82,860
Communication Specialist	Public Relations Specialists	27-3031	Communication Specialist	44,490	60,000	81,550
Software Engineer	Computer and Information Research Scientists	15-1111	Software Engineer	91,650	118,370	149,470
Software Engineer	Software Developers, Applications	15-1132	Software Engineer	79,340	103,620	130,460
Software Engineer	Software Developers, Systems Software	15-1133	Software Engineer	85,610	110,000	139,550
IT Security Specialist	Security Management Specialists	13-1199	NA	52,200	70,530	94,890
IT Systems Engineer	Computer and Information Systems Managers	11-3021	NA	110,110	142,530	180,190
IT Systems Engineer	Information Technology Project Managers	15-1199	IT Specialist	66,410	90,270	117,070
IT Project Manager	Computer and Information Systems Managers	11-3021	NA	110,110	142,530	180,190
IT Project Manager	Information Technology Project Managers	15-1199	IS/IT Project Manager	66,410	90,270	117,070

TABLE C.1 Continued

Job Title	Title	SOC	Alternative Title	25th Percentile ($)	Median Income ($)	75th Percentile ($)
Project Manager	Construction Managers	11-9021	Project Manager	70,670	93,370	123,720
Project Manager	Architectural and Engineering Managers	11-9041	Project Manager	112,400	140,760	173,180
Project Manager	Managers, All Other	11-9199	Project Manager	75,460	107,480	143,230
Project Manager	Information Technology Project Managers	15-1199	Project Manager	66,410	90,270	117,070
Project Manager	Environmental Engineers	17-2081	Project Manager	66,590	87,620	112,230
Project Manager	Wind Energy Engineers	17-2199	Project Manager	69,890	96,980	126,200
Project Manager	Environmental Restoration Planners	19-2041	Project Manager	53,580	71,130	94,590
Project Manager	Social Science Research Assistants	19-4061	Project Manager	35,450	46,640	60,830
Project Manager	Remote Sensing Technicians	19-4099	Project Manager	37,940	49,670	63,340
Project Manager	Technical Directors/Managers	27-2012	Project Manager	48,520	71,680	110,350
Project Manager	Intelligence Analysts	33-3021	Project Manager	57,560	81,920	107,000
Senior Staff	NA	NA	NA	NA	NA	NA
Policy Specialist	NA	NA	NA	NA	NA	NA
Administrative Staff	First-Line Supervisors of Office and Administrative Support Workers	43-1011	NA	42,750	55,810	71,550
Administrative Staff	Executive Secretaries and Executive Administrative Assistants	43-6011	NA	46,530	59,340	74,460
Administrative Staff	Secretaries and Administrative Assistants, Except Legal, Medical, and Executive	43-6014	NA	28,930	36,630	46,230
Administrative Staff	Business Operations Specialists, All Other	13-1199	Administrative Assistant	52,200	70,530	94,890
Administrative Staff	Billing and Posting Clerks	43-3021	Administrative Assistant	31,870	37,800	46,350
Administrative Staff	New Accounts Clerks	43-4141	Administrative Assistant	30,300	35,800	42,050
Administrative Staff	Medical Secretaries	43-6013	Administrative Assistant	29,580	35,760	43,200
Facilities Manager	General and Operations Managers	11-1021	Facilities Manager	65,650	100,930	157,120
Facilities Manager	Administrative Services Managers	11-3011	Facilities Manager	71,850	96,180	127,100
Facilities Manager	Property, Real Estate, and Community Association Managers	11-9141	Facilities Manager	41,210	58,340	85,120
Facilities Manager	First-Line Supervisors of Housekeeping and Janitorial Workers	37-1011	Facilities Manager	31,020	39,940	52,280

continued

TABLE C.1 Continued

Job Title	Title	SOC	Alternative Title	25th Percentile ($)	Median Income ($)	75th Percentile ($)
Facilities Manager	First-Line Supervisors of Office and Administrative Support Workers	43-1011	Facilities Manager	42,750	55,810	71,550
Facilities Manager	First-Line Supervisors of Mechanics, Installers, and Repairers	49-1011	Facilities Manager	51,430	66,140	83,980
Facilities Manager	Maintenance and Repair Workers, General	49-9071	Facilities Manager	29,560	38,300	50,100
Data Scientist	Computer and Information Research Scientists	15-1111	Data Scientist	91,650	118,370	149,470

NOTE: IT, information technology; SOC, Standard Occupational Classification.

TABLE C.2 Salary Ranges for Job Classifications as Defined in Chapter 2

Job Title	25th Percentile ($)	Median Salary ($)	75th Percentile ($)
Administrative Staff	28,930	37,800	94,890
Communication Specialist	44,490	60,000	81,550
Curator	38,090	53,780	80,230
Data Librarian	38,090	59,050	90,550
Data Scientist	91,650	118,370	149,470
Data Wrangler	66,410	90,270	117,070
Education Specialist	39,800	60,600	82,860
Facilities Manager	29,560	58,340	157,120
Informatician	66,410	89,505	117,070
IT Project Manager	66,410	116,400	180,190
IT Security Specialist	52,200	70,530	94,890
IT Systems Engineer	66,410	116,400	180,190
Metadata Librarian	38,090	59,050	90,550
Policy Specialist	Inf	NA	-Inf
Project Manager	35,450	87,620	173,180
Records Management Specialist	38,090	59,050	90,550
Research Domain Curator	47,500	80,300	173,180
Research Domain Project Manager	47,500	80,300	173,180
Researcher	40,670	79,945	173,180
Senior Staff	Inf	NA	-Inf
Software Engineer	79,340	110,000	149,470

TABLE C.3 Statistics Across Each of the Salary Categories (used in Chapter 2 of this report)

Relative Salary	25th Percentile ($)	Median Salary ($)	75th Percentile ($)	N	Missing
Low	28,930	37,800	94,890	7	0
Medium	29,560	61,505	173,180	34	0
High	40,670	80,300	180,190	50	1
Very High	52,200	103,620	180,190	10	1

REFERENCES

BLS (Bureau of Labor Statistics). 2019a. *Occupational Employment Statistics*. Data set. Bureau of Labor Statistics, OES Program. https://www.bls.gov/oes/home.htm.

BLS. 2019b. *Survey Methods and Reliability Statement for the May 2018 Occupational Employment Statistics Survey*. Bureau of Labor Statistics, OES Program. https://www.bls.gov/oes/current/methods_statement.pdf.

BLS. 2019c. 2018 Standard Occupational Classification System. https://www.bls.gov/soc/2018/major_groups.htm.

National Center for O*NET Development. 2019a. O*NET Data Collection Overview. https://www.onetcenter.org/dataCollection.html.

National Center for O*NET Development. 2019b. *O*NET® 23.2 Database*. Data Dictionary. O*NET Resource Center. https://www.onetcenter.org/dl_files/database/db_23_2_dictionary.pdf.

National Center for O*NET Development. 2019c. *O*NET® Database Release 23.2*. Data set. O*NET Resource Center. https://www.onetcenter.org/db_releases.html.

Robinson, D. 2019. Fuzzyjoin: Join Tables Together on Inexact Matching. https://github.com/dgrtwo/fuzzyjoin.

Winkler, W.E. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359. https://files.eric.ed.gov/fulltext/ED325505.pdf.

D

Soft Costs for Digital Preservation

Not all costs for creating and sustaining a biomedical information resource show up in the budget of the organizational units creating the data or operating the resource. Some "soft costs" might show in other parts of an organization as effort expended by users of the resource or as lost or delayed opportunities. Merrill (2017) identified seven categories of soft costs for long-term digital preservation in a corporate setting (see Box D.1).

Merrill's enumeration of soft costs is not necessarily exhaustive for the biomedical-information domain. There may be other soft costs to consider. For example, there is investigator burden: the effort required by researchers to submit their data to a repository. Another soft cost is loss of confidence: data in a resource lose value when users do not use that resource owing to a lack of trust in it or a concern about the results obtained from it. Such loss might arise because there is not enough information to replicate the process that generated the data or to audit their handling once received by a resource. Loss could also arise from lack of curation of the resource or uncertain interpretation of the data owing to no or little metadata.

DIFFICULTY IN QUANTIFYING SOFT COSTS

Merrill (2017) notes that "soft costs are important, but may be hard to isolate, define, or measure. Many soft costs are qualitative in nature. At times they can become hard costs when unusual events happen (like declaring a disaster, or in a pre-trial rush to access data). Project related benefits (staff efficiency, risk) are usually characterized by soft costs, since the IT department does not have the burden to measure or reduce these costs." Putting a dollar amount to soft costs so that they can be compared directly to hard costs does not seem feasible in most cases, but it is often possible to compare the relative soft costs of alternative approaches. For example, considering Merrill's soft cost of discovery time, there might be two approaches to supporting a repository of genetic sequences. In the first approach, the sequences can be retrieved only by accession number and organism name. In the second approach, there is an additional index that allows searching by sequence similarity. Discovery time is expected to be lower in the second approach for a task such as determining if a new sequence duplicates an existing sequence. As another example, consider Merrill's cost of performance. One option for the sequence repository would be to internally support a service for alignment of a deposited sequence to an appropriate reference sequence. A second option is not to support such a service. In that case, an investigator needing a reference alignment would need to download the sequence in question, find the appropriate reference sequence somewhere, and locate and apply an appropriate tool to perform the alignment. Clearly, the second option has a higher cost of performance.

BOX D.1
Soft Costs as Defined by Merrill (2017)

- *Provisioning time* "is the time and local effort required to acquire and present capacity for the retention period. Internal processes, procurement, provisioning steps, and delivery lead-time all contribute to this cost" (Merrill, 2017). This category does not include the direct costs for storage capacity itself, nor the ongoing costs of operating that capacity. An attraction of cloud storage is that provisioning time can be reduced from months to minutes. Long provisioning times can be detrimental to research if they delay results or access to those results by others.
- *Discovery time* "is the time it takes a person or an application to find digital content. The format may have some impact on this time, as well the robustness of the meta-data management system. Risks can arise if the time required to discover (and restore) are unsatisfactory" (Merrill, 2017). A resource might get little or no use if locating information within it is cumbersome, say because of limited search capability.
- *Time to restore* is "how long a person (or an application) has to wait for the data to be restored (to the last bit) after the request is made" (Merrill, 2017). This soft cost would pertain to cases in which data in a repository have to be moved from offline to online storage before they are accessed.
- *Cost of expansion* "of the repository must be planned since the ever-increasing growth in stored information will require future increased capacity" (Merrill, 2017). Merrill classifies expansion as a soft cost because it does not appear in current budgets. However, in considering options for a biomedical information resource, the future expansion costs for different options is an important facet.
- *Risk of loss (or loss expectancy)* "is a calculated cost associated with the probability of losing data or digital content. The loss can occur from a variety of sources, including media failure, theft, corruption, transmission error, sabotage, etc." (Merrill, 2017). While Merrill posits that this cost can be calculated in the commercial setting, it manifests as opportunity cost in the biomedical research setting, which is difficult to quantify. It is difficult to assign a dollar value to delayed or forgone discoveries, especially as the nature of those potential discoveries is challenging to foresee. In the biomedical research setting, this cost might have to be approached as setting a tolerable likelihood of loss and evaluating alternative approaches by whether they fall within that likelihood.
- *Cost of performance* "is often a perceived issue with how long IT tasks take to complete. In a few cases performance can be linked to company revenue or direct costs, but usually is a point of complaint for the IT department. If projects can demonstrate business impact due to slow or inconsistent access or retrieval, then performance can become a hard cost to the preservation architecture" (Merrill, 2017). This soft cost is one of the most relevant—but also one of the most difficult—to measure. Limitations on data, search, and services can all restrict or delay tasks that researchers want to perform with the information in a repository, thus retarding or reducing discovery (or, in cases in which a repository supports clinical uses, compromising treatment). Thus, there are at least two aspects to these costs: the additional time for tasks that are eventually completed and the lost knowledge from tasks that are forgone. While the first aspect can be characterized by the relative ease of performance for alternative approaches, even a qualitative comparison of the second aspect seems daunting. Another complication to this soft cost is that it needs to be evaluated in the context of available alternatives. In considering approaches to Repository C (or whether to establish Repository C at all), it is necessary to consider whether there is an alternative Repository D that would support at least some of the tasks that Repository C would support.
- *Cost of procurement* is "the time it takes to select, quote, bid, negotiate and purchase infrastructure for digital preservation . . . This cycle is heavily weighted with staff from procurement. This cycle tends to occur every few years when older equipment needs to be replaced. In cloud or consumption models, the provisioning process is self-serviced. This cost is different from provisioning time (see above) in that time and effort are internal to IT, and lead-times are planned such that capacity is ready when it is needed. Provisioning time for required future resources will be reduced since the procurement process

BOX D.1 Continued

is greatly simplified" (Merrill, 2017). How relevant this soft cost is to the biomedical domain depends to a large degree on the particular organizational structure around the group providing an information resource. If the group is in a large agency, then some cost of procurement could be soft costs, since some responsibility for procurement might fall to another unit within the organization. In contrast, a group within a nonprofit dedicated to maintaining a specific information resource could be managing procurement itself. In either case, future procurement costs need to be considered.

Thus, while soft costs generally cannot be quantified easily, they can be compared across approaches. The committee believes that it is possible to do so in a disciplined manner. For example, for discovery time, one could make a list of search types that a repository user might want and tabulate for each alternative approach whether or not it supports each search type. Similarly, for cost of performance, one could make a list of likely tasks that a researcher might want to perform with the data. Then for a given approach, one could determine whether it "Does Not Support," "Partially Supports," or "Supports" each particular task. With such information, one could easily determine whether Approach C "dominates" Approach D, in terms of C having equal or lower soft costs than D across all facets, or isolate the trade-off points between C and D: on what specific facets does C have higher or lower soft costs than D?

It is tempting to ignore soft costs in forecasting, since they may not be quantitative or they accrue outside the immediate organizational unit. However, they help characterize the usability and value of data for a community. Considering only hard costs might drive one to select options with low direct costs but that are difficult to use and provide little value (in which case, why support the resource at all?).

REFERENCE

Merrill, D. 2017. Economic perspectives for long-term digital preservation: Achieve zero data loss and geo-dispersion. White Paper, Hitachi Data Systems.

E

Template to Map Cost Drivers to Data Resource Properties

Table E.1 is a template that compiles all the questions regarding the cost drivers described in Chapter 4 in a single place. Information about the definitions of all terms used is found in Chapter 4.

TABLE E.1 Cost-Driver Template

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
A. **Content**			
A.1	**Size (volume and number of items)**	1. How many files will be in a single data submission? 2. How large is an average data submission in total? 3. Are the data sizes likely to stay stable over the life of the resource? 4. What is the total amount of data expected? 5. In what kind of medium will data be captured in the short and long terms?	
A.2	**Complexity and Diversity of Data Types**	1. How complex is the underlying structure of the data? 2. How are the included data to be organized? 3. How complex is the experimental paradigm that produced the data? 4. What sort of additional files might be necessary to upload with the data to properly understand them? 5. How many different data types are being produced? 6. What are the relationships among these data types (e.g., are the data correlated)?	
A.3	**Metadata Requirements**	1. How much metadata must be stored with each data object to make them findable, accessible, interoperable, and reusable (FAIR)? 2. Will metadata be entered manually by the submitter/curator? 3. Will the data to be deposited include a data schema, or will one be generated? 4. Is the provenance of a data set sufficiently described, or will it need to be? 5. How much metadata can be extracted computationally?	
A.4	**Depth Versus Breadth**	Will the repository be restricted to certain data classes or types that the repository must support?	

continued

TABLE E.1 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
A.5	**Processing Level and Fidelity**	1. Do the raw data need to be stored? 2. Do processed data need to be stored? 3. Are there compression algorithms that can reduce the file size without compromising fidelity? 4. What kind of data structure requirements will the resource have? 5. Is the data contributor or the repository responsible for any restructuring necessary? 6. How is the data structure verified?	
A.6	**Replaceability of Data**	1. Is the archive the primary steward of the data, or do copies exist elsewhere? 2. Can the data be easily recreated?	
B. **Capabilities**			
B.1	**User Annotation**	1. Will the repository have to provide user annotation capabilities? 2. What is the nature of these annotations? 3. Are they provided by humans or machines, and how will they be authenticated? 4. Are permissions required to annotate the data?	
B.2	**Persistent Identifiers**	1. What persistent identifier (PID) scheme will be used by the archive? 2. Is there a cost associated with using the PID? 3. How many objects need to be identified? 4. Who will be responsible for keeping the PIDs resolvable?	
B.3	**Citation**	1. Will users be able to create arbitrary subsets of data files and mint a PID for citation? 2. Will the repository provide machine-readable metadata for supporting data citation? 3. Will the repository provide export of data citations for use in reference managers?	
B.4	**Search Capabilities**	1. Will the repository provide a search capability for data sets? 2. How much of the metadata will be included in search? 3. How complex are the queries that will be supported? 4. What types of features for search will be provided? 5. Will the repository deploy services to search the data directly?	
B.5	**Data Linking and Merging**	1. Will the data require/benefit from linkages to other related items? 2. Will the resource provide the ability to combine data across records based on common entities/standards?	
B.6	**Use Tracking**	1. Will the resource provide the ability to track uploads, views, and downloads? 2. If so, and if made available to users, how will this information be made available? 3. Will the resource track data citations to its data?	
B.7	**Data Analysis and Visualization**	1. What types of data analyses and visualizations will the repository support? 2. What types of other data operations will the repository support (e.g., file conversions, sequence comparison)? 3. Do these services require significant computational resources? 4. Who will pay for computational resources?	
C. **Control**			
C.1	**Content Control**	1. Will all appropriate data be accepted or will there be a review process? 2. Will the review process be automated or will it require human oversight?	

TABLE E.1 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
C.2	**Quality Control**	1. What quality control process will the repository support? 2. Will these be automated or require human oversight? 3. What level of data correctness will be required, and how will it be validated? 4. What gaps in the data at the record or field level will be tolerable? 5. Will any of the data be time sensitive, and how will data currency be ensured? 6. How will duplication within or between data sets be addressed? 7. Will prevalidation guidelines or routines be distributed by the resource to the data contributors? 8. Will human curation be necessary?	
C.3	**Access Control**	1. What types of access control are required for the repository (e.g., will there be an embargo period)? 2. At what level are they instituted (e.g., individual users, individual data sets)? 3. Does use of the data require approval by a data access committee?	
C.4	**Platform Control**	Are there restrictions on the type of platform that may or must be used?	
D. **External Context**			
D.1	**Resource Replication**	Is there a requirement to replicate the information resource at multiple sites (i.e., mirroring)?	
D.2	**External Information Dependencies**	Will the resource be dependent on information maintained by an outside source?	
D.3	**Distinctiveness**	Are there existing resources available that provide similar types of data and services?	
E. **Data Life Cycle**			
E.1	**Anticipated Growth**	1. Is the repository expected to continuously grow over its lifetime? 2. Is the likely rate of growth in data and services known? 3. Is the use of the repository likely to grow over time? 4. Is the likely growth of the user base known?	
E.2	**Update and Versions**	1. Will the deposited data require updates (e.g., in response to new data or error corrections)? 2. Will prior versions of the data need to be retained and made available locally or in a different resource? 3. How frequently will individual data sets be updated?	
E.3	**Useful Lifetime**	1. Are the data to be housed likely to have a limited period of usefulness? 2. Does the resource have a defined period of time for which it will operate? 3. Does the resource have to provide a guarantee that the data will be available for a finite period of time (e.g., 10 years)?	
E.4	**Offline and Deep Storage**	1. Can the resource take advantage of offline storage for data that are not heavily used? 2. Does the resource have a plan for moving unused data to deep storage (i.e., State 3)?	

continued

TABLE E.1 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
F. **Contributors and Users**			
F.1	**Contributor Base**	1. Is the number of contributors known? If not, can it be estimated? 2. Are all the data originating from the same source (e.g., a single instrument or a single organization)? 3. How will data be transferred into the data resource (e.g., periodic large batches, more frequent smaller data sets, constantly streamed, by physical transfer)? 4. Will the data be pushed by the contributor or pulled by the resource? 5. Are there direct or indirect fees associated with acquiring the data from a source? 6. Will a data steward be available from among the contributors to assist with any data integration into the data resource?	
F.2	**User Base and Usage Scenarios**	1. How many users will likely access the data? 2. What will be the frequency of access? 3. How will users access the data? 4. Will the resource be building analysis tools? 5. Will the resource support individual file download or bulk download? 6. Will there be any fees for downloading/accessing the data? 7. How many different types of users must be supported?	
F.3	**Training and Support Requirements**	1. Will training for resource use be offered? 2. What form will the training take? 3. Will a “help desk” be provided? 4. When does live help need to be available? 5. What is the expected skill level of the user base?	
F.4	**Outreach**	1. Does the existence of the repository need to be advertised? 2. How many conferences per year should resource representatives attend? 3. Will the resource have a booth at the conference for live demos or conduct hands-on tutorials? 4. Are users required by funders or journals to deposit data in the repository?	
G. **Availability**			
G.1	**Tolerance for Outages**	1. What is the tolerance for outages of the resource? 2. What measures will be taken to avoid and mitigate outages? 3. How quickly and completely does the resource need to recover from an outage?	
G.2	**Currency**	1. How often will the data be released? 2. How soon do data need to be made available after they are received?	
G.3	**Response Time**	1. Are there requirements for response time for service? 2. Are there requirements for responses from humans?	
G.4	**Local Versus Remote Access**	1. Does the resource require that any data be shipped via physical media? 2. Will the resource be built using commercial clouds? 3. Do users have to travel to the resource to use the data?	
H. **Confidentiality, Ownership, and Security**			
H.1	**Confidentiality**	1. Will any of the data require special protections? 2. Will any of the data have embargo periods or embargo-related limitations that may entail costs? 3. Are there any audit requirements for who has accessed or downloaded the data?	
H.2	**Ownership**	1. If data are contributed from multiple sources, will there be a need to process multiple kinds of release forms? 2. Will all the data be released by the data resource under the same license, or will different permissions be assigned to different data sets? 3. Will data submission agreements be necessary?	

TABLE E.1 Continued

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
H.3	**Security**	1. What measures need to be taken to ensure the integrity and availability of the data? 2. Do these measures require using protected computing, storage, or networking platforms?	
I. **Maintenance and Operations**			
I.1	**Periodic Integrity Checking**	1. What processes will be put in place for checking the integrity of the hardware, software, and data? 2. How frequently will these checks be performed?	
I.2	**Data-Transfer Capacity**	Will the bandwidth available to the resource be sufficient for the data sizes and rates required?	
I.3	**Risk Management**	1. Will the repository be solely responsible for risk mitigation? 2. Is a response plan for unexpected termination required?	
I.4	**System-Reporting Requirements**	What types of system reporting will the resource be required to do?	
I.5	**Billing and Collections**	Will there be charges for use of the resource?	
J. **Standards, Regulatory, and Governance Concerns**			
J.1	**Applicable Standards**	1. How many different standards will the resource have to support? 2. Do these standards exist? a. If not, is the resource expected to lead their development? b. What is the plan for accepting data while standards are in development? c. If so, are the standards mature (i.e., how much are they expected to evolve)? 3. Are the data validators and converters available for the standards, or do they have to be developed? 4. What is the plan for “retrofitting” data that have been uploaded without the standards in place? 5. How frequently will the standards update? 6. Do the standards require spatial transformations (e.g., will they need to be aligned to a common coordinate system)? 7. How many file formats will be supported? 8. Is there an open file format available?	
J.2	**Regulatory and Legislative Environment**	1. What laws and regulations cover the data and operation of the resource? 2. Is the resource covered by an open-records act?	
J.3	**Governance**	1. Does the resource need to maintain an external advisory board? 2. Does the resource set policy for itself, or is it part of a larger organization?	
J.4	**External Consultation**	1. Will external stakeholders be consulted for initial design? 2. Will external stakeholders be consulted on an ongoing basis?	

F

Comparison of the Contents Across the Three Data States

Three hypothetical data resources—one representing each of the three data states described in Chapter 2—are described in Box F.1, which provides descriptions of the three hypothetical data states used in tabulated comparisons. Table F.1 then describes characteristics of those hypothetical data resources for the purpose of comparison.

BOX F.1
Comparison of the Contents Across the Three Data States

Three examples of hypothetical biomedical information resources are provided below, one for each of the data states described in Chapter 2. Characteristics of data that might be found in each of the data state platforms are described.

State 1 Example:

A research group is studying the effects of mutation of a particular gene in a model organism—say, mouse. The group is collecting several kinds of data, including exome data, gene sequences, cell and tissue images, history and treatment of the individual mice used, and biosample tracking data. The group is also downloading data on the specific gene under study and a reference mouse genome (or portion thereof). The group gathers these data in a collective workspace, perhaps on a local network drive, or is using a commercial service, such as Dropbox.

State 2 Example:

A public repository holds longitudinal human genome, exome, and ribonucleic acid (RNA) sequence information. Each individual in the collection has been sequenced with one or more modes at multiple points in time. Such data could be used for various studies, such as early disease markers, onset of mutations, and results of drug treatment. Contributors of data are expected to "reconsent" any participant before his or her information is submitted to the repository. The repository applies standard processing pipelines to certain uploaded data, such as generating variant calls from genome sequence data. The information in the repository is "data at work" in the sense that users can perform certain operations on the data within the repository.

State 3 Example:

A university maintains a data archive for projects completed on campus to meet university, government, and research sponsor data-retention requirements. The archive might be viewed as one holding data that are not expected to be used in place, rather than the active data described in the previous examples. Investigators wanting to use data from the archive will generally download them into their own computing environments and interact with them there. Data contributors are investigators at the university, but the potential users may be quite broad.

TABLE F.1 Characteristics of Hypothetical Information Resources for Comparison

State 1 (A) Content Characteristics	State 2 (A) Content Characteristics	State 3 (A) Content Characteristics
Small numbers of items and total storage; moderately diverse data types held. Metadata requirements, if any, are informal. Coverage is broad enough to include the types of information generated or downloaded. Raw and more processed versions of data stored. The data are replaceable, but rerunning experiments would be required.	Modest number of items in the repository: thousands of individuals, with 1-3 sequence types at 2-10 time points. Some items are large (e.g., full-genome sequences). Limited number of data types—a few types of sequence data plus text for medical histories. Certain demographic and medical data in a specific structured format, to support searching expected from contributors. Repository is narrowly focused on sequence data. Submissions to the repository have some level of processing (assembled sequence or RNA abundance, for example, rather than raw sequence reads). Since the data reflect the past states of individuals, they are replaceable only if biospecimens have been retained.	Large archive (in bytes), but the number of items will be proportional to the number of projects—a unit of storage may include all the data deposited from a single project. Must accept data of any type but does not have to perform type-specific operations (e.g., gene-sequence matching). There will be metadata requirements for data sets deposited, although they might not be extensive (e.g., information on the depositor, project, sponsor, citations to related papers, textual description). Archive holdings will be broad and correspond to the range of investigations at the university but generally not deep unless there is a large concentration of work in one area. The data span a wide range of processing levels and fidelities; some may be unique and nonreplaceable.
State 1 (B) Capability Characteristics	**State 2 (B) Capability Characteristics**	**State 3 (B) Capability Characteristics**
Passive repository with few capabilities; users likely extract data from it and work with it on their own computers. Some keyword-based searching might be possible.	Supports user annotation, has persistent identifiers, and provides means to cite both the repository and the original contributors of data. Supports searches based on the structured metadata supplied by contributors and on data characteristics (e.g., number of time points for an individual or type of sequence data). All data items for a given individual linked together. Data usage is tracked at per-item and per-user levels. Supports a range of analyses and visualizations locally in the repository (e.g., extracting a time series for a gene across all items for an individual or charting the differences between two items).	User annotation is unlikely; data sets not expected to change or be augmented after deposit. University might provide persistent identifiers for data sets, but persistent identifiers (e.g., Digital Object Identifiers associated with a data set may already exist. Citation supported only full-data set level. Search capabilities limited to faceted search over metadata, augmented with keyword search of textual data elements. Hierarchical browsing of data sets along thematic lines possible. Linking and merging of data items not expected. Data set download numbers will be tracked. Analysis and visualization of data not supported on the platform.
State 1 (C) Control Characteristics	**State 2 (C) Control Characteristics**	**State 3 (C) Control Characteristics**
Informal content control (e.g., laboratory policies on what is appropriate for shared workspace). Quality control focused on experiment protocols for collecting project data. Access control restricted to project members and likely relies on file-system permissions. No platform restrictions.	All repository submissions are curated for formatting, conformance to appropriate standards, quality issues, and completeness of metadata. Access is carefully controlled. Users and their proposed studies using the data must be approved by a review board. Since individuals can be reidentified from genomic information, the repository needs to run on a high-trust platform.	Mainly related to whether data are appropriately packaged and documented for deposit; potential data size limits accepted. Quality control is limited to metadata completeness and correctness. Only appropriate university community members may deposit data sets. Access for downloading may be limited if sensitive or proprietary information is in any of the data sets. Simply searching the archive will not require authorization. Possible restrictions on platform dictated by the university if the school wants to keep the archive on its own server or if there is an institutional arrangement with a particular commercial cloud provider.

TABLE F.1 Continued

State 1 (D) External Context Characteristics	State 2 (D) External Context Characteristics	State 3 (D) External Context Characteristics
No requirements to replicate data outside resource (workspace is internal to project). Possible dependencies on external data sets (e.g., gene variant cells relative to a reference mouse genome). Data distinctiveness dependent on whether other groups are studying the same gene.	To retain access control, the repository not replicated elsewhere. Data are mostly self-contained, although some of the metadata may be drawn from controlled vocabularies. The repository is fairly unique, with its focus on the time dimension.	No obligation to replicate the archive as a whole at other sites; specific data sets may be required to be replicated offsite, but data creator responsible for such replication. Generally, no dependencies on external data sets; archive unable to track or maintain external dependencies (data in read-only mode). No general characterization of distinctiveness of archive as a whole; that will vary among data sets.
State 1 (E) Data Life-Cycle Characteristics	**State 2 (E) Data Life-Cycle Characteristics**	**State 3 (E) Data Life-Cycle Characteristics**
Likely steady growth over course of project; growth after project if used for new studies. Rare updates to raw data; some data processing might be repeated with different parameters or reference sets; formal versioning unlikely. Useful lifetime of data might extend to follow-on studies in the same laboratory. If data useful to wider community, then they will probably be made available through a public repository. Fully processed might be moved to archival or offline storage if storage needs of project outstrip workspace size.	Repository will likely grow at an increasing rate each year as more individuals are added and new sequences are submitted for existing individuals. Repository updated incrementally as new data arrive and are approved. Certain information may be versioned, such as different versions of variant calls relative to different versions of the human reference genome. Given the temporal aspect, the data are unlikely to be superseded by other sources. Data in the repository are expected to stay online.	Archive growth will occur at accelerated rate; number of data sets deposited annually may be stable, but amount of data collected per project is expected to increase. Data sets in archive not typically updated (except, perhaps, if corrected or withdrawn). New versions of data sets may be deposited. Useful lifetime of data sets varies. Large portions of the archive may be held in offline or deep storage, with only metadata kept online for searching.
State 1 (F) Contributor and User Characteristics	**State 2 (F) Contributor and User Characteristics**	**State 3 (F) Contributor and User Characteristics**
Project members; no requirements for accommodating large numbers of contributors or users. No outreach required. Informal training and support provided by existing laboratory members.	Contributors and users from the same research community. The number of contributors will be limited compared to general sequence repositories (most sequencing projects not collecting data across time). The user base is relatively larger and could include most of the contributors, as they may want to compare their data to other sources. Most users will carry out initial data search and analysis on platform, downloading only small subsets or analysis results. Contributors and users will require training and support, and there will be outreach to both groups.	Contributors include investigators currently or formerly with the university. Users could be almost anyone. Deposits on the order of tens per week. Archive searches common, but downloads infrequent for most data sets. Training and support requirements skewed toward data contributors (e.g., education on what can and should be deposited; consultation on data preparation). Outreach activities focused within the university to make researchers aware of archive and what should be deposited.

continued

TABLE F.1 Continued

State 1 (G) Availability Characteristics	**State 2 (G) Availability Characteristics**	**State 3 (G) Availability Characteristics**
Short-duration outages inconvenient but tolerable. Currency of data is important. Interactive response times likely not necessary. Local access the norm, but remote access might be desirable for off-site collaborators.	Short outages are permissible for maintenance or upgrades. Currency of the data is not critical; contributors may be submitting data well after the collection point. Response time approval of submissions should be within a week. Searches should run interactively, but some of the more complex analyses might take minutes or hours. It is important that there be enough computing resources to support a modest number of simultaneous users. The resource is available remotely via a web interface.	Must not lose submissions, but brief outages acceptable. Currency not critical (data deposited at project end). Some requirements of investigators for data sharing within a time frame following publication or project end; deposits may need to be vetted and brought online quickly. Archive searching will be interactive, but downloads within hours or days tolerable, especially for off-line material. The archive is remotely accessible.
State 1 (H) Confidentiality, Ownership, and Security	**State 2 (H) Confidentiality, Ownership, and Security**	**State 3 (H) Confidentiality, Ownership, and Security**
No personal health information (no human subjects). Potential ownership concerns for data downloaded from elsewhere, but not for data generated in the laboratory. Security important for keeping data and results private until publication and preventing loss or damage to data by unauthorized users.	Data in the repository come from human subjects and are confidential. All data usage should be auditable, to document compliance with access policies. Inclusion of data in the repository is consented, but participants allowed to revoke consent. Thus, submitted data and any additional results derived therefrom must be traceable to a participant or participants. Repository operators will arrange periodic external audits of their security practices.	Data sets may contain personal or proprietary information and thus are confidential. Processes necessary for reviewing data access requests. Ownership of most data resides with the university (or possibly the investigator), but data produced on contract research might be externally owned and require tracking. Security against unauthorized modifications and against unauthorized access (if confidentiality or ownership issues involved) required.
State 1 (I) Maintenance and Operations Characteristics	**State 2 (I) Maintenance and Operations Characteristics**	**State 3 (I) Maintenance and Operations Characteristics**
Hardware and software integrity checked by others if workspace is on a network drive or commercial service. Data integrity checks (if any) performed by project members. Backups will be managed by project team or affiliated staff if workspace is on a local server; otherwise, others will manage backups to mitigate hardware failure risk. Minimal system-reporting requirements (e.g., lists of space usage by user; monitoring remaining free space).	Integrity checks on the data are conducted monthly, as well as a report produced on any anomalies. Monthly reports required on size of holdings, per-item and per-user access frequency, and compute usage. The operators of the resource cannot assume that contributors retain their data on a long-term basis; hence, they are responsible for risk management.	University responsible for integrity checking of hardware and software if archive is maintained locally. Accidental corruption of data sets unlikely (given no data set updating). Offline data sets should be checked periodically for readability. Risk-of-loss management necessary if archive contains copies of record of data sets. Minimal systems-reporting requirements (e.g., monthly download summaries and space usage).
State 1 (J) Standards, Regulatory, and Governance Concerns	**State 2 (J) Standards, Regulatory, and Governance Concerns**	**State 3 (J) Standards, Regulatory, and Governance Concerns**
Relevant standards for some types of data (e.g., sequencing data and their analysis products), but common software tools are available to generate data according to standards. Possibly some future requirements from project sponsors or host institutions for sharing and archiving data. Likely no governing body for project-specific resource.	Community standards exist for all the main types of sequence-based data hosted; repository conforms to these. Sequence data may not be explicitly categorized as personally identifiable information in some government regulations (e.g., Health Insurance Portability and Accountability Act); they might be in the future and repository operators treat it as such. An advisory board helps develop repository acquisition and use policies.	Archive will enforce archive-level standards for metadata on deposited data sets but will not check or enforce data set-specific standards. Main source of regulation and governance will be based on university rules, policies, and possible oversight from offices or committees on campus.

G

Committee Biographical Information

DAVID S. C. CHU, *Chair*, served as president of the Institute for Defense Analyses (IDA) 2009-2020. IDA is a nonprofit corporation operating in the public interest. Its three federally funded research and development centers provide objective analyses of national security issues and related national challenges, particularly those requiring extraordinary scientific and technical expertise. Dr. Chu served in the Department of Defense as Under Secretary of Defense for Personnel and Readiness from 2001 to 2009 and earlier as Assistant Secretary of Defense and Director for Program Analysis and Evaluation from 1981 to 1993. From 1978 to 1981, he was the Assistant Director of the Congressional Budget Office for National Security and International Affairs. Dr. Chu served in the U.S. Army from 1968 to 1970. He was an economist with the RAND Corporation from 1970 to 1978, director of RAND's Washington Office from 1994 to 1998, and vice president for its Army Research Division from 1998 to 2001. He earned his doctorate in economics, as well as a bachelor of arts in economics and mathematics, from Yale University. Dr. Chu is a member of the Defense Science Board and a fellow of the National Academy of Public Administration. He is a recipient of the Department of Defense Medal for Distinguished Public Service with Gold Palm, the Department of Veterans Affairs Meritorious Service Award, the Department of the Army Distinguished Civilian Service Award, the Department of the Navy Distinguished Public Service Award, and the National Academy of Public Administration's National Public Service Award.

ILKAY ALTINTAS is the chief data science officer at the San Diego Supercomputer Center (SDSC), University of California, San Diego (UCSD), where she is also the founder and director for the Workflows for Data Science Center of Excellence and a fellow of the Halicioglu Data Science Institute. In her various roles and projects, she leads collaborative multidisciplinary teams with a research objective to deliver impactful results through making computational data science work more reusable, programmable, scalable, and reproducible. Since joining SDSC in 2001, she has been a principal investigator and a technical leader in a wide range of cross-disciplinary projects. Her work has been applied to many scientific and societal domains including bioinformatics, geoinformatics, high-energy physics, multiscale biomedical science, smart cities, and smart manufacturing. She is a co-initiator of the popular open-source Kepler Scientific Workflow System and the co-author of publications related to computational data science at the intersection of workflows, provenance, distributed computing, big data, reproducibility, and software modeling in many different application areas. Among the awards she has received are the 2015 Institute of Electrical and Electronics Engineers (IEEE) Technical Committee on Scalable Computing Award for Excellence in Scalable Computing for Early Career Researchers and the 2017 Association for Computing Machinery (ACM)

Special Interest Group on High Performance Computing's Emerging Woman Leader in Technical Computing Award.

GOLAM SAYEED CHOUDHURY is the associate dean for research data management and Hodson Director of the Digital Research and Curation Center at the Sheridan Libraries of Johns Hopkins University. Choudhury is also a member of the executive committee for the Institute of Data Intensive Engineering and Science based at Johns Hopkins. Choudhury is a President Obama appointee to the National Museum and Library Services Board. He was a member of the National Academies Board on Research Data and Information and the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. He has testified for the Research Subcommittee of the Congressional Committee on Science, Space, and Technology. He was a member of the board of the National Information Standards Organization, OpenAIRE2020, DuraSpace, the Inter-university Consortium for Political and Social Research (ICPSR) Council, Digital Library Federation advisory committee, Library of Congress National Digital Stewardship Alliance Coordinating Committee, Federation of Earth Scientists Information Partnership Executive Committee, and the Project MUSE Advisory Board. Choudhury was a member of the EDUCAUSE Center for Analysis and Research Data Curation Working Group. He has been a Senior Presidential Fellow with the Council on Library and Information Resources, a lecturer in the Department of Computer Science at Johns Hopkins, and a research fellow at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. He is the recipient of the 2012 Online Computer Library Center, Incorporated/Library and Information Technology Association Kilgour Award. Choudhury has served as principal investigator for projects funded through the National Science Foundation, Institute of Museum and Library Services, Library of Congress National Digital Information Infrastructure and Preservation Program, Alfred P. Sloan Foundation, Andrew W. Mellon Foundation, Microsoft Research, and a Maryland-based venture capital group. He is the product owner for the Data Conservancy, which focuses on the development of data curation infrastructure, and the Public Access Submission System, which supports simultaneous submission of articles to PubMed Central and institutional repositories. He has oversight for data curation research and development and data archive implementation at the Sheridan Libraries at Johns Hopkins University. Choudhury has published articles in journals such as the *International Journal of Digital Curation*, *D-Lib*, the *Journal of Digital Information*, *First Monday*, and *Library Trends*. He has served on committees for the Digital Curation Conference, Open Repositories, Joint Conference on Digital Libraries, and Web-Wise. He has presented at various conferences including EDUCAUSE, the Coalition for Networked Information, Jisc-Coalition for Networked Information, Digital Library Federation, American Library Association, Association of College and Research Libraries, and international venues including the International Federation of Library Associations and Institutions, the Kanazawa Information Technology Roundtable, eResearch Australasia, the North America-China Conference, eResearch New Zealand, and the Arabian-Gulf Chapter of the Special Libraries Conference.

MARGARET LEVENSTEIN is director of ICPSR; research professor at the Institute for Social Research and the School of Information; and adjunct professor of business economics and public policy at the Stephen M. Ross School of Business. She has taught economics at the University of Michigan since 1990. She serves as co-executive director of the Michigan Federal Statistical Research Data Center (FSRDC) and co-chair of the Executive Committee of the FSRDC national network. She is the associate chair of the American Economic Association's Committee on the Status of Women in the Economics Profession and past president of the Business History Conference. She is principal investigator of CenHRS, a Sloan Foundation–funded project building an enhancement to the Health and Retirement Study based on linkages to administrative and survey data on Health and Retirement Study employers and co-workers. She is principal investigator of a National Science Foundation (NSF)-funded project to establish a repository of linked data and data linkage algorithms at ICPSR; a Sloan and NSF-funded effort to establish a Researcher Passport using open badges for credentialed, trusted researchers to access restricted data; and an NSF-funded project conducting experiments to encourage citizen-scientists to improve research metadata. She received a Ph.D. in economics from Yale University and a B.A. from Barnard College, Columbia University. She is the author of numerous studies on competition and collusion, the development of information systems, and using "organic" data to improve social and economic measurement.

CLIFFORD LYNCH has been the executive director of the Coalition for Networked Information (CNI) since 1997. CNI, jointly sponsored by the Association of Research Libraries and EDUCAUSE, includes about 200 member organizations concerned with the intelligent uses of information technology and networked information to enhance scholarship and intellectual life. CNI's wide-ranging agenda includes work in digital preservation, data intensive scholarship, teaching, learning and technology, and infrastructure and standards development. Prior to joining CNI, Lynch spent 18 years at the University of California Office of the President, including the last 10 as Director of Library Automation. Lynch, who holds a Ph.D. in computer science from the University of California, Berkeley, is an adjunct professor at Berkeley's School of Information. He is both a past president and recipient of the Award of Merit of the American Society for Information Science and a fellow of the American Association for the Advancement of Science, the ACM, and the National Information Standards Organization. He served as co-chair of the National Academies Board on Research Data and Information from 2011 to 2016, and he is active on numerous advisory boards and visiting committees. His work has been recognized by the American Library Association's Lippincott Award, the EDUCAUSE Leadership Award in Public Policy and Practice, and the American Society for Engineering Education's Homer Bernhardt Award.

DAVID MAIER is Maseeh Professor of Emerging Technologies at Portland State University. Prior to his current position, he was on the faculty at the State University of New York, Stony Brook, and Oregon Graduate Institute. He has spent extended visits with Inria, the University of Wisconsin, Madison, Microsoft Research, and the National University of Singapore. He is the author of books on relational databases, logic programming, and object-oriented databases, as well as papers on database theory, object-oriented technology, scientific databases, and dataspace management. He is a recognized expert on the challenges of large-scale data in the sciences. He received an NSF Young Investigator Award in 1984 and was awarded the 1997 ACM Special Interest Group on Management of Data's Innovations Award for his contributions in objects and databases. He is also an ACM Fellow and IEEE Senior Member. He holds a dual B.A. in mathematics and in computer science from the University of Oregon (Honors College, 1974) and a Ph.D. in electrical engineering and computer science from Princeton University (1978).

CHARLES MANSKI has been Board of Trustees Professor in Economics at Northwestern University since 1997. He previously was a faculty member at the University of Wisconsin, Madison (1983-1998), the Hebrew University of Jerusalem (1979-1983), and Carnegie Mellon University (1973-1980). He received his B.S. and Ph.D. in economics from the Massachusetts Institute of Technology (MIT) in 1970 and 1973, respectively. He has received honorary doctorates from the University of Rome "Tor Vergata" (2006) and the Hebrew University of Jerusalem (2018). Manski's research spans econometrics, judgment and decision, and analysis of public policy. He is author of *Public Policy in an Uncertain World* (Harvard, 2013), *Identification for Prediction and Decision* (Harvard, 2007), *Social Choice with Partial Knowledge of Treatment Response* (Princeton, 2005), *Partial Identification of Probability Distributions* (Springer, 2003), *Identification Problems in the Social Sciences* (Harvard, 1995), and *Analog Estimation Methods in Econometrics* (Chapman and Hall, 1988); co-author of *College Choice in America* (Harvard, 1983); and co-editor of *Evaluating Welfare and Training Programs* (Harvard, 1992) and *Structural Analysis of Discrete Data with Econometric Applications* (MIT, 1981). He has served as director of the Institute for Research on Poverty (1988-1991), chair of the Board of Overseers of the Panel Study of Income Dynamics (1994-1998), and chair of the National Research Council Committee on Data and Research for Policy on Illegal Drugs (1998-2001). Editorial service includes terms as editor of the *Journal of Human Resources* (1991-1994), co-editor of the *Econometric Society Monograph Series* (1983-1988), member of the editorial board of the *Annual Review of Economics* (2007-2013), member of the Report Review Committee of the National Research Council (2010-2018), and associate editor of the *Annals of Applied Statistics* (2006-2010), *Econometrica* (1980-1988), *Journal of Economic Perspectives* (1986-1989), *Journal of the American Statistical Association* (1983-1985, 2002-2004), and *Transportation Science* (1978-1984). Manski is an elected member of the National Academy of Sciences. He is an elected fellow of the American Academy of Arts and Sciences, the Econometric Society, the American Statistical Association, and the American Association for the Advancement of Science, distinguished fellow of the American Economic Association, and corresponding fellow of the British Academy.

MARYANN MARTONE is a professor emerita at UCSD but still maintains an active laboratory and currently serves as the chair of the University of California Academic Senate Committee on Academic Computing and Communications. She received her B.A. from Wellesley College in biological psychology and ancient Greek and her Ph.D. in neuroscience from UCSD. She started her career as a neuroanatomist, specializing in light and electron microscopy, but her main research for the past 15 years focused on informatics for neuroscience (i.e., neuroinformatics). She led the Neuroscience Information Framework (NIF), a national project to establish a uniform resource description framework for neuroscience, and the National Institute of Diabetes and Digestive and Kidney Diseases Information Network (dkNET), a portal for connecting researchers in digestive, kidney, and metabolic disease to data, tools, and materials. She just completed 5 years as editor-in-chief of *Brain and Behavior*, an open-access journal, and has just launched a new journal as editor-in-chief, *NeuroCommons*, with BMC. Dr. Martone is past president of FORCE11, an organization dedicated to advancing scholarly communication and e-scholarship. She completed 2 years as the chair of the Council on Training, Science, and Infrastructure for the International Neuroinformatics Coordinating Facility and is now the chair of the Governing Board. Since retiring, she served as the director of biological sciences for Hypothesis, a technology nonprofit developing an open annotation layer for the web (2015-2018), and founded SciCrunch, a technology start-up based on technologies developed by NIF and dkNET.

ALEXA T. McCRAY 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 NIH, she directed the design and development of a number of national information resources, including ClinicalTrials.gov. Before joining NIH, she was on the research staff of IBM's T.J. Watson Research Center. She received her Ph.D. from Georgetown University and for 3 years was on the faculty there. She conducted predoctoral research at MIT. McCray joined Harvard Medical School in 2005, where she was founding co-director of the Center for Biomedical Informatics and associate director of the Francis A. Countway Library of Medicine. McCray was elected to the National Academy of Medicine in 2001. She is chair of the National Academies Board on Research Data and Information. She is a fellow of the American Association for the Advancement of Science, a fellow of the American College of Medical Informatics (ACMI), an honorary fellow of the International Medical Informatics Association, and a founding fellow of the International Academy of Health Sciences Informatics. She is a past president of ACMI and a past member of the boards of both the American Medical Informatics Association and the International Medical Informatics Association. She is a former editor-in-chief of *Methods of Information in Medicine*, and she is a past member of the editorial board of the *Journal of the American Medical Informatics Association*. She chaired the 2018 National Academies of Sciences, Engineering, and Medicine consensus study entitled Open Science by Design: Realizing a Vision for 21st Century Research.

MICHELLE MEYER is an assistant professor and associate director, research ethics, in the Center for Translational Bioethics and Health Care Policy at Geisinger, a large, integrated health system in Pennsylvania and New Jersey, where she chairs the Institutional Review Board (IRB) Leadership Committee and directs the Research Ethics Advice and Consultation Service. She is also faculty co-director of Geisinger's Behavioral Insights Team (a.k.a. "nudge unit") in Geisinger's Steele Institute for Health Innovation. Her empirical and normative research focuses on judgment and decision making by patients, clinicians, research participants, and IRBs that has implications for law, ethics, and policy. She has served on the advisory board of the Social Science Genetic Association Consortium; the board of directors of the Open Humans Foundation (formerly PersonalGenomes.org); the Ethics and Compliance Advisory Board of PatientsLikeMe; the American Psychological Association's Commission on Ethics Processes; the ClinGen Working Group on Complex Diseases; a National Academy of Medicine/Patient-Centered Outcomes Research Institute working group on generating stakeholder support and demand for health data sharing, linkage, and use; and a Defense Advanced Research Projects Agency–funded technical exchange on complex social systems. She developed a commissioned white paper addressing ethical issues raised by plans for developing a new data-sharing institute. In most of those roles, she has focused on consent; data privacy; and data access and use, especially with respect to genomic data. Immediately before joining the faculty at Geisinger, Meyer was an assistant professor and director of bioethics policy in the Clarkson University-Icahn School of Medicine at Mount Sinai School of Medicine Bioethics Program and adjunct faculty at Albany Law School. Previously, she was an academic fellow at the Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics at Harvard Law School; a Greenwall Fellow in Bioethics and Health Policy at The Johns Hopkins and Georgetown Universities; and a research fellow at the John F. Kennedy School of Government at Harvard. She earned a Ph.D. in religious studies, with a focus on practical ethics, from the University of Virginia under the supervision of James F. Childress and a J.D. from Harvard Law School, where she was an editor of the *Harvard Law Review*. Following law school, she clerked for Judge Stanley Marcus of the U.S. Court of Appeals for the Eleventh Circuit. She graduated summa cum laude from Dartmouth College.

WILLIAM STEAD is chief strategy officer for Vanderbilt University Medical Center (VUMC). In this capacity, he facilitates structured decision making to achieve strategic goals and concept development to nurture system innovation. Dr. Stead received his B.A., M.D., and residency training in internal medicine and nephrology from Duke University. He remained on Duke's faculty in nephrology as the physician in the physician-engineer partnership that developed The Medical Record, one of the first practical electronic medical record systems. He also helped Duke build one of the first patient-centered hospital information systems (IBM's PCS/ADS). He came to VUMC in 1991 and holds appointments as the McKesson Foundation Professor of Biomedical Informatics and Professor of Medicine. For two decades, he guided development of the Department of Biomedical Informatics and operational units providing information infrastructure to support health care, education, and research programs of the Medical Center. He aligned organizational structure, informatics architecture, and change management to bring cutting-edge research in decision support, visualization, natural language processing, data mining, and data privacy into clinical practice. His current focus is on system-based care, learning and research leading toward personalized medicine, and population health management. Dr. Stead is a founding fellow of both the American College of Medical Informatics and the American Institute for Engineering in Biology and Medicine. He served as founding editor-in-chief of the *Journal of the American Medical Informatics Association*. His awards include the Collen Award for Excellence in Medical Informatics and the Lindberg Award for Innovation in Informatics. Most recently, the American Medical Informatics Association named the Award for Thought Leadership in Informatics in his honor. He served as president of the American College of Medical Informatics, chairman of the Board of Regents of the National Library of Medicine, presidential appointee to the Commission on Systemic Interoperability, chair of the National Research Council Committee on Engaging the Computer Science Research Community in Health Care Informatics, and co-chair of the Institute of Medicine Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records. He chairs the National Committee for Vital and Health Statistics of the Department of Health and Human Services and the Technical Advisory Committee of the Center for Medical Interoperability. He is a member of the Council of the National Academy of Medicine and the American Medical Association's Journal Oversight Committee. In addition to his academic and advisory responsibilities, Dr. Stead is a director of HealthStream.

LARS VILHUBER is presently on the faculty of the Department of Economics at Cornell University, executive director of the ILR School's Labor Dynamics Institute, a senior research associate at the ILR School at Cornell University, Ithaca, and affiliated with the U.S. Census Bureau (Center for Economic Studies). He holds a Ph.D. in economics from Université de Montréal, Montreal, Canada, having previously studied economics at the Universität Bonn, Germany, and Fernuniversität Hagen, Germany. He has worked in both academic and government research positions and continues to consult and collaborate with government and statistical agencies in Canada, the United States, and Europe. His research interests lie in the dynamics of the labor market: working with highly detailed longitudinally linked data, he has analyzed the effects and causes of mass layoffs, worker mobility, and the interaction between housing and the local labor market. Over the years, he has also gained extensive expertise on the data needs of economists and other social scientists, having been involved in the creation and maintenance of several data systems designed with analysis, publication, replicability, and maintenance of large-scale code bases in mind. His research in statistical disclosure limitation issues is a direct consequence of his profound interest in making data available in a multitude of formats to the broadest possible audience. His knowledge about various data enclave systems comes from both personal experience and the desire to improve the experience of others. He is data editor of the *American Economic Association* and managing editor of the *Journal of Privacy and Confidentiality*; chair of the Scientific Advisory Committee of the Centre d'accès sécurisé aux données in France; and senior advisor of the New York FSRDCs in the United States.

H

Acronyms

ABIDE	Autism Brain Imaging Data Exchange
ADNI	Alzheimer's Disease Neuroimaging Initiative
AI	artificial intelligence
BCBC	Beta Cell Biology Consortium
BIDS	Brain Imaging Data Structure
BLAST	Basic Local Alignment Search Tool
BRAIN	Brain Research Through Advancing Innovative Neurotechnologies
BVI	business value of information
CCPA	California Consumer Privacy Act
CERN	European Organization for Nuclear Research
CITI	Collaborative Institutional Training Initiative
COMBINE	Computational Modeling in Biology Network
CVI	cost value of information
DICOM	Digital Imaging and Communications in Medicine
DMP	data management plan
DNA	Deoxyribonucleic acid
DOI	Digital Object Identifier
DUOS	Data Use Oversight System
EAB	external advisory board
ERC	European Research Council
FAIR	findable, accessible, interoperable, and reusable
FASTQ	A text-based format used for storing biological sequence data and quality scores

FedRAMP	Federal Risk and Authorization Management Plan
FISMA	Federal Information Security Act
fMRI	functional magnetic resonance imaging
FSRDC	Federal Statistical Research Data Center
GB	gigabyte
GDPR	General Data Privacy Regulation
HDF	hierarchical data format
HIPAA	Health Insurance Portability and Accountability Act of 1996
HSR	human-subjects research
IEDA	Interdisciplinary Earth Data Alliance
IGERT	Integrative Graduate Education and Research Traineeship
INCF	International Neuroinformatics Coordinating Facility
IRB	Institutional Review Board
IT	information technology
IVI	intrinsic value of information
maDMP	machine-actionable data management plan
MeSH	Medical Subject Headings
NCBI	National Center for Biotechnology Information
NDA	National Institute of Mental Health Data Archive
NEMO	Neuroscience Multi ‘Omic Archive
NIA	National Institute on Aging
NIDDK	National Institute of Diabetes and Digestive and Kidney Diseases
NIDM	Neuro Imaging Data Model
NIF	Neuroscience Information Framework
NIH	National Institutes of Health
NIH RePORTER	NIH Research Portfolio Online Reporting Tools
NIMH	National Institute of Mental Health
NITRC	NeuroImaging Tools and Resources Collaboratory
NLM	National Library of Medicine
NNLM	National Network of Libraries of Medicine
NSF	National Science Foundation
NWB	Neurodata Without Borders
O*NET	Occupational Information Network
OASIS	Open Access Series of Imaging Studies
ORCID	Open Researcher and Contributor Identifier
PB	petabyte
PID	persistent identifier, personal identifier
RFA	request for application
RNA	ribonucleic acid

SOC	Standard Occupational Classification
TB	terabyte
UMLS	Unified Medical Language System
XML	Extensible Markup Language