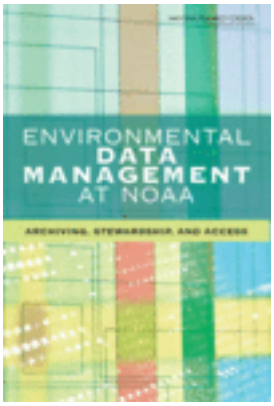

Environmental Data Management at NOAA: Archiving, Stewardship, and Access

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

ISBN: 0-309-11210-9, 130 pages, 6 x 9, (2007)

ENVIRONMENTAL DATA MANAGEMENT AT NOAA

ARCHIVING, STEWARDSHIP, AND ACCESS

Committee on Archiving and Accessing
Environmental and Geospatial Data at NOAA

Board on Atmospheric Sciences and Climate

Division on Earth and Life Studies

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

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

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

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

This study was supported by Contract/Grant No. DG133R04CQ0009 between the National Academy of Sciences and the National Oceanic and Atmospheric Administration. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number-13: 0-309-11209-3
International Standard Book Number-10: 0-309-11209-5

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON ARCHIVING AND ACCESSING ENVIRONMENTAL AND GEOSPATIAL DATA AT NOAA

DAVID A. ROBINSON (*Chair*), Rutgers University, Piscataway, New Jersey
DAVID C. BADER, Lawrence Livermore National Laboratory, Livermore, California
DONALD M. BURGESS, University of Oklahoma, Norman
KENNETH E. EIS, Colorado State University, Fort Collins
SARA J. GRAVES, University of Alabama, Huntsville
ERNEST G. HILDNER III, NOAA Space Environment Center (retired), Boulder, Colorado
KENNETH E. KUNKEL, Illinois State Water Survey, Champaign, Illinois
MARK A. PARSONS, University of Colorado, Boulder
MOHAN K. RAMAMURTHY, University Corporation for Atmospheric Research, Boulder, Colorado
DEBORAH K. SMITH, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts
JOHN R. G. TOWNSHEND, University of Maryland, College Park
PAUL D. TRY, Science and Technology Corporation, Williamsburg, Virginia
STEVEN J. WORLEY, National Center for Atmospheric Research, Boulder, Colorado
XUBIN ZENG, The University of Arizona, Tucson

NRC Staff

IAN KRAUCUNAS, Study Director
ELIZABETH A. GALINIS, Research Associate
ROB GREENWAY, Senior Program Assistant

Preface

The National Research Council (NRC) impaneled this committee in response to a request from the National Environmental Satellite, Data, and Information Service (NESDIS) of the National Oceanic and Atmospheric Administration (NOAA) to provide high-level advice on how to archive and provide access to the broad range of environmental data NOAA and its partners collect. With finite resources and enormous growth in data volumes, NOAA seeks input on how to identify the observations, model output, and derived products that must be preserved in perpetuity and made readily accessible versus those that require shorter storage life spans or more limited access. The committee's full statement of task is included in Appendix A.

NOAA is to be commended for addressing its data archiving and access challenges and for seeking external advice for such critical endeavors. This final report expands on an interim report in which the committee proposed preliminary principles and guidelines for archiving the environmental data that NOAA currently collects (NRC, 2006). These preliminary ideas, which focused mainly on data archiving, have been refined and expanded to encompass issues of data discovery, access, and integration, as well as data stewardship. Our intent is to provide a foundation for NOAA, its partners, and its community of data users to build on as they continue to develop their data archiving and access systems. For example, NOAA's Data Archiving and Access Requirements (DAAR) Working Group (see Appendix D) will use these principles and guidelines

as they provide ongoing advice to NOAA's Science Advisory Board on specific data management activities.

As part of its deliberations in preparing its interim and final reports, the committee met five times over a 12-month period. At several meetings, representatives from across NOAA provided briefings on their data management and archiving activities. A "NOAA Data Users–Data Managers Forum," held in October 2006, brought together individuals from NOAA, other federal agencies, educational institutions, and the private sector to discuss issues of data and product discovery, access, and delivery. This event proved to be exceedingly beneficial to the committee, as were the briefings we received at other meetings regarding NOAA's Data Management Report (NOAA, 2006a) and the plans of the DAAR Working Group. We also received input describing NOAA's legal requirements as they relate to data management and archiving; the relative costs of saving certain types of derived data products versus regenerating these products from archived first-stream data; the challenges associated with interagency cooperation on data management issues; and the value to society of archiving and facilitating the discovery, access, and delivery of a broad variety of environmental data.

The primary result of the committee's effort, and the main focus of this report, is a series of high-level principles that are essential for effective environmental data management. Many of these principles are accompanied by explanatory guidelines intended to help NOAA and its partners apply these principles to their ongoing data management planning. The three main data management functions—data stewardship, data archiving, and data access—are explored in individual chapters. The final chapter describes the essential components and attributes of an integrated, enterprise-wide data management system at NOAA. Background and discussion are provided to help readers understand the context, basis, and implications of each principle and guideline, and case studies are used to illustrate key points where appropriate.

The committee received input from a large number of individuals as it prepared its interim and final reports. The committee would especially like to thank Tom Karl (Director, National Climatic Data Center), Chris Fox (Director, National Geophysical Data Center), Zdenka Willis (former director, National Oceanographic Data Center), and their administrative staffs. Thanks are also offered to the following individuals within NOAA: John Bates, Richard Brooks, Kurt Schnebele, Bonnie Ponwith, Susan McLean, Richard Beeler, Donald Collins, Steve Murawski, Dru Smith, Marc Hodges, Mark Anderes, Maureen Kenny, James O'Sullivan, Rick Visbulis, Glenn Rutledge, Steve DelGreco, Neal Lott, Tim Owen, Fred Branski, Robert Bunge, Tim Birdsong, Tony LaVoi, Roy Mendelsson, Keith Dixon, Jaeson Abraham, and Steve Hankin. Our gratitude to those from

other agencies and organizations who participated in the Data Users–Data Managers Forum, including Mark Russo, Michael Reiter, John Schalles, Kathy Strebe, Lee Branscome, Robert Chadduck, Dave Jones, Tamara Ledley, Patricia Liggett, Martha Maiden, Paul Pisano, and Larry Robinson. The insights of Robert Serafin, then Chair of the NRC's Board on Atmospheric Sciences and Climate (BASC), BASC member Roger Wakimoto, and DAAR Working Group chair Ferris Webster are also greatly appreciated. On behalf of the entire committee, I would also like to express gratitude to Ian Kraucunas, Elizabeth Galinis, and Rob Greenway of the NRC staff for their excellent support throughout our endeavor.

Finally, I would like to thank my fellow committee members for their enthusiastic participation and excellent contributions. This report reflects their dedication to the science community and illustrates their belief that by providing high-level advice on NOAA's data management activities, they can make a difference.

David A. Robinson, *Chair*
Committee on Archiving and Accessing
Environmental and Geospatial Data at NOAA

Acknowledgments

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's (NRC's) Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report:

Peter Backlund, National Center for Atmospheric Research, Boulder, Colorado
Roberta Balstad, Columbia University, Palisades, New York
Suzanne Carbotte, Lamont-Doherty Earth Observatory, Palisades, New York
Peter Cornillon, University of Rhode Island, Narragansett
William J. Emery, University of Colorado, Boulder
Bryan Lawrence, British Atmospheric Data Centre, Chilton, Oxfordshire, UK
Vincent V. Salomonson, University of Utah, Salt Lake City
Donald Sawyer, National Aeronautics and Space Administration, Greenbelt, Maryland
Tamara Shapiro Ledley, TERC, Cambridge, Massachusetts

Soroosh Sorooshian, University of California, Irvine
Jose Zero, Intel Corporation, Cloverdale, California

Although the reviewers listed above have provided constructive comments and suggestions, they were not asked to endorse the report's conclusions or recommendations, nor did they see the final draft of the report before its release. Lee Branscome, Climatological Consulting Corporation, oversaw the review of this report. Appointed by the NRC, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring panel and the NRC.

Contents

Summary

The National Research Council (NRC) impaneled this committee to provide high-level advice on how to archive and provide access to the broad range of environmental data collected by the National Oceanic and Atmospheric Administration (NOAA) and its partners. The data managed by NOAA provide a broad range of benefits to society, but rapid increases in data volumes, as well as demand for these data, have created a first-order data management challenge. NOAA has asked for principles and guidelines to help them identify the observations, model output, and other environmental information that must be preserved in perpetuity and made readily accessible, as opposed to data with more limited storage lifetime and accessibility requirements.

This report offers nine general principles for effective environmental data management, listed below, along with a number of guidelines and examples that describe and illustrate how these principles could and should be applied at NOAA. This guidance is based on the following categories of information reviewed: NOAA's legal and administrative mandates; the data archiving requirements and access capabilities required by NOAA's mission objectives; the findings and recommendations from many previous reports; the documents and feedback received from NOAA, its agency partners, and its data users regarding current and planned data management activities; and the knowledge and experience of the members of this committee.

PRINCIPLES

1. **Environmental data should be archived and made accessible.**
2. **Data-generating activities should include adequate resources to support end-to-end data management.**
3. **Environmental data management activities should recognize user needs.**
4. **Effective interagency and international partnerships are essential.**
5. **Metadata are essential for data management.**
6. **Data and metadata require expert stewardship.**
7. **A formal, ongoing process, with broad community input, is needed to decide what data to archive and what data not to archive.**
8. **An effective data archive should provide for discovery, access, and integration.**
9. **Effective data management requires a formal, ongoing planning process.**

Although the principles are numbered for convenience, all nine should be regarded as equally important for effective environmental data management. The explanations that follow include key guidelines that explain how these principles could and should be applied to improve the effectiveness, reliability, and utility of NOAA's data management activities. The numbers in brackets indicate the chapter(s) of the report where each principle is discussed.

PRINCIPLE #1: Environmental data should be archived and made accessible. [3, 5, 6]

In the view of this committee and many other groups, full and open access to data should be a fundamental tenet at all US federal agencies, including NOAA. The environmental data[1] collected by NOAA and its partners constitute an invaluable resource that should be securely archived and made broadly accessible so that a diverse group of users can conduct the analyses and generate the products necessary to describe, understand, and predict changes in the Earth's environment. **Although it is impossible to save everything, the goal of NOAA's data management enterprise should be to ensure that the broadest possible collection of environmental data is archived and made discoverable and accessible to**

[1]Throughout this report, the term "environmental data" is used broadly to indicate all types of Earth System observations (including physical samples as well as in situ and remotely sensed data), model output, and synthesized products derived from these data.

the widest possible range of users. Doing so will maximize the value of collected data and guarantee that they are available to meet the nation's current and future economic, social, and environmental needs.

PRINCIPLE #2: Data-generating activities should include adequate resources to support end-to-end data management. [3, 7]

At the outset of any activity that will generate environmental data, end-to-end data management should be planned and budgeted. It is also essential to plan and secure adequate funding for the management of environmental data streams that were collected or generated without sufficient resources reserved for long-term data management. Data management planning should begin when plans are first made to acquire a new sensor system or other source of environmental data and continue until the archived data are no longer useful. Plans for data management should explicitly address data storage, preservation, and stewardship responsibilities, and sufficient funds should be reserved to acquire, process, store, maintain, update, and provide access to the data for extended periods of time. These activities all require funding for personnel as well as for hardware. If additional resources can be secured, they should be used to bring additional environmental data into NOAA's archives, to improve the quality and usefulness of the data already in the archive, and to provide more effective access to both new and existing data sets.

PRINCIPLE #3: Environmental data management activities should recognize user needs. [3, 4, 5, 6, 7]

The success of any data management enterprise is judged by its usefulness to current and future users. However, for environmental data, both user needs and the data themselves are constantly evolving. **Thus, all environmental data management activities, including decisions regarding archiving and access of specific data sets as well as the development of the enterprise-wide data management plan, should incorporate substantial and ongoing user input.** Ideally, this input should include both informal feedback from a broad range of users, for instance through user logs, and more formal forms of feedback such as external advisory groups and stakeholder panels. User involvement is critical to make certain that essential data are properly archived and that data access activities are as practical, efficient, and cost-effective as possible. Ongoing stakeholder involvement in the planning process is also essential to ensure adequate integration and coordination across the broad spectrum of NOAA's data management activities.

PRINCIPLE #4: Effective interagency and international partnerships are essential. [3, 5]

Any environmental data management planning process or system needs to include substantial coordination and agreement among the relevant federal agencies and international partners in order to achieve maximum cost-effectiveness and to ensure proper data stewardship. Different agencies have different missions with respect to collecting, archiving, and providing access to different types of data, and these missions sometimes overlap or leave gaps in responsibility for critical data sets; for instance, the transition from research to operations has been a longstanding challenge. **Although no agency can or should be expected to assume complete responsibility for all types of environmental data, NOAA should consider taking a leadership role in archiving and providing access to a broad range of environmental data, including data not traditionally regarded as falling under its operational mission. At the very least, NOAA should improve its relationship with several important partners to guarantee that all essential environmental data are archived and made available to users.** Leadership on data management issues is important if NOAA wants to continue being regarded as the lead federal agency for Earth System research, environmental prediction, and other important data-driven activities such as climate services.

PRINCIPLE #5: Metadata are essential for data management. [3, 4, 5, 6]

To ensure the enhancement of knowledge for scientific and societal benefit, metadata that adequately document and describe archived environmental data should be created and preserved. **Whenever possible, NOAA's metadata protocols should be made compatible with international standards and should be expanded, in coordination with other federal agencies and international entities, to ensure proper preservation and stewardship of data and to facilitate data discovery, access, and integration.** Metadata ideally should include not only descriptive information about the data and how they were originally collected or produced, but also information on the quality and appropriate uses of the data, how they have been managed and processed, and the relationship of the data to data in other collections. Some of this information could be gleaned from user feedback logs, which should be considered for inclusion with the standard metadata associated with each data set.

PRINCIPLE #6: Data and metadata require expert stewardship. [4]

Scientific data stewardship encompasses all activities that preserve and improve the information content, accessibility, and usability of data and their associated metadata. These activities include maintaining a scalable and reliable infrastructure to support long-term access and preservation, preserving data access and archive integrity during media migration and software evolution, providing effective data support services and tools for users, and enhancing data and metadata by adding information that is established throughout the data life cycle. By adding a data management framework that maintains and improves the archive and assures access and understanding for users, the stewardship function supplements and enhances data collection, spans data archiving and access to enable discovery and integration, and facilitates the realization of societal benefits. **Scientific data stewardship, with assigned organizational responsibility, should be applied to all environmental data sets managed by NOAA and to their associated metadata to make sure this information is preserved, remains continually accessible, and can be improved as future discoveries build understanding and knowledge.** In particular, each data set should have an identified steward who understands the data and is responsible for working with data providers and data users to maintain the archive, assess and improve the data and associated metadata, and make certain that data access systems meet user needs.

PRINCIPLE #7: A formal, ongoing process, with broad community input, is needed to decide what data to archive and what data not to archive. [5]

NOAA needs to establish a high-level, enterprise-wide approach to decide what data to include in their archives. It is not possible to save everything, so at some point certain data will need to be designated for reduced archiving and/or access requirements. Original observations, being irreplaceable, are the most important type of data to preserve. Candidates for reduced archiving requirements include data that are obsolete, redundant, or clearly have only short-term uses. Because it is impossible to predict all future applications of a data set, however, it is always preferable to reduce accessibility but preserve the data in some form than to consider disposal when resources for data management are limited. For data that can be reproduced or replicated, such as multiple versions of reprocessed data or model output, it may sometimes be more cost-effective to regenerate the data on demand than to archive all versions of the data. **In all cases, the decision to archive data (or not to archive data)**

should follow established procedures, should be driven by scientific and societal benefits, and should explicitly incorporate broad community engagement and coordination with other agencies.

PRINCIPLE #8: An effective data archive should provide for discovery, access, and integration. [6]

To the maximum extent possible, users should have full, open, and timely access to all of NOAA's environmental data and metadata; data and metadata should be made easily discoverable to a broad range of users; and the data management system should be designed to allow the integrated exploitation of data from multiple sources within and outside of NOAA to answer environmental questions and support NOAA's mission. Discoverability and integration depend critically on metadata to guide search results and on compatible data types and structures, while accessibility is predicated on minimizing administrative, technical, and systematic barriers. In addition, all aspects of data access must take the needs of users into account. Since environmental data are generally managed in a distributed environment and user capabilities and needs vary widely, different access mechanisms are needed to accommodate different data presentations and user requirements. **In general, all aspects of environmental data access would benefit from an enterprise-wide effort to improve and expand existing data access portals, increase the linkages between different portals, and create new portals targeted at particular applications or user groups.** Search tools and other discovery-enhancing features could also be improved at many data access points, and an ongoing process is needed to evaluate user needs, capture user feedback, and use this feedback to build metadata and improve data access tools.

PRINCIPLE #9: Effective data management requires a formal, ongoing planning process. [7]

NOAA faces a formidable data management challenge due to rapidly growing data volumes and data access demands, a diverse spectrum of data types, a heterogeneous population of users and potential users asking increasingly inter- and multidisciplinary environmental questions, and an environment of rapidly evolving technologies and constant budgetary pressures. NOAA also needs to continue to ensure the consistency and continuity of numerous data sets in order to meet its legal, mission, and administrative requirements and to fulfill its interagency and international agreements. However, the scale and complexity of NOAA's data archiving and access requirements have reached the point where the ad hoc, discipline-specific, instrument-oriented data management systems

and decision-making processes relied on in the past are unlikely to be able to keep up with future demands.

To improve the coordination, integration, flexibility, adaptability, and cost-effectiveness of its future data management activities, NOAA should establish and codify an enterprise-wide data management plan that explicitly incorporates all the principles in this report and emphasizes the role of environmental data in supporting NOAA's mission, vision, and goal statements. This plan should formally delineate individual and shared responsibilities for the archiving, stewardship, and access of all environmental data sets that fall under NOAA's purview. Creating a complete, publicly available inventory of NOAA's data would be a logical first step. The data management plan should also include a formal, ongoing planning process—developed and integrated across NOAA with substantial user involvement and coordination with other agencies—to prioritize data management activities and to make sure the system keeps pace with changes in observing systems, models, data storage and access technologies, scientific understanding, user demands, and available resources.

Fortunately, NOAA has the opportunity to build on a number of successful data management activities that are already providing reliable data archiving, stewardship, and access capabilities for many user communities, as well as several prototype projects designed to improve data discovery, utilization, and integration. Recent noteworthy accomplishments include the Global Earth Observation Integrated Data Environment (GEO-IDE) concept, the formation of the Data Access and Archiving Requirements (DAAR) Working Group, and the implementation of the Comprehensive Large-Array [data] Stewardship System (CLASS). All these activities will need to be expanded, integrated under a formal enterprise-wide data management plan, and accompanied by broad ongoing support in order to create a truly integrated and cost-effective system that meets user needs as well as all legal, administrative, and mission requirements.

NOAA would be well advised to continue moving toward a federated but integrated "system-of-systems" to manage its environmental data. Such an approach would leverage existing capabilities and facilitate the inclusion of users and other stakeholders in the planning process, while still providing the top-level support and strategic guidance required to improve integration, coordination, and cost-effectiveness. This approach would also mirror and support the evolution toward a system-of-systems concept in the observational network, namely, the Global Earth Observing System of Systems (GEOSS). Implementation will be a difficult and challenging process, particularly in light of projected data volumes and present resource allocations, but it is critical for NOAA to establish and

maintain a coordinated, enterprise-wide data management plan to ensure that it continues to meet its legal and mission requirements as well as user needs.

———

Understanding and predicting environmental change are likely to remain critically important activities throughout the 21st century. Hence, it is critical to archive and provide access to the environmental data needed to describe, understand, and predict changes in the Earth System and the impacts of these changes on ecosystems, society, and other elements of the system. NOAA is clearly dedicated to providing excellent service for its many user communities, and the agency is to be commended for soliciting external advice to make sure it continues to provide effective stewardship of important national assets. We hope that NOAA and its partners find the guidance offered in this report useful as they continue to improve their data archiving, data stewardship, and data access capabilities.

1

Introduction

NOAA is a mission agency with legal, administrative, and mission mandates that require it to archive and provide access to a wide variety of environmental data, including model output.[1] The data managed by NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) stretch from the surface of the sun to the core of the Earth and provide a broad range of benefits to society. Customers of this investment include other offices within NOAA, other federal agencies, state and local governments, industry and business interests, scientists, educators, and the public, including the international community. The needs of these users are diverse, making it difficult to assess the value of any particular environmental data stream, much less the sum of all data managed by NOAA. However, the substantial resources typically required to collect or generate environmental data imply that they have significant social, economic, and/or scientific value.

As described in Chapter 2, the volume and variety of data collected by NOAA and its partners have increased dramatically over the past decade, as has the demand for access to these data by an increasingly diverse range of users. These trends are expected to continue over the next few decades as new satellite systems, modeling capabilities, and other data-generating activities come on line. It is essential to provide effective

[1]Throughout this report, the term "environmental data" is used broadly to indicate all types of Earth System observations (including physical samples as well as in situ and remotely sensed data), model output, and synthesized products derived from these data.

stewardship for these data sets and to expand and improve access to them in order to fully realize their social, economic, and environmental benefits. Unfortunately, the resources available for data management activities at NOAA have remained relatively flat in recent years. In addition, NOAA has yet to implement a coordinated, enterprise-wide plan that explicitly spells out the data management responsibilities of various program elements, or how these activities will mesh with data management activities at other agencies and international groups to ensure the long-term preservation and enhancement of all important environmental data sets, as well as access to these data by a broad range of users.

NOAA deserves praise for the steps that it has already taken to evaluate and improve its data management activities, such as the report *Integrating the Pieces: Assessment of NOAA's Environmental Data and Information Management* (NOAA, 2006a), the formation of the Data Archiving and Access Requirements (DAAR) Working Group, and the conceptual plan for the Global Earth Observation Integrated Data Environment (GEO-IDE) (NOAA, 2006b), all of which are discussed in Chapter 2. These preliminary steps will need to be connected and extended by an effective, integrated, enterprise-wide data management plan, and accompanied by broad support and sustained funding, in order to meet the formidable challenge posed by anticipated increases in data archive volumes and data access demands and to ensure that NOAA's environmental data continue to be available to meet societal needs. This report provides the foundation for such a plan by describing the principles that underlie effective environmental data management and providing guidelines and examples that explain and illustrate how these principles could and should be implemented within an integrated data management framework.

It should be emphasized from the outset that the sheer volume and diversity of NOAA's environmental data holdings, coupled with the difficulty in predicting the future value of these data, demands that a structured but flexible process be used when making data management decisions. For example, as discussed in later chapters, the decision to upgrade legacy systems to meet the demands of increasing data volume and complexity should be made on a case-by-case basis, with input from the broad user community. NOAA's end-to-end data management process, which includes acquiring, processing, storing, maintaining, updating, and providing access to data, represents an interconnected series of activities involving a large number of different people and systems and serving a broad range of stakeholders. Hence, in addition to providing principles related to the types of scientific data that should be archived and how to provide the best possible access to these data, this report addresses some of the key organizational, management, and planning activities, including inter- and intra-agency relationships, that are critical for making effective

data management decisions. A key theme that emerges from these considerations is that NOAA needs to establish, codify, and maintain a formal, flexible, ongoing process for data management decision making.

This report is organized as follows. Chapter 2 provides an overview of current data management activities at NOAA and highlights some key findings and recommendations from previous reports that discuss archiving and access of environmental data at NOAA and other federal agencies. Chapter 3 presents five overarching principles that underlie effective environmental data management. Chapters 4, 5, and 6 define and describe in greater detail the three fundamental functions of a data management system: data stewardship; the decision-making process for data archiving; and data discovery, access, and integration. Each of these three chapters is organized around a single major principle, supplemented with more detailed guidelines and examples that describe and illustrate how these principles could be applied to current and future environmental data sets at NOAA. The final chapter provides one final principle and outlines how the nine principles and guidelines offered in this report, when considered in the context of NOAA's existing data management activities and accompanied by sufficient ongoing support, could yield a reliable and cost-effective "system-of-systems" for data management that meets or exceeds anticipated user requirements.

2

Background

The first part of this chapter summarizes current data management activities at NOAA as well as the various requirements, mandates, and national and international agreements that are most relevant to these activities. It also describes how rapid increases in data volume and data diversity are straining existing resources. The following section summarizes findings and recommendations from a number of previous reports that address various aspects of environmental data management. This background material forms the basis for this committee's understanding of current data management practices at NOAA and other federal agencies, which leads directly to the principles and guidelines provided in the chapters that follow.

DATA MANAGEMENT AT NOAA

Historically, NOAA's data management activities were driven primarily by the agency's meteorological, oceanographic, and geophysical operational mission requirements. In recent years, the agency has taken on an increasing role in collecting, archiving, and providing stewardship and access for a broad range of environmental and geospatial data, including data collected internationally and by other agencies for a wide variety of purposes. Currently, NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) operates three National Data Centers—the National Climatic Data Center (NCDC), the National Geophysical Data Center (NGDC), and the National Oceanographic Data

Center (NODC)—along with a number of smaller "centers of data." The National Data Centers are large repositories whose primary functions are to archive, disseminate, and provide stewardship for the data that fall under their purview. "Centers of data" provide specialized data or expertise not readily available from the large data centers (Mock, 2001); they range from well-established archive and access points for specific environmental parameters (for example, the National Snow and Ice Data Center) down to small facilities that maintain certain retrospective records for research or operational applications. For example, the Great Lakes Environmental Research Laboratory holds several specialized data sets, such as hydrology and hydraulics data for the Great Lakes, that are used in research activities and made available to external users.[1] Each data center and center of data represents an important component of NOAA's data management enterprise, and collectively they are responsible for "acquiring, integrating, managing, disseminating, and archiving environmental and geospatial data and information obtained from worldwide sources to support NOAA's mission."[2]

As discussed in the paragraphs that follow, some of NOAA's data management activities are codified in the form of legislative mandates, administrative orders, or agreements with other entities, but these documents typically do not spell out specific requirements and responsibilities for individual data sets or derived products, and in many cases they are not accompanied by dedicated funding to accomplish the required activities. Additionally, little formal guidance is available to help data managers decide on the appropriate level of archiving and access to apply to different kinds of data, such as model output or multiple versions of reprocessed satellite data. A significant fraction of NOAA's data are thus collected, archived, and disseminated on an ad hoc basis using limited discretionary funds. The dedication and resourcefulness of NOAA personnel have thus far allowed this approach to work, but the ever-increasing complexity of environmental data coupled with the anticipated explosion in data volumes over the next decade have created a situation where NOAA may not be able to provide the data archiving, stewardship, and access services required to realize the full societal benefits promised by current and future data-generating activities. The background provided in the remainder of this section explains the scope of the data management challenge currently faced by NOAA and describes some of the preliminary steps that the agency has taken to meet this challenge.

[1]http://www.grerl.noaa.gov/data/.
[2]NOAA Administrative Order 212-15.

Requirements, Mandates, and Agreements

Federal legislation contains numerous mandates requiring NOAA to archive data. Some of these laws were originally established for entities that subsequently became parts of NOAA. For example, the Federal Records Act of 1950 established the National Weather Records Center, which later became the NCDC, as the official depository of all U.S. weather records. The importance of understanding climate change has led to additional legislation focusing specifically on climate-relevant data. The National Climate Program Act of 1978 authorized "management and active dissemination of climatological data,"[3] which led to a corresponding increase of the scope of the data managed by NCDC. The full scope of activities performed by NOAA's three National Data Centers is described in detail in the U.S. Code of Federal Regulations.[4]

In the broader context of global change research, the Global Change Research Act of 1990 requires "an early and continuing commitment to the establishment, maintenance, global measurements, establishing worldwide observations . . . and related data and information systems."[5] The Climate Change Science Program (CCSP), which bridges all government agencies involved in climate change research, including NOAA, includes the following guidance in its Strategic Plan (CCSP, 2003): "Preservation of all data needed for long-term global change research is required. For each and every global change data parameter, there should be at least one explicitly designated archive." However, it should be noted that the CCSP Strategic Plan contains no mention of data management. Other pieces of legislation implicitly require archiving through mandates to improve monitoring, analyses, and forecasts for the atmosphere, the ocean, and the land surface (for example, the Ocean and Coastal Observation System Act of 2005 and the Weather Service Organic Act[6]).

In addition, a variety of legislation applies to federal records management at all federal agencies. For instance, the Federal Data Quality Legislation (Act) of 2001 states that "government must assure the quality of the information disseminated."[7] Similarly, all data maintained for legal purposes or in the national interest must be archived using the stringent standards of the National Archive and Records Administration. This requirement applies to data held at all Federal Records Centers, including the three NOAA National Data Centers (but not the smaller centers of

[3]15 U.S.C. CH29 P.L. 95–357; see also Appendix C.
[4]CFR Title 15, Chapter IX, Part 950.
[5]P.L. 101-606(11/16/90) 104 Stat. 3096–3104.
[6]15 U.S.C. 313 et seq.
[7]P.L. 106-554 Section 515.

data). Other relevant laws include the Paperwork Reduction Act of 1995,[8] the Freedom of Information Act,[9] various U.S. National Archives and Records Administration (NARA) Records Management Regulations,[10] and other U.S. laws.[11] NOAA's environmental and geospatial data must also be maintained in accordance with applicable Office of Management and Budget (OMB) regulations,[12] as well as with Federal Geographic Data Committee (FGDC) approved data standards. All these regulations and pieces of legislation influence data management activities at NOAA and need to be kept in mind when considering any changes to the agency's data management enterprise.

There are also internal mandates and directives regarding NOAA's environmental data management activities. NOAA's 2006 Annual Guidance Memorandum articulates the need for "Integrated data assimilation and management: archived, interoperable, accessible, and readily usable observations and data products." NOAA Administrative Order (NAO) 212-15 is the most relevant document for guiding environmental data management across the agency. It contains an assortment of important and encouraging mandates such as "NOAA data management planning will include end-to-end data stewardship"; that NOAA program managers should "ensure that during the initial planning of new programs NOAA Line and Staff Office requirements for new data are identified"; and that "data are considered, and are to be treated as, corporate assets." However, the requirements and mandates provided in this and other documents are often not very specific, leaving data managers with considerable flexibility, but also little guidance, for determining which data sets to archive. For example, NAO 212-15 specifies that centers of data should transfer their data holdings to one of the three National Data Centers "when continued storage at the Center of Data is no longer appropriate," but the order offers no further explanation of what constitutes an appropriate transition.

NOAA has established working relationships with several other federal agencies to address certain data management issues. For example, NGDC archives marine seismic reflection data originally collected by the U.S. Geological Survey (USGS), while NCDC shares data distribution responsibilities for the data generated by the Defense Meteorological Satellite Program (DMSP), which is operated by NOAA on behalf of the Department of Defense. In terms of data retention and archiving, all of

[8]44 U.S.C. 3501 et seq.
[9]5 U.S.C. § 552, as amended in 2002; http://www.usdoj.gov/oip/foiastat.htm.
[10]36 C.F.R. 1220–1238.
[11]For example, 44 U.S.C. 3101–3107.
[12]For example, OMB Circulars A-16 and A-130.

NOAA's National Data Centers are considered Agency Record Centers and are therefore required to follow NARA's stringent archiving standards and to provide disposition schedules for their records. For weather and climate data, NCDC and NARA have signed agreements designating NCDC as the operational archive. As discussed in *Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government: Working Papers* (NRC, 1995a), "NCDC has its special statutory authority because of the great importance of weather and climate records, the great length of time for which those records remain important, [and] the particular expertise with respect to those records found within NOAA." These same considerations also apply to NGDC and NODC.

Other federal agencies have data collection programs and data management responsibilities that overlap to varying extents with activities at NOAA. For example, the Data and Sample Policy at the National Science Foundation's (NSF's) Division of Ocean Sciences requires its principal investigators to submit all collected environmental data to designated National Data Centers, four out of five of which are operated by NOAA. USGS archives and distributes terrestrial global data sets derived from NOAA's Advanced Very High Resolution Radiometer (AVHRR) satellite. One of the most notable environmental data archiving activities of the past decade has been the evolution of NASA's Earth Observing System Data and Information System (EOSDIS), which was developed over many years—and at great expense—to serve the needs of NASA's Earth Observing System. The successes, failures, and "lessons learned" from the development of EOSDIS have been explored in a number of reports (e.g., NRC, 2004a). One of the most serious challenges for EOSDIS, which was so severe that it led to several mission delays (Asrar 1998), was that the information technology and personnel available at the time had difficulty dealing with the volume and complexity of the data; this problem was exacerbated by data processing requirements and heavy user demands. Fortunately, technological advances have made some of these particular challenges less daunting today than they were when EOSDIS was originally being planned and implemented, although other issues remain. For example, EOSDIS was originally conceived as a short-term archive, so the responsibility for long-term management of EOSDIS data remains a subject of major concern and discussion (Hrastar, 2003).

Archiving and providing access to satellite data, which tend to be especially voluminous and complex, has been a particularly challenging area for interagency cooperation. NOAA and NASA signed memorandums of understanding (MOUs) in 1989 and 1992 stating that the two agencies will "exchange near real time data"; that NASA will fund "active short-term archives" and "transfer to NOAA responsibility for active long-term archiving . . . under plans to be developed under this

MOU"; and that NASA's "access to and use of non-real time archive data and products under NOAA's responsibility will be in accordance with legislative mandates."[13] Unfortunately, as discussed in further detail in Chapters 3 and 5, these agreements appear to have fallen short of original expectations, due in part to a lack of agreement regarding the types of data that fall under each agency's mission but also to ongoing difficulties in securing adequate resources for archiving and provide access to large amounts of data for extended periods of time. It should also be noted that data management responsibilities are just one of the issues that have impeded the transition from research to operations, as discussed in the next section.

The United States is also a signatory to a number of international agreements relating to climate and global change that either explicitly require NOAA to archive certain data or call for actions that can be carried out only if reliable data archives are in place. Several World Data Centers operating in the United States under the International Council for Science are housed within NOAA. For example, the World Data Center for Paleoclimatology is operated by NCDC and collocated with NGDC in Boulder, Colorado. World Data Centers operate under rules that require sustained archiving and convenient access to data. Other international agreements, both global and regional, are also predicated on the existence of archived data sets to perform analyses and assessments. For example, the Global Climate Observing System's (GCOS's) Second Adequacy Report and the GCOS Implementation Plan were both requested and accepted by the Conference of the Parties to the United Nations Framework Convention on Climate Change, of which the United States is a signatory, to assess and improve global observations of climate change. The Committee on Earth Observation Satellites (CEOS), of which NOAA is a member, provided a coordinated response by space agencies for mitigation of the inadequacies identified by GCOS; these included several actions with respect to NOAA sensors and missions (CEOS, 2006).

More recently, the United States has joined many other countries in the formation of the Earth Observation Summit and the subsequent Group on Earth Observations (GEO). The United States has played a leading role within the GEO framework in designing the Global Earth Observation System of Systems (GEOSS). Its 10-Year Implementation Plan, adopted in February 2005, includes the task "To exchange, disseminate, and archive shared data, metadata, and products." The Global Earth Observation Integrated Data Environment (GEO-IDE) concept of operations (NOAA, 2006c), discussed in further detail later in this chapter, specifically notes that "NOAA must be able to successfully integrate information from all of

[13]NOAA Solicitation No. EA133E-07-RP-0041, issued March 27, 2007.

its goal areas and exchange data with partners in the national US-Global Earth Observation System (US-GEO) and the international Global Earth Observation System of Systems (GEOSS)." These agreements place additional burdens on NOAA's data centers, but they also offer opportunities to collaborate with other countries in order to increase the realized benefits of archived data.

Expanding Data Volumes

In the NOAA 2007 budget request,[14] it is noted that "Collectively, the three national data centers acquire over one petabyte (10^{15} bytes) of new data annually, provide access to an archive exceeding 3.5 petabytes, and support over 100 million worldwide [individual data requests] per year, providing data transfers to over two million customers." NOAA's data archive and access volumes are growing rapidly, as demonstrated by the NCDC data access statistics shown in Figure 2-1. This explosive growth is expected to continue over the next decade and beyond, even with the recent de-scoping of the National Polar Orbiting Environmental Satellite System (NPOESS) and Geostationary Operational Environmental Satellite (GOES) programs (NASA-NOAA, 2007), as illustrated in Figure 2-2. Improvements in data storage technology, data processing speeds, and other technological advances (e.g., data compression techniques) will help NOAA accommodate the anticipated increase in data volumes, assuming that data management budgets and historical rates of technological change either stay the same or increase. It is less clear, however, if improvements in internet bandwidth will be able to keep pace with increasing data access demands, especially since NOAA's user base is expanding. The challenges associated with providing effective data integration strategies, discovery tools, and support services for such a large and diverse volume of data are also likely to persist. Together, these challenges will make it difficult for NOAA's data centers to continue to meet their mission requirements, much less expand their archival and access capabilities, under current and projected funding levels.

The archive volume growth shown in Figure 2-2 is driven mainly by a few large but relatively homogeneous data streams, such as satellite observations, model output, and radar data. Simply archiving and providing access to these "large-array" data streams poses a significant data management challenge, but this challenge is magnified when considered in the context of NOAA's other data management activities. For example, many of the nation's most pressing environmental problems require

[14]NOAA 2007 budget request "blue book," available at http://www.corporateservices.noaa.gov/~nbo/07bluebook_highlights.html.

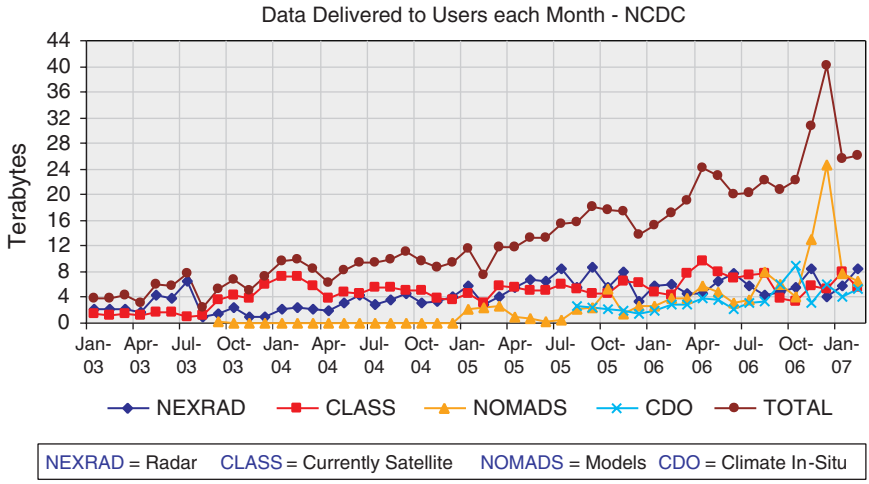

FIGURE 2-1 Monthly data downloads from the National Climatic Data Center (NCDC), in terabytes, for the period January 2003–March 2007: total (brown dots and line) and broken into four categories—NEXRAD, or NEXt-generation RADar data (blue); CLASS, the Comprehensive Large-Array [data] Stewardship System, which currently archives only satellite data (pink); NOMADS, or National Operations Model Archive Distribution System (yellow); and CDO, or *in situ* climate data available through Climate Data Online (light blue). (SOURCE: Thomas Karl, NCDC, personal communication.)

integrated access to multiple data sets, so the large-array data need to be made available to users in formats compatible with corresponding data from medium- and small-volume data sets. Data also need to be made available at varying levels of complexity and resolution to meet the requirements of a user base with a wide range of technical and scientific sophistication.

Data diversity is a further challenge. NOAA's consolidated observation requirements include more than two thousand diverse variables, ranging from hyperspectral satellite imagery to the stomach contents of fish, and these data come from a broad range of platforms, including (but not limited to) satellites, fixed and mobile radars, research aircraft, buoys, ships of opportunity, and mesoscale networks. NOAA also generates and collects a large and rapidly growing volume of reanalysis data and model output, as well as multiple versions of various reprocessed observational data, each of which raises its own data management issues (see Chapter 5). In addition, NOAA produces a wide variety of informational and other

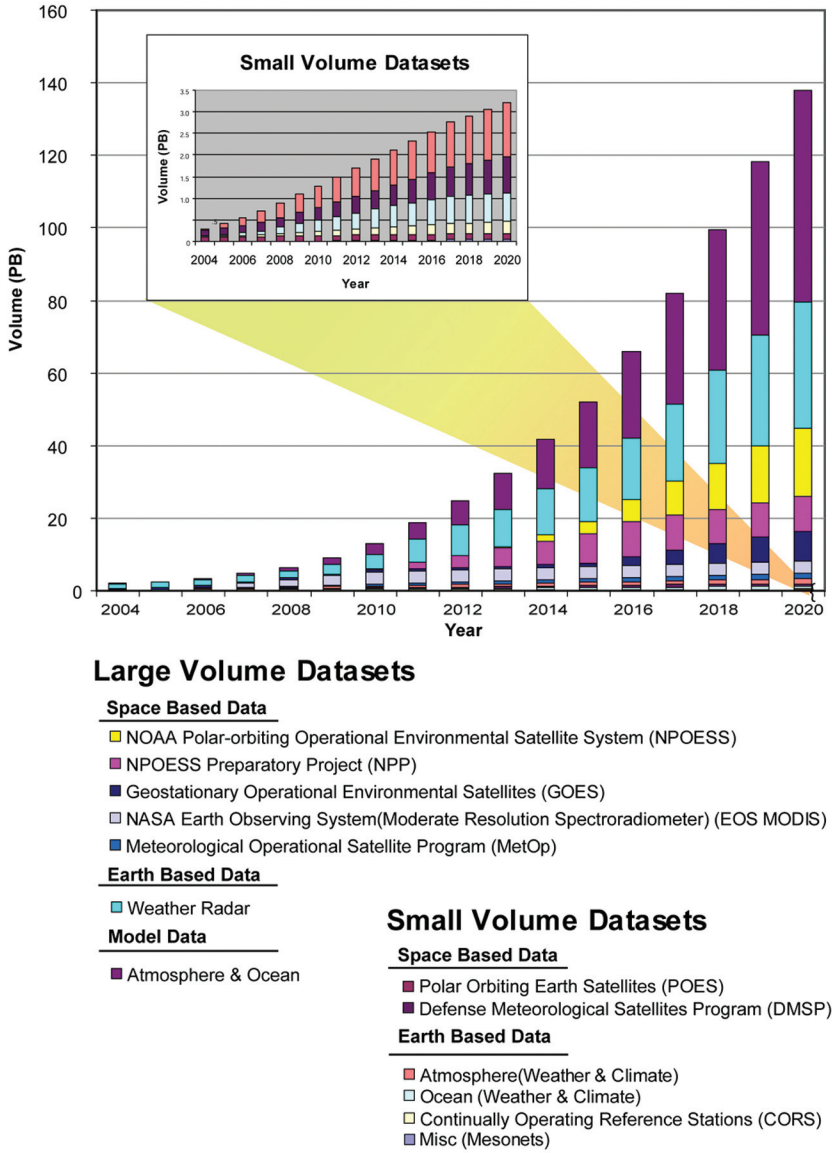

FIGURE 2-2 NOAA NESDIS data archive volume projections, including backup copies, in petabytes. (SOURCE: Updated March 2007 from NOAA, 2003.)

products (see Appendix C), many of which need to be archived and made accessible for a variety of applications.

All the different data streams managed by NOAA are important for answering environmental questions, and each is associated with unique data management challenges. The data centers have evolved in part to address these unique challenges. Each center serves an important and diverse user community and has developed data management strategies designed to best meet the needs of their customers. The data centers also have a long history of providing excellent, discipline-specific user support. For example, NCDC has worked with a consortium of government agencies, under the auspices of the American Society of Civil Engineers (ASCE), to produce a national climatology of ice thickness due to freezing rain. Ice thickness (and therefore weight) is a key engineering design consideration in the construction of many structures that are subject to outdoor weather. The resulting design values of ice thickness, computed on the basis of an extreme value analysis, were included in ASCE Standard 7-05, *Minimum Design Loads for Buildings and Other Structures* (ACSE, 2006).

The value of NOAA's current data archiving and access activities cannot be overemphasized. However, data management activities across NOAA could be better coordinated and integrated to help future users address increasingly inter- and multi-disciplinary environmental problems. For example, the solution of environmental problems in coastal or estuarine areas often requires the integration of diverse terrestrial, hydrological, biological, and physical oceanic observations as well as atmospheric data. Similarly, the response of permafrost areas to global warming will require the integration of atmospheric observations with surface and subsurface cryospheric properties as well as surface hydrological and ecological observations. Chapter 7 provides additional examples and further discussion of how the coordination and integration of NOAA's data management activities could be improved.

Addressing the Data Challenge

The diverse nature and rapidly growing volume of NOAA's data holdings, in an environment of increasing user demands and flat or declining overall agency funding, represents a formidable data management challenge. NOAA clearly recognizes the magnitude of this challenge and has already taken several significant steps to address the ability of its data centers to handle the current and future needs of their users. The NOAA Observing Systems Council (NOSC) serves as the principal advisory body to the NOAA Administrator and is the focal point for the agency's observing system activities and interests. An additional purpose

of NOSC is to coordinate data management activities across NOAA.[15] The NOAA Data Management Committee (DMC) is the NOSC agent for coordinating the development and implementation of data management policy across NOAA. The DMC has broad latitude and authority to coordinate data management activities and make data management decisions for NOAA.[16] One of the primary objectives of this report is to provide the NOSC and DMC with the information they need to integrate sound data management principles with the NOAA Observing Systems Architecture (NOSA) and throughout the entire NOAA data management enterprise.

NOAA has submitted several reports to Congress that assess its data management activities and describe its future plans (NOAA, 2003; NOAA, 2005; NOAA, 2006b). The most recent of these reports, *NOAA's Environmental Data Management: Integrating the Pieces* (NOAA, 2006b), provides a resource needs assessment for each of NOAA's mission goals and individual programs across a broad range of end-to-end data management functions. These reports have raised the level of awareness—both inside and outside of NOAA—of the challenges faced by the agency. While examining these reports reveals that considerable progress has been made over the past several years, the 2006 report concludes that "NOAA [still] faces ongoing challenges in managing increasing diversity and volumes of data, enabling data integration, and addressing real-time dissemination demands."

NOAA has also started to recognize the value of seeking external help when making important data management decisions. In addition to this report, which was requested by NOAA to provide high-level principles and guidelines that it can use as it continues to develop its data management plans, NOAA has formed a Data Archive and Access Requirements (DAAR) Working Group, organized under its Science Advisory Board, to provide ongoing advice on data management activities. The DAAR Working Group is tasked to prioritize the data sets and products that NOAA should archive and to provide specific recommendations for data access, using the principles and guidelines enumerated in this report, along with other materials, to aid their decision process (see Appendix D). NOAA has also attempted to communicate more effectively with its user communities via Web surveys and annual user workshops. The continuation and expansion of these types of activities are essential to ensure that NOAA achieves its vision of "an informed society that uses a comprehensive understanding of the role of the oceans, coasts and atmosphere in the global ecosystem to make the best social and economic decisions."

[15]http://www.nosc.noaa.gov/purpose.html.
[16]http://www.nosc.noaa.gov/dmc/dmc_tor.html.

GEO-IDE and CLASS

NOAA's GEO-IDE is envisioned as a "system-of-systems" that will improve the exchange and integration of data between NOAA and its partners and thus provide easier and more cost-effective access to their diverse environmental data holdings by users (Figure 2-3). This framework parallels the international effort, spearheaded by NOAA, to build and maintain GEOSS, a federated but coordinated system of systems for global observations (Group on Earth Observations, 2005). The goals of GEO-IDE, as outlined in the GEO-IDE Concept of Operations (NOAA, 2006c), are to (1) "take full advantage of the opportunities presented by internet technology to make access to environmental data and information as easy and effective as access to digital documents over the Web is today," and (2) "improve efficiency and reduce costs by bridging the barriers between existing, independent 'stovepipe' systems and integrating the data management activities of all NOAA programs, while avoiding a fully centralized approach." The GEO-IDE concept of operations specifically advocates a federated approach to achieve these goals and

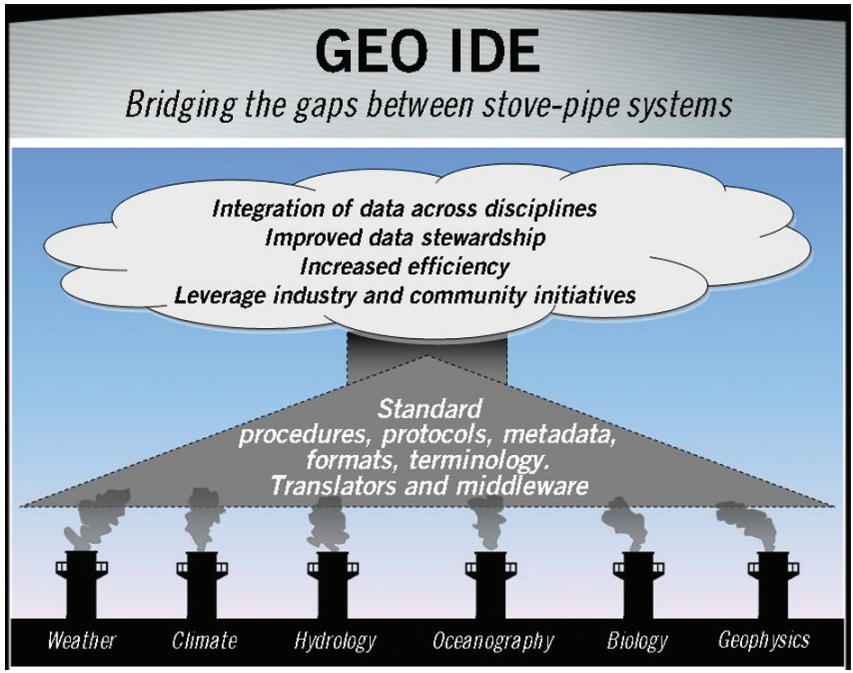

FIGURE 2-3 A vision for GEO-IDE (SOURCE: Lautenbacher, 2006.)

emphasizes the importance of using thoroughly documented and supported standards.

The vision for GEO-IDE articulated in the Concept of Operations document is similarly ambitious. It includes provisions for access ("friendly and flexible mechanisms to locate and access data and data products"); integration ("products in multiple formats and communication protocols"); standards ("utilization of current information technology standards where they are mature, and best practices where accepted standards are still evolving"); user needs ("satisfying the diverse requirements of operations, research, monitoring, and archives"); and user feedback ("a continuous, vigorous outreach process to identify and remedy difficulties encountered by any users"). NOAA has also developed a draft implementation plan for GEO-IDE (NOAA, 2006b) that addresses, among other areas, management and leadership needs, standards (and processes to arrive at standards), forward-looking technology, and a vision for user access.

The vision, goals, and overall concept of GEO-IDE appear to be both well conceived and, for the most part, compatible with the principles and guidelines offered in the following chapters, particularly those that pertain to data discovery, access, and integration. However, GEO-IDE remains very much in the conceptual phase, as can be inferred from the first two steps of the implementation plan, which are to "establish the project management structure" and to "secure funding." It should also be noted that the "hood" in Figure 2-3 represents a significant challenge due to the diversity of the incoming data, the projected increases in future data volumes, and the difficulty in making these data discoverable, accessible, and understandable to a broad range of users. In addition, the GEO-IDE planning documents were developed primarily by a single person, and only one full-time employee is currently allocated for implementation. Thus, GEO-IDE appears to be an important and well-thought-out project that could eventually lead to the effective, integrated, end-to-end data management system demanded by NOAA's mission, but the project is significantly under-resourced given the scope of the effort.

CLASS currently provides storage and online access to NOAA's large-array satellite data and satellite-derived products, but CLASS is planned to evolve into a "unified enterprise data access system that centralizes NOAA's numerous data systems" and provides "long-term, secure storage and access to data, information, and metadata of NOAA's archived assets" (NOAA, 2006d). Hence, CLASS will first need to handle the high-volume data streams depicted in Figure 2-2, but it will have the potential to eventually handle a broader spectrum of NOAA's environmental data. Rather than a static, top-down approach, the current development plans for CLASS (NOAA, 2006d) call for "an open system architecture

whose design will easily accommodate interoperability with existing and new systems as they come on line." CLASS is also expected to "support NOAA's archive and science data stewardship missions by providing the IT [information technology] portion of an archive." Hence, CLASS will provide the hardware, software, and related support services needed to store and provide access to incoming data, while stewardship responsibilities presumably remain the province of experts at NOAA's data centers.

The vision for CLASS has evolved rapidly since the formation of this committee from a holistic, top-down system providing a variety of archive, stewardship, and access functions toward a more limited, federated system that will primarily provide archiving and access services for the high-volume data streams shown in Figure 2-2. The current plans for CLASS (NOAA, 2006a, 2006c, 2006d) appear to be consistent with the goals and vision of GEO-IDE and with many, but not all, of the principles and guidelines provided in this report. For example, it is not clear how data sets will be evaluated for inclusion into CLASS as the system evolves; the principles introduced in the next chapter, and explored in further detail in the chapters that follow, demand that these decisions should be made on a case-by-case basis, with input from users. The extent to which the data handled by CLASS can be effectively integrated with other environmental data streams, including those collected by other federal agencies, also remains to be seen. Finally, the huge data volumes projected to be handled by CLASS will require substantial, ongoing resources to support the hardware and personnel needed to provide reliable archive and access capabilities.

In addition to the challenge of scaling, the collective activities required to move from the successful, but not well-coordinated or integrated, user-support capabilities at the various data centers toward the integrated, mission-focused applications encapsulated by the GEO-IDE vision will require both outside guidance and focused funding. These requirements have proven elusive in the past. Given the rapid evolution of GEO-IDE and the many changes in the scope of CLASS that have taken place, the required integration of GEO-IDE, CLASS, and other data management efforts does not appear to be very well coordinated across NOAA. A related concern is that the rapid evolution of both GEO-IDE and, especially, CLASS might cause the scientific focus of these projects to become lost within the acquisition processes. Improved mechanisms will be needed to ensure that data management activities receive sufficient resources and that they are effectively coordinated across all elements of NOAA's data management enterprise. NOAA will also need to solicit and incorporate input from disciplinary experts both inside and outside of the agency, as well as other potential users, in order to make sure that the

integrated data management system effectively translates environmental data into the societal benefits required to meet NOAA's mission requirements through the GEO-IDE framework.

PREVIOUS WORK

The importance of long-term data collection, data stewardship, the need to archive and provide access to environmental data for societal benefit, the importance of consulting with users when making data management decisions, and the recognition of the significant costs and other challenges involved in these activities are all issues that the scientific and data management communities have discussed for a number of years and for a variety of reasons. Previous reports from the National Research Council (NRC) and other entities have helped set the stage for this report by documenting and analyzing data management practices from various perspectives and by providing findings and recommendations that can be readily applied to the environmental data streams and data management activities at NOAA in particular. For example, *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources* (NRC, 1995b) provided a wide range of advice to NARA and other federal agencies on the long-term retention of scientific and technical data, particularly in electronic formats. This advice included fundamental principles for retaining scientific information, many of which have direct analogs in this report, as well as an extensive discussion of general retention criteria for many different types of data.

A more recent report, *Climate Data Records from Environmental Satellites* (NRC, 2004a), focuses more specifically on generating, analyzing, and archiving the records that are most useful for understanding climate variability and change. That report, while focusing almost exclusively on climate data, offers many findings and recommendations that could be applied more broadly to environmental data from diverse sources with a simple change of wording from "climate data record" to "environmental data management," such as this paragraph from the executive summary.

> *Underlying many of the elements of success is early attention to data stewardship, management, access and dissemination policies, and the actual practices implemented. Because a successful* ***climate data record*** *program will ultimately require reprocessing, data sets used in their creation, such as metadata, should be preserved indefinitely in formats that promote easy access. The ultimate legacy of long-term* ***climate data record*** *programs is the data left to the next generation, and the cost of data management and archiving must be considered as an integral part of every* ***climate data record*** *program.* [Emphasis added]

Two recent reports from the NRC's Committee on Earth Science and Applications from Space have highlighted the critical issues facing the U.S. environmental satellite program. The overall conclusion of that committee's interim report, *Earth Science and Applications from Space: Urgent Needs and Opportunities to Serve the Nation* (NRC, 2006), is that budgetary pressures have severely altered the scope of future environmental satellite missions, particularly NPOESS, to the point where the overall continuity and security of entire classes of space-based environmental observations are at risk of collapse. However, that committee also offered the following recommendation regarding data management at NOAA:

> *The committee recommends that NOAA, working with the Climate Change Science Program and the international Group on Earth Observations, create a climate data and information system to meet the challenge of ensuring the production, distribution, and stewardship of high-accuracy climate records from NPOESS and other relevant observational platforms.*

Awareness about the threatened satellite systems has expanded beyond the agencies and is now the focus of discussions throughout academia and the U.S. Congress. The same committee's final report, *Earth Science and Applications from Space: National Imperatives for the Next Decade and Beyond*, (NRC, 2007), proposes a series of satellite missions designed to preserve the continuity of essential space-based observations and to meet the most essential national needs in a cost-effective manner. The extent to which the recommendations of this "Decadal Survey" will be embraced by the current administration and relevant federal agencies remains unclear. It should be noted, however, that data access improvements that bring current satellite data archives to bear more effectively on problems with demonstrable societal benefits would support the argument to maintain an adequate space-based environmental observing system. The future evolution of U.S. satellite observations will no doubt have a significant impact on future archiving and access requirements at NOAA, NASA, and other federal agencies. However, even if the increases in the volume of incoming satellite data are not as large as originally projected, NOAA still faces an acute data management challenge. In fact, the de-scoping of NPOESS may represent a window of opportunity to make fundamental improvements to existing observation and information systems.

Several reports specific to government data centers, including those at NOAA, have also been published. The report *Government Data Centers: Meeting Increasing Demands* (NRC, 2003a) summarizes a workshop exploring how the increasing volume and number of data sets, coupled with greater demands from more diverse users, are making it difficult for data centers at a number of federal agencies to maintain records of envi-

ronmental change. The report focuses on technological approaches that could enhance the ability of environmental data centers to deal with these challenges and to improve the ability of users to find and use the information held in these centers. The NRC has also provided more focused reviews of NOAA's NGDC (NRC, 2003b), NOAA's NCDC (NRC, 1994), and NASA's Distributed Active Archive Centers (NRC, 1998), which offer a variety of insights on how to best archive and provide access to environmental data. NRC (1998) recommends that "to function optimally, the Data Archive and Access Centers need to be intimately involved with the scientific community they serve. The Data Archive and Access Centers should deliberately pursue and improve routine, daily interactions with active scientists who use their data holdings." NRC (2003b) states that "NGDC should develop an integrated approach to the stewardship of environmental data and operate in such a way that shareable services and functions (for example, database management, software development) serve all NGDC disciplines" and "improve scientific involvement of center personnel with the data sets by recruiting scientists to work with the data, establishing a vigorous program of external visiting scientists, and/or creating strong partnerships with other agencies, industry, and academia to supplement staff expertise."

Because the rapid increase in satellite data volumes is one of the most important forcing factors for NOAA's data management planning, a relevant and timely NRC report is *Utilization of Operational Environmental Satellite Data: Ensuring Readiness for 2010 and Beyond* (2004b), which offers findings and recommendations aimed at defining specific approaches to resolving the potential overload faced by two agencies—NOAA and NASA—responsible for satellite data. The report focuses on the end-to-end utilization of environmental satellite data by characterizing the links from the sources of raw data to the end requirements of various user groups. It is an important foundation document because it addresses three areas that still present challenges: (1) the value of and need for environmental satellite data; (2) satellite data distribution and storage; and (3) satellite data access and utilization. Some of its findings are particularly relevant in the context of this report, including (in brief):

- Improved and continuous access to environmental satellite data is of the highest priority for an increasingly broad and diverse range of users.
- The national and individual user requirements for multiyear climate system data sets from operational satellites are placing special demands on current and future data archiving and utilization systems.
- Data from diverse satellite platforms and for different environmental variables must often be retrieved from different sources, and these

retrievals often yield data sets with different formats, resolutions, and/or grid properties. The multiple steps currently required to retrieve and manipulate environmental satellite data sets are an impediment to their use.

- Early and ongoing cooperation and dialog among users, developers of satellite remote sensing hardware and software, and U.S. and international research and operational satellite data providers are essential for the rapid and successful utilization of environmental satellite data. Many of the satellite data utilization success stories have a common theme: the treatment of research and operations as a continuum, with a relentless team focus on excellence with the freedom to continuously improve and evolve.

Several other NRC reports have focused on the challenges, many of which remain unresolved, associated with the transition from research to operations in satellite data. Among those reports are *From Research to Operations in Weather Satellites and Numerical Weather Prediction: Crossing the Valley of Death* (NRC, 2000) and *Satellite Observations of the Earth's Environment: Accelerating the Transition of Research to Operations* (NRC, 2003c). There are still many benefits to be gained by promoting the effective transition of research satellite data to operational usage, and one way these benefits could be more fully realized is through more effective data access and integration. Satellite data can seldom stand accurately on their own; other *in situ* observations are critical to provide "ground truth" to calibrate and supplement space-based observations. Since NOAA is obliged to be a steward for both types of data, it is therefore well positioned to realize the potential benefits of improved data access and integration for a broad user community.

A number of other reports are of particular importance. *Global Change Science Requirements for Long-Term Archiving* (USGCRP, 1999) provides high-level guidance on long-term archiving of Earth observation data and derived products, including the guiding principles and essential functionality necessary for any long-term data archive to be successful, along with lessons learned from current and past experiences. Many of the principles offered in that report have direct analogs in the principles offered in this report, and the report also inspired some of the more detailed definitions and guidelines offered in the chapters that follow.

The report *Recommendation for Space Data System Standards: Reference Model for an Open Archival Information System* (Consultative Committee for Space Data Systems, 2002) is an internationally developed set of recommendations about how open archival information systems should be structured. Among other tasks, the "OAIS" document defines a model system that

- provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access;
- provides the concepts needed by non-archival organizations to be effective participants in the preservation process;
- provides a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future archives;
- provides a basis for comparing the data models of digital information preserved by archives and for discussing how data models and the underlying information may change over time;
- provides a foundation that may be expanded by other efforts to cover long-term preservation of information that is not in digital form; and
- expands consensus on the elements and processes for long-term digital information preservation and access.

Similarly, the National Science Board, at the request of NSF, has reported on the importance of long-term archives in *Long-lived Digital Data Collections: Enabling Research and Education for the 21st Century* (NSB, 2005). Here there is recognition of the need for formal and established data policies and procedures, a clear technical and financial strategy, and a broad dialogue among agencies that collect and archive data. Data archiving and access are also identified as critical components in the multiagency Data Management and Communications (DMAC) subsystem of the Integrated Ocean Observing System (IOOS) plan.[17] As implementation of IOOS/DMAC takes shape, two things are obvious: data archiving and access capabilities are required components in all end-to-end data management systems; and all agencies that collect data need to ensure these capabilities through a formalized data management plan. Other relevant reports pertaining to the challenges of data archiving include the International Council for Science (ICSU) Report of the CSPR Assessment Panel on Scientific Data and Information (ICSU, 2004) and the Final Report from the Workshop on Research Challenges in Digital Archiving and Long-Term Preservation (NSF, 2003).

Together, these documents and many others—including those available on the Web sites of facilities responsible for data management—provide a solid foundation on which to develop principles and guidelines for archiving and providing access to environmental data at NOAA. In the present report, the concepts that have gained broad acceptance in the data management community have been synthesized down to the most

[17]http://dmac.ocean.us/dacsc/imp_plan.jsp.

essential elements and then applied to NOAA's current data management challenge. The challenges faced by NOAA include not only how to be a wise steward of such a broad range of data managed at its three National Data Centers and centers of data, but also how to meet its growing data archive and data access demands with limited resources. The principles and guidelines offered in the chapters that follow are designed to be applicable across the broadest possible range of circumstances, yet with attention to the practical considerations that should always be kept in mind when managing such an important national asset.

3

Overarching Principles

The critical functions of an integrated, end-to-end data management system can be broken into three main elements: *archiving*, which includes ingestion, storage, and preservation; *access*, which includes discovery, delivery, and integration; and *stewardship*, which spans archiving and access and includes all the processes and procedures that preserve and improve the information content of individual data sets and their associated metadata, as well as assuring access and understanding for users. These three functional elements are interdependent, and they need to be coordinated through a formal, ongoing planning process to ensure that the data management system is integrated, cost-effective, and successful in meeting user needs and requirements. In addition, there are certain overarching principles, laid out in this chapter, that apply to all aspects of data management. Many of the concepts introduced by these overarching principles, such as the importance of user engagement, are also discussed in subsequent chapters that explore the three main functional elements of data management (stewardship, archiving, and access) and, in the final chapter, the need for a formal ongoing planning process to integrate these elements into an effective end-to-end data management system.

The nine principles offered in this report, including the five overarching principles introduced in this chapter, are numbered sequentially for convenience, but all nine should be regarded as *equally important* for effective environmental data management. Some principles are accompanied by guidelines intended to provide specific recommendations for how NOAA and its partners could and should apply these principles to

their current and future data management activities. These principles and guidelines are based on a variety of sources: the large body of previous work described in Chapter 2; NOAA's legal and administrative mandates; the data archiving and access requirements demanded by NOAA's mission objectives; information and feedback received from NOAA, its agency partners, and its data users regarding current and planned data management activities; and the knowledge and collective experience of the members of this committee.

PRINCIPLE #1: Environmental data should be archived and made accessible.

The environmental data (including model output, derived products, and other information) collected by NOAA and its partners are an invaluable resource that should be archived and made accessible in a form that allows a diverse group of users to conduct analyses and generate products necessary to describe, understand, and predict changes in the Earth's environment. Full and open access to data should be a fundamental tenet of all US federal agencies, including NOAA.

NOAA's mission is "to understand and predict changes in the Earth's environment and conserve and manage coastal and marine resources to meet our nation's economic, social and environmental needs." Since the Earth and its environment represent a complex, interconnected biogeochemical system, describing the current state of the system or its variability over time requires a large number of observations, and predicting its future behavior often demands the use of models built around a detailed understanding of various system components. Thus, any observational data stream, model output array, or other environmental data set that contributes to the description, understanding, or prediction of the Earth system (including derived products) should, in principle, be archived by NOAA. Likewise, in order to realize the full benefits of Earth system measurements, analyses, and predictions, these data should be made accessible to the broadest possible range of users in a form that allows them to make informed economic, social, and environmental decisions.

Although "save all environmental data and disseminate it to all possible users" is a worthwhile goal for a data archival and access system, a number of practical considerations make this goal impossible to fully achieve. First among these is the reality of limited resources, which places restrictions on the volume and quality of data that can be archived, the ability to make these data accessible to a wide range of users, and the number of personnel dedicated to ensuring reliable data stewardship. Normally, cost-benefit analysis would be a powerful tool for identifying the most appropriate way to allocate limited resources and prioritize data

sets. However, it is difficult to assess the current value of any particular environmental data stream (see, for example, Millard et al., 1998), and impossible to anticipate all its potential future uses. In addition, it is usually impossible to resample environmental data, although there are exceptions (for example, some types of geodetic data or model output). Finally, we agree with the many previous groups (e.g., NRC, 1997) who have argued that full and open access to data should be a fundamental tenet at all federal agencies. These considerations support the notion that NOAA should strive to archive the broadest possible collection of environmental data and disseminate these data to the broadest possible range of users, and they also support the following principle.

PRINCIPLE #2: Data-generating activities should include adequate resources to support end-to-end data management.

End-to-end data management (which includes acquiring, processing, storing, maintaining, updating, and providing access to data) should be planned and budgeted at the outset of any activity that will generate environmental data. This planning should explicitly address data archiving, data stewardship, and data access responsibilities, and sufficient funds should be provided to archive and provide ready and easy access to the resulting data for extended periods of time.

In general, the cost of archiving and providing access to environmental data represents only a small fraction of the total resources invested in originally collecting or generating the data. For example, NOAA's FY2007 budget request includes $965 million for satellite observation systems alone (acquisitions plus operations), compared to $69 million for all of NOAA's data management activities combined (although some of the satellite observing system funding is spread out over several years to support data acquisition, processing, analysis, and distribution).[1] For digital data, hardware for data storage and for data distribution constitutes a sizable fraction of data management costs, but funding to support data management personnel is also critically important. The return on these investments in terms of societal benefits is difficult to quantify, although the value of environmental data can sometimes be inferred; for example, the economic value of weather forecasts for the household sector alone has been estimated to exceed $109 per household (or $11.4 billion total) per year, compared to an average expenditure of $13 per household per year to support the National Weather Service (Lazo and Chestnut, 2002).

Effective data management planning is needed to ensure that the nation receives an appropriate return on the substantial investments made

[1]http://www.corporateservices.noaa.gov/~nbo/07bluebook_highlights.html.

in observing systems, environmental models, and other data-generating activities. Most agencies, including NOAA,[2] have begun to recognize that data management is far easier to plan and more cost-effective to implement at the outset of data-generating activities. Unfortunately, many *in situ* data sets, as well as high-volume radar and satellite data, have been collected in support of NOAA's operational missions with little initial provision made for long-term preservation and access. However, data once perceived to be of little use can sometimes become quite valuable as a result of advancing technology or changing user needs. For example, many atmospheric data sets that previously had little use beyond satisfying operational needs can now be ingested by models using new data assimilation methods to yield improved atmospheric reanalyses (see, for example, Kalnay et al., 1996). **It is thus essential to secure adequate funding to manage the environmental data streams that were collected or generated without sufficient resources reserved for long-term data management.** A number of previous reports, including many of those summarized in Chapter 2, have highlighted additional benefits that might be realized when sufficient resources are available to support effective and proactive environmental data management.

Despite these demonstrable societal benefits, ensuring adequate and sustained levels of funding to support data management activities remains a major ongoing challenge for both NOAA and other agencies and international groups involved in archiving and providing access to environmental observations, model output, and other environmental information. This challenge has arisen largely for three reasons: in part because data management activities require continuing costs that extend long after the data are originally collected or generated, in part because data management often requires considerable coordination among different agencies or groups, and in part because the benefits of effective data management are difficult to quantify. Considerable progress has been made in some areas, such as the recent expansion of the Comprehensive Large-Array [data] Stewardship System (CLASS), but many other potential environmental data users and applications would benefit from improved archiving and access capabilities at NOAA, especially improvements in data discoverability and integration across NOAA's diverse data streams.

Any additional resources that can be secured to support long-term data management at NOAA should be incorporated into a comprehensive, integrated, and flexible end-to-end data management plan that is coordinated across NOAA to support Earth system research, improved environmental predictions, and other societal benefits. Additional

[2]See, for example, NOAA Administrative Order 212-15.

resources can be expected if the data management system is shown to be cost-effective and to have significant, demonstrable societal benefits. Chapter 7 describes some of the potential applications and societal benefits that could be realized with additional investments in data archiving and access capabilities.

PRINCIPLE #3: Environmental data management activities should recognize user needs.

The success of any data management enterprise is judged by its usefulness to current and future users. However, for environmental data, both user needs and the data themselves are constantly evolving. Thus, all environmental data management activities, including data archiving and access decisions for specific data sets as well as the development of the overall data management system, should incorporate substantial and ongoing user input.

The ultimate measure of any data management enterprise is its benefit to society, and the most straightforward way of both ensuring and assessing these benefits is via close and continuous interaction with users. User participation in the planning and in the ongoing management process helps to ensure that the system meets societal needs, and user feedback provides a mechanism to evaluate individual system components and to identify and develop potential system improvements. Since user needs are continually evolving and new applications for data are constantly being developed, user participation is most effective when it is regular and scheduled, rather than episodic or ad hoc, and when it is solicited from the user community most familiar with the environmental data under review. Ideally, user input should include both informal feedback from a broad range of data users, for instance through user logs, and more formal forms of feedback such as external advisory groups and stakeholder panels. Regular communication with users likewise promotes a proactive approach toward data access improvements.

To the maximum extent possible, current efforts to engage users in specific data archiving and access decisions, as well as in the planning and implementation process for an enterprise-wide data management system at NOAA should be continued and expanded. The recent formation of the Data Access and Archiving Requirements (DAAR) Working Group represents an important step toward expanded user involvement; it will be critical for the DAAR Working Group to remain actively involved in the planning and implementation process for both CLASS and the Global Earth Observation Integrated Data Environment (GEO-IDE), and for that working group, or a similar entity, to be actively engaged in guiding the future evolution of NOAA's enterprise-wide data management activities.

Because the DAAR Working Group is only a single, ad hoc group with limited time and resources, they will not be able to offer specific advice on the large and diverse range of data and products that NOAA collects and generates. It will therefore also be essential to establish a mechanism for ongoing discipline-specific user feedback to help inform individual data archiving and access decisions at NOAA's National Data Centers and other centers of data. Fortunately, these centers already have considerable experience working with their user communities, especially when developing and improving data access capabilities. This experience should be leveraged to facilitate further user engagement and participation in individual data management decisions.

The importance of user feedback and involvement for data stewardship, data archiving decisions, and designing effective data discovery, access, and integration tools are discussed in further detail in Chapters 4, 5, and 6, respectively.

PRINCIPLE #4: Effective interagency and international partnerships are essential.

Any environmental data management planning process or system needs to include substantial coordination and agreement among the relevant federal agencies and international partners in order to achieve maximum cost-effectiveness and to ensure proper data stewardship. NOAA should take steps to improve its relationship with several important partners and consider taking a leadership role in archiving and providing access to a broad range of environmental data, including data not traditionally regarded as falling under its operational mission.

Many different entities collect environmental data, and scientific understanding and other societal benefits cross over political and agency boundaries. Thus, NOAA needs to work closely with other agencies and international groups to make certain that all important environmental data are archived and made available to users. In particular, there should be a clear understanding of both individual and shared responsibilities among the various agencies (including NOAA, NASA, the U.S. Geological Survey, the Department of the Interior, the Department of Defense, and others) involved in archiving and providing access to environmental data collected both domestically and internationally. Even though NOAA is probably the largest environmental data steward in the world, as measured collectively by data diversity and volume, it can and should engage other similar organizations to discuss common problems, share lessons learned, and find areas where collaborations result in mutual benefits (for example, providing mutual backup services).

Interagency coordination is fraught with practical difficulties. Different agencies have different missions with respect to collecting, archiving, and providing access to different types of data, and these missions sometimes overlap or leave gaps in responsibility for critical data sets. The transition from research to operations is one notable and longstanding challenge (see, e.g., NRC, 2000). The concept that a single agency should archive *all* environmental data for the United States, regardless of the source or primary users of a data set, does not appear to be a fiscally, bureaucratically, or politically feasible solution, although the concept of a single repository for all scientific data generated by U.S. agencies is apparently still being discussed (Butler, 2007). Our view is that NOAA should consider taking a leadership role in archiving and providing access to a broad range of environmental data, including data not traditionally regarded as falling under its operational mission. Leadership on data management issues is important if NOAA wants to continue to be regarded as the lead U.S. agency for Earth System research, environmental prediction, and other important data-driven activities, such as climate services; it is recognized, however, that additional resources would be required in order for NOAA to assume such a role.

At the very least, NOAA should work to improve its relationship with its agency partners to ensure that all environmental data derived from publicly funded research are archived and made available to users. Inventories and formal agreements can be time consuming and difficult to complete, but they form the foundation for all future data management decisions. It is also essential to periodically review these data inventories and agreements and to establish extensive and ongoing coordination efforts in order to make sure that new data streams are properly archived and that the evolving set of user needs continues to be satisfied. Facilitating international cooperation is associated with additional complications, as noted in Chapter 2, but also brings the potential for more dynamic collaboration, as illustrated by the effort to bring about the Global Earth Observing System of Systems (GEOSS). Of course, funding is always an important consideration in any international or interagency agreement process, a consideration that provides further support for the principle that "data-generating activities should include adequate resources to support end-to-end data management." Box 3-1 illustrates some of the problems that can arise when interagency coordination is lacking, and additional discussion of interagency and international issues is included in several of the chapters that follow.

BOX 3-1
Archiving Responsibilities for EOS Data

A recent example illustrates the perils of inadequate coordination of data management activities. At one of the meetings of this committee, it was revealed that NASA officials were unaware of NOAA's decision to archive only a small subset of Earth Observing System (EOS) data (in particular, MODIS "Level 1b" data). NOAA representatives explained that NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) will archive only those data that fall under its climate monitoring or operational missions, citing both funding constraints and memorandums of understanding (MOUs) signed by NASA and NOAA in 1989 and 1992 (see Chapter 2). NASA has subsequently decided to archive the rest of the EOS data on an ad hoc basis, and there appears to be little discussion between the two agencies to guarantee that all EOS data are archived and made accessible for the remaining portion of their life cycle. This situation raises serious concerns about the stewardship and long-term preservation of data that are critical for global change research but which NOAA does not consider relevant to its mission. While NOAA's decision to archive and provide access only to data directly relevant to its operational and climate monitoring activities may be based on very real funding constraints, federal agencies should not be forced to rely on ambiguous, nonbinding 15-year-old agreements to make critical data management decisions. Of particular concern is the absence of any formal mechanism for NOAA and NASA to periodically review their agreements in order to ensure that important environmental data are continually archived and made available to users. These issues are explored in further detail in Chapter 5.

PRINCIPLE #5: Metadata are essential for data management.

Metadata that adequately document and describe each archived data set should be created and preserved to ensure the enhancement of knowledge for scientific and societal benefit. NOAA and its partners should continue and expand their usage of metadata standards.

Metadata are all the pieces of information necessary for data to be independently understood by users, to ensure proper stewardship of the data, and to allow for future discovery. This information should include, at a minimum, a thorough description of each data set, such as its spatial and temporal resolution and how it was originally collected or produced, as well as thorough documentation of how it has been managed and processed. Ideally, metadata should also describe appropriate applications of the data, the relationship between the data and other data within and outside of the archive, and enough high-level information to allow even inexperienced users to find, understand, and use the data. Metadata

are thus the essential data management component that makes an environmental data archive useful. Whenever possible, NOAA's metadata protocols should be made compatible with international standards to facilitate discovery and coordination with other archives, and NOAA should work with its users and partners to expand and improve these standards. The contents and utility of metadata for data stewardship, archiving, and access are described in additional detail in Chapters 4, 5, and 6, respectively.

4

Data Stewardship

PRINCIPLE #6: Data and metadata require expert stewardship.

Scientific data stewardship, with assigned organizational responsibility, should be applied to all environmental data sets and their associated metadata to ensure that this information is preserved, remains continually accessible, and can be improved as future discoveries build understanding and knowledge.

The importance of data stewardship has been examined at length in many previous reports, including many of those discussed in Chapter 2. NOAA has also recognized the value of providing effective and ongoing stewardship for its environmental data, as evidenced by the explicit inclusion of the word in the name Comprehensive Large-Array [data] Stewardship System (CLASS), in the Global Earth Observation Integrated Data Environment (GEO-IDE) planning documents, in the terms of reference for the Data Access and Archive Requirements (DAAR) Working Group, and in the statement of task for this requested report. While there is general agreement that sustained data preservation and effective data access require careful stewardship, the term tends to be used somewhat differently in different contexts. We propose the following definition, which is adhered to throughout this report:

> **Data stewardship** encompasses all activities that preserve and improve the information content, accessibility, and usability of data and metadata. These activities include maintaining a scal-

able and reliable infrastructure to support long-term access and preservation, preserving data access and archive integrity during media migration and software evolution, providing effective data support services and tools for users, and enhancing data and metadata by adding information that is established throughout the data life cycle.

Environmental data archives are not static collections, so data stewardship is an ongoing, iterative process involving users, servers, storage systems, software, interfaces, and analysis tools. Since analysis and usage can improve the accuracy and understanding of data and metadata, stewardship activities also include capturing these advances and carrying them forward for future generations. By adding a data management framework that maintains and improves the archive and ensures access and understanding for users, the stewardship function supplements and enhances data collection, spans data archiving and access, enables discovery and integration, and facilitates the realization of societal benefits. Together, these activities support the vision introduced in *Global Change Science Requirements for Long-Term Archiving* (USGCRP, 1999) of "a continuing program for the preservation and responsive supply of reliable and comprehensive data products and information on the Earth system for use in building new knowledge to guide public policy and business decisions."

Stewardship should be a formal and emphasized component of NOAA's data management planning process. In particular, each data set should have a designated data steward; in most cases, teams of skilled professionals with complementary skills are required to enable the level of stewardship needed to support NOAA's mission. These teams include information technology, computer, database, and software specialists; disciplinary science experts; curators; project managers; and a variety of other professionals to support the full range of NOAA's data management activities. Multidisciplinary teams are critical to stewardship, and it has been recommended that data technology specialists receive recognition for this work as "data scientists" (NSB, 2005). Since people retire and organizations evolve, the stewardship plan for each data set also needs to provide for the continuity of stewardship throughout the data life cycle, which can last many decades.

Even though our definition of data stewardship spans a broad range of data management activities and responsibilities, the remainder of this chapter focuses on three fundamental stewardship functions: data and metadata preservation, scientific assessment and improvement, and data support services and tools. In combination, these fundamental stewardship functions ensure that the archive is maintained and that the flow of

information to users supports the realization of societal benefits. Thus, data stewardship provides the foundation for the additional, cross-cutting data management functions of data archiving and data discovery, access, and integration, topics that are touched on here and discussed in detail in Chapters 5 and 6, respectively.

DATA AND METADATA PRESERVATION

Guideline: Metadata that adequately document and describe each archived data set should be created and preserved to ensure the enhancement of knowledge for scientific and societal benefit.

Metadata are all the pieces of information necessary for data to be independently understood by users, to ensure proper stewardship of the data, and to allow for future discovery. This information should include, at a minimum: a thorough description of each data set, including its spatial and temporal resolution; the time and location of each measurement, and how the data were originally collected or produced; and a thorough documentation of how the data have been managed and processed, including information about any media and format migrations, the accessibility of the data, and the algorithms or procedures used for any reprocessing, revisions, or error corrections. Collectively, these pieces of information are what make the data in an archive useful. Ideally, metadata should also describe appropriate applications of the data, the relationship between the data and other data both within and outside of the archive, and enough high-level information to allow different types of users to find and understand the data. Adding these additional pieces of information would help support the discovery and integration of data across different archives and disciplines.

In addition to supporting user needs, maintaining accurate and extensive metadata can improve the cost-effectiveness of an archive by providing for more efficient and effective data storage and access. For example, data accessed less frequently or only by a subset of users may be more cost-effectively archived and made available in a different manner than data that are more widely used. Effective metadata management could thus be expected to help NOAA meet the challenge of increasing data volume and diversity. NOAA's data management system should also be designed to facilitate the sharing of metadata across systems, disciplines, and programs to build more complete data catalogs. This would improve the ability of disparate users to find and use a wide variety of data. Further, because data systems will evolve to incorporate new information and to take advantage of technological improvements, the data system

philosophy needs to account for the reality that metadata continually evolve, expand, and mature. As discussed in the next guideline, these considerations necessitate the use of existing standards and the adoption of new ones, when appropriate, to facilitate services and integration across data sources and disciplines.

Guideline: The application and expansion of metadata and related standards are essential for good stewardship; NOAA and its partners should continue and expand their usage of standards and reference models.

Where possible and practical, it is advisable to use established metadata standards in order to improve cost-effectiveness and coordination with other activities, to guarantee the integrity and security of the archive, and to facilitate the crosscutting stewardship functions of discovery, access, and integration. The PREservation Metadata: Implementation Strategies (PREMIS) Working Group recently developed standards for preservation metadata in accordance with the Open Archival Information System (OAIS) Reference Model (PREMIS Working Group, 2005). The OAIS Reference Model provides detailed guidance on data preservation and should serve as a primary reference for archiving activities at NOAA and other agencies. It is encouraging that the NOAA National Data Centers and CLASS have already adopted the OAIS model to guide their archive development. We recommend that NOAA and its partners make the OAIS model even more of a standard across their data management enterprises and routinely assess their practices against community-developed reference models.

NOAA and other agencies should also be encouraged to expand their metadata protocols beyond mandatory requirements. For example, as discussed previously, standard information about each data set could be expanded to include its relationship to other data sets both within and outside of the archive. In the next section, it is also suggested that user feedback logs be included as part of the metadata for each data set to facilitate data and metadata improvement. In general, any information that could be useful for improving the integrity of the archive, data access capabilities, or data quality should be collected, retained, and utilized. Metadata are also essential for data discovery, access, and integration and are thus critical for facilitating the translation of environmental data to societal benefits; Chapter 6 discusses these benefits of metadata in further detail.

Guideline: NOAA should develop and maintain a scalable and reliable infrastructure that ensures long-term access and preservation of digital data assets.

Data are often subjected to substantial "environmental" changes over time, such as evolutions in storage and access technologies, changes in stewardship responsibility, relocation and repurposing, and—possibly—disruptions due to system failures, operational errors, and natural hazards and disasters. The digital preservation infrastructure for a long-term environmental data archive should therefore be able to handle the following requirements:

- Each preserved digital object should contain sufficient information to enable the application of long-term preservation policies and to handle its life cycle management (that is, an Archive Information Package, as described in the OAIS Reference Model).
- Efficient management of technological innovations and evolutions in both hardware and software, especially when technology begins to become obsolete.
- Efficient risk management and disaster recovery mechanisms from technology failures or degradation; natural disasters such as fires, floods, and earthquakes; or human-induced operational errors.
- Efficient mechanisms to guarantee the authenticity of content, context, and structure of archived information throughout the preservation period.
- The ability to provide for secure discovery and access, with an automatic enforcement of authorization and security policies, throughout the life cycle of each object.
- Scalability in terms of data ingestion rates, storage capacity, processing power, and the speed at which users can discover and retrieve information.

Guideline: NOAA should establish and maintain data and metadata migration plans for all current and future long-term archive systems to adapt to information technology evolution.

All data, including digital information, must be periodically migrated to new storage media to make sure that they remain retrievable for long periods of time. When the data archive is large and growing and data transfer rates are near the upper bounds in storage systems, media migration can be a continual process; that is, the time needed to transfer data from one media to the next approaches the reliable life expectancy of the legacy media itself. Data stewards need to make certain that no informa-

tion is lost, that data access remains uncompromised, and that errors may be corrected whenever the data are migrated to new media, translated to another form, or otherwise modified from their original format during the normal course of technology and software evolution. Irreplaceable data should carry so-called fixity information (for instance, digital signatures), which can be tested during any sort of media migration or backup process to verify that the content has not been altered (CCSDS, 2002).

Because it changes the actual digital structure of the data, format migration can be more problematic than media migration. The best archive formats are those where the digital content of each data record can be described in elementary terms (for example, number of bytes, numeric type, character string, pixel, etc.). This is one feature of an open format standard that helps minimize software and computer operating system dependencies that could render the data inaccessible in the worst case. So-called proprietary formatted data (non-open format description) should in general not be considered as a good candidate for long-term archiving unless a plan and a process are in place to translate the data to an open format standard. Metadata should be stored in similarly open formats and should be tightly coupled with and managed in conjunction with the data so both are always readily available to the user.

The top priority of the data steward should always be the security and integrity of the archive. Thus, the archiving procedures of NOAA and its partners should include storage in multiple independent locations, regular tests on fixity information, and disaster recovery exercises that certify system reliability. Systems that track changes made to digital objects, often used in software development, are also applicable across data and metadata management, and they can be used to guard against irrecoverable errors. NOAA should explore the possibility of expanding its use of change control software, although the long-term viability of the software should also be considered. Finally, data stewards should maintain open lines of communication to their users and their colleagues at other data repositories to ensure the consistency and coordination of archiving procedures across NOAA's data management enterprise.

SCIENTIFIC ASSESSMENT AND IMPROVEMENT

Guideline: Good stewardship requires systematic, ongoing assessment and improvement of data and metadata.

Properly stewarded data collections almost always improve over time. Throughout the data life cycle, which is illustrated in Figure 4-1, archived information may be enhanced by recalibration, error correction,

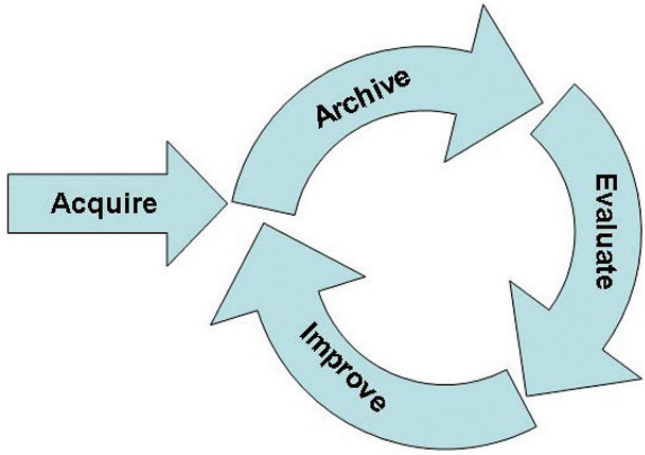

FIGURE 4-1 The data management life cycle. Data and associated metadata are acquired and integrated into the archive and access system. Ideally, data stewards, in collaboration with users, evaluate the data over time, leading to additional knowledge about the data, including potential improvements that might be made to the data or metadata. If improvements are possible, they should be applied, at which point the archiving, evaluation, and improvement cycle begins again.

and/or the addition of supplemental data or metadata derived from scientific research, user feedback, and other knowledge-building processes. For example, calibration and validation studies typically lead to successively higher-quality versions of a given data set via improvements in processing algorithms or error-handling procedures. Of course, these repeated improvements tend to exacerbate the challenge of increasing data archive volumes, particularly with multiple reprocessed data sets. Chapter 5 contains guidelines that can help data managers deal with this issue, noting, for instance, that demand will typically be far greater for the most recent version of each data set and so it may not be necessary to provide immediate access to older versions. It is vital that data stewards be engaged in these decisions.

As with most aspects of data management discussed in this report, the assessment and improvement function of data stewardship is most effective when it occurs on a regular basis and under a flexible but systematic set of rules and requirements. These rules should be advertised both to users and to data providers, who should in turn be given a chance to provide input to the process. Similarly, these rules should explicitly take into account the estimated costs and likely benefits of improvement efforts.

BOX 4-1
Understanding Sea Ice

Sea ice data archived and managed by the National Snow and Ice Data Center (NSIDC) provide an example of how data stewardship can lead to increased value and better understanding of environmental data. The primary sea ice products at NSIDC are derived from passive microwave remote sensing instruments on satellites. These sensors were originally used to support operational weather forecasting, but as a result of an ad hoc process of scientific assessment, researchers now use these data for many other applications, including sea ice research and monitoring. In the late 1980s, NSIDC established a program for the stewardship of passive microwave data and other related sea ice products. Here we provide examples of how the three major aspects of stewardship were applied to these data.

To ensure robust *preservation* of sea ice data, NSIDC established routine off-site backup procedures, as well as an ongoing media migration program, to guarantee data integrity. The center has developed and maintained detailed metadata that document the data and how it has been archived. More recently, data stewards at NSIDC have been exploring a preservation metadata scheme to improve compliance with the OAIS reference model.

The *scientific assessment and improvement* aspect of NSIDC's stewardship program has been even more dynamic. The instruments and satellites used to collect passive microwave data constantly evolve, making it difficult to construct consistent time series or climate data records, so NSIDC documentation includes detailed information on the known errors and uncertainties in the data. For example, in 1999 NSIDC was notified by a user of errors in the latitude, longitude, and pixel area files supplied with some of the passive microwave data. These errors would have biased analyses of the data if left uncorrected, but NSIDC was able to

Scientific data stewards should be encouraged to engage both expert and nonexpert users to understand how data are being used and should make sure that the metadata for each data set describe the appropriate uses and limitations of each data set, in addition to formal uncertainties and errors (see Box 4-1). It is also important to realize and communicate (using metadata) that despite continued improvements, some data may never be appropriate for certain applications.

Guideline: Assessment and improvement should be based on expert knowledge and user feedback.

The assessment and improvement of environmental data sets, like

correct the error and notify users of the changes. Other scientific assessment and improvement activities have required more sophisticated scientific analyses; for example, NSIDC and other scientists have prepared three special reports devoted to deeper analysis of passive microwave-derived sea ice products (Maslanik et al., 1998; Stroeve et al., 1997; and Stroeve and Smith, 2001). All told, these in-house analyses and community interactions have led to the creation of time series that are consistent and well characterized across sensors and platforms.

Those involved in data support services at NSIDC have also been quite active with sea ice data sets. NSIDC archives several dozen sea ice data sets, one dozen of which are derived from passive microwave remote sensing. This is a confusing array of data, even for a specialist in sea ice research. NSIDC also recognizes that data stewards need to guard against scientifically inappropriate use of their data (Parsons and Duerr, 2005). In total, NSIDC provides hundreds of pages of data documentation along with a Web site that compares and contrasts their sea ice products, including descriptions of the strengths, weaknesses, and appropriate applications of each product.[a] These tools—which were developed in response to feedback from the user community, both informally and through an established user working group—are intended to guide scientific users to the product that most suits their application. NSIDC has also developed specific products geared to nonexpert users, each of which is accompanied by a description of appropriate uses for the data. The best known among these is the Sea Ice Index, which provides images of average monthly ice conditions and trends, along with anomalies that compare recent conditions with the long-term averages (Fetterer and Knowles, 2004). NSIDC also provides higher-level informative products geared specifically to educational users, the media, and the general public.

[a]http://nsidc.org/data/seaice/.

other stewardship functions, requires communication among data stewards, data users, and data set experts. Typically, the scientific assessment and improvement function is initiated by data set experts as they use data and metadata and uncover problems such as bias offsets, missing records, or systematic errors in spatial or temporal identification. However, input from nonexpert users can also be extremely helpful. In addition to being valuable to the experts responsible for data set maintenance, assessment, and improvement, feedback logs would provide valuable information to users about the quality, utility, and appropriate applications for the data, as well as helping them find related data sets. Chapter 6 discusses these additional benefits of user feedback and recommends including user feedback logs as a regular component of the metadata associated with

each data set. Formal efforts to engage users in data management activities, such as user panels, can also be employed to improve all the major data stewardship functions. To the maximum extent possible, these user feedback mechanisms should be made a regular part of NOAA's data management activities.

Guideline: Stewardship plans should be consistent but flexible so that improvements in data and metadata can be captured and included in the data systems.

In order to realize the scientific and societal benefits of improvements in data and metadata over time, the data management system should capture and reflect the information that leads to increased knowledge about each data set and its uses. Data management systems should therefore be designed to support the growth of archived information that occurs during the data life cycle, including metadata that is outside of current standards and requirements. As an example, consider a data steward who comes to recognize that a particular data set has broader relevance than is currently being realized and collaborates with other data stewards to achieve wider data distribution. These stewards might work together to transfer the data from one of NOAA's centers of data to one of the National Data Centers, either once or on a routine basis if the collection is continuously growing. Following the transfer, the data should be verified for completeness and accuracy; the metadata should be captured and standardized; an optimal organization should be established to serve current and future users; and dataset provenance should be documented. Finally, the collection is offered to users by the best available methods and scaled appropriately for the data set.

In the broader context of data discovery, access, and integration, it will also be necessary to design data management systems that can accept and disseminate a wider array of metadata as it expands over time. Many user benefits can be derived from quality metadata, including informative Web interfaces for users and increased interoperability through implementation of standards and programming interfaces. The importance of continued metadata improvement cannot be overemphasized and is an important key to enabling integration of data services across NOAA.

DATA SUPPORT SERVICES AND TOOLS

Guideline: Each data set should have an identified expert who understands the data and who is responsible for working with data

providers and data users to maintain the archive and make certain that data access systems meet user needs.

Data support services work at the interface between the data archive and the data user, while data support tools are the applications used to analyze and understand the data. Favorable user experiences are strongly linked to how well these stewardship functions are provided. Of particular importance are the personnel who work both individually and collectively to acquire, preserve, improve, and disseminate the data and metadata under their purview. These data stewards are critical for successful and effective service because they serve as the bridge between data collectors/generators and end users. In addition to understanding both the data and its end uses, they can often anticipate user needs and organize the information in a manner that helps meet those needs, can work with technology experts to design optimal archival and access systems, and can understand and help resolve user problems and questions.

Stewardship starts with the investigator or investigators who first collect or obtain the data. For some data sets (for example, time series from a cruise, aircraft, or other *in situ* measurement) the investigators might collect the data directly, while for others (for example, surface weather observing networks), the investigators might include both volunteer observers and the NOAA staff who routinely monitor the instruments, collect the data from a number of locations, and perform initial quality control. In many cases, one or more of the investigators are already located at a center of data or a National Data Center.

For all situations, there should be transfer protocols in place that describe how the data should move from the investigator to an archive. Such protocols assure mutual understanding and increase the likelihood that a complete set of data will be archived for future use. Depending on the situation, the transfer protocol may range from agreements between individuals who routinely transfer data and certify each other's actions with a data receipt to formal interagency or international submission agreements. The same principles used to transfer data from investigators into the NOAA archive system should be applied and formalized for the cases where data from centers of data need to be transferred to one of the three NOAA National Data Centers.

Since data stewards are the primary contact point for information entering and exiting the archive, they should be familiar with the full spectrum of capabilities and standards of the archive and access system, as well as all legal, mission, and administrative requirements (see Chapter 2). Well-designed systems will satisfy a majority of users with automatic data access, but there will always be questions from users, so person-to-person data usage consultation remains an important data support

service. An active consultation service has several benefits: it helps detect problems with the archive; it helps data stewards gauge how the user community is evolving; and it improves user satisfaction. By maintaining substantial and ongoing communication with users, stewards can build and maintain more effective access tools and facilitate the discovery and integration of data to meet new user needs. (These topics are covered in additional detail in Chapter 6.)

Data stewards should also work collaboratively with data experts, data providers, and principal investigators to establish agreed-upon processes for maintaining the current flow of information into the archive, identifying and planning for new data streams, and, as discussed in the next section, assessing and improving the data and metadata already in the archive. For future data-generating activities, it is beneficial to involve data stewards early in the planning process so that data archiving, stewardship, and access can be planned ahead of time and thus more economically. Finally, data stewards need to collaborate with their counterparts at other agencies and international groups to ensure that all important environmental data are archived and made accessible.

Guideline: Decisions to upgrade legacy data support systems should be thoroughly planned, in consultation with users, with particular attention paid to preserving Web-based metadata.

Many of NOAA's legacy data archiving and access systems are providing stable and reliable services to users. Decisions to upgrade these systems should be thoroughly planned, with input from users, to avoid costly interruptions in service and to make sure that upgrades actually result in demonstrable improvements in access, discovery, and integration. A related issue is that Web-based infrastructure currently holds vast quantities of context information for data sets and projects. Much of these metadata are not being archived in a manner that permits easy association with the data for purposes of long-term preservation. This issue is complicated by the fact that Web-based information can easily be linked to information elsewhere on the Web, making it difficult to capture a complete set of metadata for long-term preservation. New methods are needed to address this issue.

DATA STEWARDSHIP AT NOAA

Establishing and maintaining effective data stewardship is a crucial but challenging task for any program or agency with data management responsibilities, but this task is especially challenging for NOAA because of the volume and diversity of its data holdings and the number and

diversity of its users. While stewardship is mentioned in virtually all of NOAA's data management planning documents, there may not yet be an agency-wide appreciation of the difficulty and challenge of data stewardship. Various guidance documents also tend to use the term stewardship in different and sometimes inconsistent ways, which can lead to confusion in determining the appropriate roles for various components of NOAA's data management enterprise. For example, the word "stewardship" is part of the CLASS acronym, yet current plans for CLASS indicate that the data centers are expected to provide the stewardship function. While this latter division of responsibilities seems most appropriate due to the close relationship between the data centers and their users, in the absence of a comprehensive agency-wide data management plan, confusion over stewardship responsibilities is likely to persist.

Based on our review of current NOAA practices, a synthesis of previous reports, and application of concepts that have gained broad acceptance in the data management community, we believe the time is ripe for the development of specific guidelines, based on the general guidelines provided in this chapter, to develop a more consistent and integrated stewardship program at NOAA. Such a common understanding, coupled with a goal of being good stewards, can help guide NOAA in making appropriate data archiving and access decisions. Box 4-2 illustrates the potential benefits of an effective data stewardship program. The next two chapters discuss these aspects of data management in further detail, while Chapter 7 describes the overarching data management plan that is needed to bring together all three functional elements of data management.

BOX 4-2
NCEP-NCAR Global Atmospheric Reanalysis: An Example of the Benefits of Data Stewardship

The National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) Reanalysis provides a standardized time series of the evolution of the global atmosphere starting in 1948 (Kalnay et al., 1996). This suite of data products, which was made possible by scientific community foresight and resources from an operational numerical weather prediction center, has proven to be an extremely valuable resource for a broad range of scientific studies, and in many respects it serves as a model for effective stewardship practices.

More than 40 year ago, long before data stewardship became part of the data management vocabulary, Roy Jenne, then at NCAR, and other like-minded data managers and providers at NOAA and NASA began sharing, collecting, documenting, and preserving a wide variety of *in situ* and satellite observations of atmospheric conditions. These data and metadata preservation activities established long-term collections that ultimately became a major input data component for the data assimilation and forecast processes used to create the NCEP–NCAR Reanalysis. In fact, it would be fair to say that reanalysis would not have been possible over the nearly 60-year period of record or at the fidelity we have come to expect today without the activities of Jenne and his colleagues.

Stewardship is much more than data preservation. A typical dataset used in the NCEP-NCAR Reanalysis has gone through several phases of preparation, usage, and improvement, and careful stewardship is required during each of these phases. For example, through user studies that analyze the data for scientific content and complimentary examinations and quality control by data stewards, fixable problems and systematic errors are occasionally uncovered. Data stewards are critical during this stage because they work with scientists to fully define the problem and to design processes to fix the data. After the data are improved, the metadata are updated and user access is refreshed with a new version of the data set. In this way historical data sets improve in quality through a stewardship-supported life cycle (see Figure 4-1). Many data sets used in the NCEP-NCAR Reanalysis have experienced several such iterations, and this process has also benefited subsequent reanalyses, such as the European Center for Medium-range Weather Forecasts (ECMWF) ERA-40 data set. The importance of data stewardship for the success of these efforts has been documented, and community-based plans have been developed to make further improvements (Schubert et al., 2006).

Stewardship also supports the NCEP-NCAR Reanalysis (and other reanalyses) by providing data security, multiple access points for products, user consulting, and routine time series extension. NCEP does not archive the Reanalysis for public access; rather, it distributes copies to both NCAR and NCDC to ensure that the data are securely preserved. These two institutions also provide data support services and a variety of access methods to obtain the NCEP-NCAR Reanalysis. NOAA's Earth Systems Research Laboratory (ESRL) provides additional support, such as transforming some of the most popular data to different formats so users can employ their favorite tools—for example, direct access by remote applications—to exploit the data. NCEP's collaboration with these other groups results in extensive and well-supported data availability for thousands of users worldwide.

5

What to Archive

PRINCIPLE #7: A formal, ongoing process, with broad community input, is needed to decide what data to archive and what data not to archive.

NOAA needs to establish a high-level, enterprise-wide approach to decide what data to include in their archives. The decision to archive (or not to archive) data should be driven by societal benefits, should ensure that irreplaceable environmental data are preserved, and should explicitly incorporate broad community engagement and coordination with other agencies.

Before a diverse group of users can be provided with a wide spectrum of useful and reliable environmental data, it must first be decided what data to include in the data archive. The decision to archive or not to archive a data set is often made before the data actually become available (for instance, as part of the requisite data management planning now required for all new NOAA data streams), and the decision must be repeated each time the data are revised or reprocessed in accordance with the data stewardship guidelines and practices described in Chapter 4. The iterative nature of this decision-making process is illustrated in Figure 5-1, which extends Figure 4-1 to indicate the two points during the data life cycle when the decision to archive or not to archive data is made.

For some data sets, archiving requirements are explicitly spelled out in legislation, administrative orders, or agreements with other agencies. However, as noted in Chapter 2, these requirements are often not very

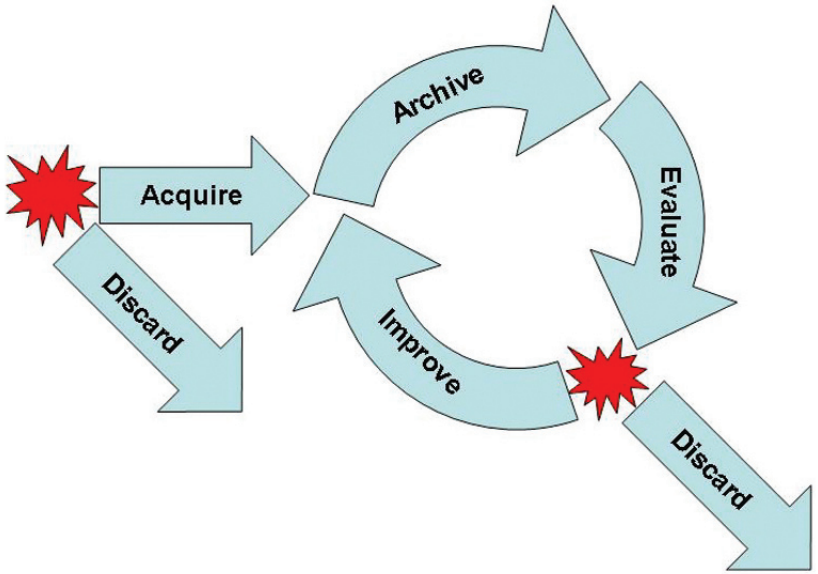

FIGURE 5-1 The data management life cycle, including decision points. Data and associated metadata are brought to the attention of data stewards, at which point a decision is made either to discard the data and metadata or to integrate them into the archive and access system. As described in Chapter 4, data stewards work with users to evaluate the data over time, leading to additional knowledge about the data. This knowledge might include fixable problems with the data or scientific findings that add to the metadata. If improvements are possible, they should be applied, at which point the archiving, evaluation, and improvement cycle begins again. If the problems are not fixable, or if the data are determined to be no longer useful, the data may be discarded in accordance with the guidelines described in this chapter.

specific, leaving data managers with considerable flexibility, but also little guidance, for determining which data sets to archive. Based on this committee's review of current data management practices at NOAA, it appears that most archiving decisions are made in a deliberate manner and with a concerted effort to meet user needs. There have also been some preliminary, sporadic attempts to involve users more directly and to improve communication with other agencies. However, the decision-making process is largely ad hoc from an enterprise-level perspective, and stakeholders are not engaged in a systematic way. The lack of a complete, publicly available inventory of all federal environmental data holdings and a formalized, inclusive process for making archival decisions not

only decreases the efficiency and the effectiveness of NOAA's entire data management enterprise, but also leads to a situation where valuable environmental data could potentially "fall between the cracks" (see Box 3-1). In addition, stakeholders often develop their own products from NOAA data sources. NOAA should remain engaged with these stakeholders and incorporate these supplementary products into the agency's archives if they have the potential to prove useful for a significant number of users.

To ensure that all important environmental data are archived and improve the overall efficiency and effectiveness of NOAA's data management enterprise, the committee recommends that NOAA establish a formal, ongoing process to evaluate the need for original and continued archiving of all environmental data sets that may fall under NOAA's mission, considered in the broadest possible context. This principle represents a more focused manifestation of principle #9 (explored in further detail in Chapter 7) that "effective data management requires a formal, ongoing planning process." As with other aspects of data management, the decision-making process for data archiving should be driven by the value to society, should incorporate periodic input from users, and should include ongoing coordination and collaboration with other agencies and international groups. Creating a complete, publicly available inventory of NOAA's data would be a logical first step, and the resulting guidelines and procedures for data archiving would need to be sufficiently flexible and adaptable to accommodate the ongoing nature of the data life cycle and the wide spectrum of different kinds of environmental data managed by NOAA's National Data Centers and centers of data (the entities that would ultimately be responsible for implementation). In the remainder of this chapter we explore the nature and nuances of data archiving decisions in additional detail and provide general guidelines that NOAA and its partners can use to develop more specific rules and procedures for deciding which current and future data sets to include in their archives.

DATA TYPES

Throughout this report, the term "environmental data" is used broadly to indicate all types of environmental Earth System observations (including physical samples as well as in situ and remotely sensed data), model output, and synthesized products derived from these data. NOAA has defined seven different general types of data (see Appendix C), of which five are likely to contain candidate data sets for archiving under NOAA's mission. Those five are: (1) Original Data; (2) Synthesized Products; (3) Hydrometeorological, Hazardous Chemical Spill, and Space Weather Warnings, Forecasts, and Advisories; (4) Natural Resource Plans; and (5) Experimental Products. The guidelines that follow address the

archiving requirements for some of these data types in detail, while others are examined in a more general sense. As part of the formal, ongoing planning process that we recommend NOAA develop and implement, NOAA may wish to revise their data type definitions to include more specific guidance on the types of data that should be archived, thus improving enterprise-wide coordination.

Guideline: It is especially important to save the most primitive useful forms of all environmental data.

Original Data, which represent the most primitive useful form of environmental observations, should always be considered for long-term archiving. Most environmental observations are irreplaceable, since resampling is usually impossible, and are expensive to obtain, especially compared to typical data management costs. Although it is extremely difficult to assess the quantitative value of most Original Data, especially for future applications, access to a broad and continuous collection of high-quality observational data is essential to long-term environmental monitoring, model development and testing, and many other areas of Earth System research. For instance, in the report *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences* (CCSP, 2006), archived radiosonde and satellite measurements were used to resolve a debate concerning the amount of warming observed in the troposphere. Since they cannot be regenerated or reproduced from other archived data, Original Data should always be considered for archiving and should be preserved even when improved or reprocessed versions of the same data are later developed. In fact, it is advisable to begin each new iteration with Original Data so that the quality and error assessment of new versions is independent of previous data manipulations.

Although it is critical to plan for the archive growth associated with current and future data streams, many applications, such as climate change detection and attribution, require data collected over extended periods of time. A number of Original Data sources critical to NOAA's mission are currently at risk or are essentially inaccessible, often because they are stored on deteriorating, substandard, or outdated media. Although some of these data might be considered "permanently stored," they cannot be considered properly archived if they are inaccessible to researchers should an important future application arise. NOAA should make certain that critical at-risk observations are at least migrated to stable media; it would be preferable to add these data to NOAA's archives. The Climate Data Modernization Program[1] has already initiated a number of effective data

[1]http://www.ncdc.noaa.gov/oa/climate/cdmp/cdmp.html.

"rescue and recovery" efforts that have resulted in drastically improved archiving and access of many at-risk data sets. NOAA's National Data Centers should continue and expand these efforts by working with users and other agencies to identify, obtain, and archive at-risk environmental data.

Guideline: It may be more cost-effective to regenerate certain kinds of environmental data on demand.

Some types of Synthesized Products, such as model output, reprocessed data, and analyzed or derived products, could potentially be regenerated from other archived data. In some cases, it may be more cost-effective to regenerate these products on demand, rather than archiving every single version of every single data set associated with a particular data stream. Proper care needs to be taken to ensure that the hardware and software needed to perform the requested operation remains supported and that it provides the exact same high-quality information as a true archive. These two requirements are challenging because processing algorithms—especially model code—are often complicated and machine dependent. For on-demand processing or reprocessing, it is also imperative to maintain an easily accessible archive of "first-stream" data, which could be the Original Data (such as radiosonde data) or a product generated from Original Data (such as radar precipitation data). The evolution of the data should also be fully documented, including records of successive improvements, recalibrations, and other modifications (including changes to the model code) that impact regeneration. Proprietary algorithms should also be avoided unless they are fully transparent and reusable.

Box 5-1 contains an example that illustrates how processing on demand has proven to be useful in reducing the archive volume for Moderate Resolution Imaging Spectroradiometer (MODIS) "Level 1b" data. NOAA should consider employing similar techniques for some of its data, especially the data expected from the National Polar Orbiting Environmental Satellite System (NPOESS) and NPOESS Preparatory Project (NPP). This will require careful planning and close collaboration with federal agency partners (NASA and the Department of Defense) and private contractors. However, some caution is also warranted since there is not a general consensus within the data management community about the true utility and reliability of processing or regenerating data on demand versus archiving all levels of each data set. In addition, even though some applications may benefit from a process-on-demand approach, there may be legal concerns with failing to archive all intermediate versions of a data set. For example, certain versions of climate data sets are cited in a

BOX 5-1
Regenerating Versus Archiving MODIS Data

NASA has taken responsibility for archiving and stewardship of Earth Observing System (EOS) satellite data. In order to improve the cost-effectiveness of the Moderate Resolution Imaging Spectroradiometer (MODIS) data archive, the data will be moved from the central Goddard Earth Sciences Distributed Active Archive Center (DAAC) to the decentralized MODIS Adaptive Processing System (MODAPS). In addition, the "Level 1b" data (calibrated, geo-located radiances based on "Level 1a" data of radiance counts) will be changed from "permanent archive" to "regenerate on demand" status (Maiden, 2006). This change will be based on an extension to existing MODAPS science team production systems that produced only higher-level (higher than Level-1b) MODIS land and atmosphere products in the past, and it will use commodity-based systems for processing and online data storage and access. Anticipated benefits of this transition include:

- Reduction in archive growth through on-demand processing;
- Faster access to products due to reduced processing time resulting from all online storage;
- Reduced costs due to use of commodity disks and reduction in operating costs at the DAAC;
- 10 percent fewer products and 90 percent reductions in total archive volume and archive growth at the Goddard Earth Sciences DAAC; and
- Closer involvement and control by the science community, which is expected to make the products, tools, and processing capabilities at the archive more responsive to scientific needs.

variety of peer-reviewed publications and assessments. Future requests to access these data for verification purposes need to be accommodated, even though more recent versions of the data are available. The next two guidelines address some of these concerns.

Guideline: The most obvious candidates for reduced archiving requirements are data that are obsolete or redundant, that could be regenerated on demand, or that clearly have only short-term uses. This includes older versions of reprocessed data and model output.

One strategy that could help data managers make practical but careful archiving decisions is to first identify those data that clearly have only short-term uses. Candidate data sets include those whose original and predicted purpose has been satisfied, those that may be cost-effectively regenerated from archived first-stream input (that is, when the cost of

storing the data exceeds the cost of reproducing or regenerating the data), or those that are obsolete or redundant. These data may include Experimental Products (see Appendix C) or data collected for specific short-term applications such as near-term operational decisions, because these data typically have little value after the decision, product, or improvement has been made and can often be reproduced if necessary. Other candidates may include high-resolution data intended for modeling or image display applications that would be adequately served with lower-resolution versions of the data, or data used in field research programs (for example, satellite data over a limited area) that are redundant with data held in global archives. Some of the data used to create Hydrometeorological, Hazardous Chemical Spill, and Space Weather Warnings, Forecasts, and Advisories (Appendix C) may also fall under this guideline, such as long-range and intermediate forecast model output or high-volume data (sub-second wind or radar returns, for instance) used to generate longer averages of specified parameters for operational reporting purposes.

In situations where multiple versions of derived products have been generated, it would be helpful to have a defined process in place to determine which versions need to be archived. The following three questions, for example, could form the basis for such decisions:

1. Is it feasible to retain multiple versions of the data?
2. Are the differences among the various versions sufficiently large and scientifically important to make it worth preserving multiple versions?
3. Is it too technically difficult to regenerate earlier versions?

If the answer to all three questions is positive, then multiple versions should be archived. For example, NCEP has produced and is currently extending both the NCEP-NCAR Reanalysis (see Box 4-2) and the NCEP/Department of Energy (DOE) Reanalysis (Kanamitsu et al., 2002). These two data sets could be considered different versions during their period of overlap (1979 onward). However, they have some large and scientifically important differences: for example, the NCEP/DOE Reanalysis is generally viewed to have superior surface radiation analyses, while the NCEP-NCAR Reanalysis is critical for climate assessments because it begins in 1948. In addition, many scientific papers, assessments, and other publications have been produced based on earlier versions of the NCEP-NCAR Reanalysis. Furthermore, due to the complexity of the models used to perform these reanalyses, archiving earlier versions would be much more cost-effective than regenerating them from archived first-stream input. Comparative use of multiple reanalyses, including multiple versions of the same reanalysis, is also likely to be helpful to those responsible for generating future versions of the reanalysis. Thus, none of the criteria for

designating the earlier version as a candidate for disposal is presently satisfied.

In contrast, the MODIS team (see Box 5-1) has produced four versions of MODIS data, with the fifth version due to be delivered in 2007. All of the derived products included in the initial two versions would be candidates for disposal, because all three of the above criteria are met: as initial products, their original purpose has been largely satisfied; the archive cost of these products is quite high; and they have become obsolete due to their known serious deficiencies. Similarly, the volume of ensemble weather prediction model output may be too large to retain complete data on all members of the ensemble, although there have been a variety of plans put forward to archive subsets of these forecasts for research purposes. An even more complicated example is described in Box 5-2.

Guideline: Archiving and access decisions are closely related. In general, when resources are limited, access to older or less commonly used data should be scaled back rather than removing data from the archive.

As noted previously, not all data sets are of equal value, and practical constraints prevent all data from being archived and made readily accessible, so at some point certain data will need to be designated for reduced archiving and/or access requirements. Ideally, this decision would be made based on the current utility and potential future value of the data, but as noted at several points in this report, it is extremely difficult to assess even the current value of any particular environmental data stream (see, for example, Millard et al., 1998). Likewise, it is virtually impossible to anticipate its potential future uses. The decision-making process also needs to be ongoing, with data managers/stewards continually reviewing the data holding under their purview to determine the appropriate level of service for each data set given legal and mission requirements, user needs, and available resources.

A difficult task, but one that could yield major economic benefits for NOAA, is to determine the realistic and required level of service for each archived data set. For example, large but little-used large data sets could be placed in a "deep archive" that is secure and offers a clearly defined access path that potential users could discover, but the data would not be readily available on demand. This strategy would be particularly appropriate in cases where the decision to stop archiving a data set would be irreversible (that is, the data could not be regenerated or resampled). One such example might be the hard copies of original high-resolution

BOX 5-2
Multiple Versions of MSU Data

A Microwave Sounding Unit (MSU) is a satellite instrument that measures radiation emitted from the Earth's atmosphere, with different frequencies corresponding to different atmospheric layers. Data from these instruments, which have been installed on a series of polar-orbiting satellites operated by NOAA since 1979, have been analyzed and pieced together to yield a global record of atmospheric temperature commonly referred to as the "MSU temperature record." Not only is this record constantly being extended, but a number of subtle errors have been detected and removed, and different versions are being produced by different groups using different analysis techniques (Christy et al., 2000, Christy et al., 2003, Spencer et al., 2006, etc.). The Climate Change Science Program's (CCSP's) Synthesis and Assessment Product 1.1, *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences* (CCSP, 2006) discusses these and other issues associated with constructing records of global temperature change in the lower atmosphere.

The MSU temperature record poses an interesting and illustrative challenge for data archiving because it is not immediately clear how many different versions of the data should be archived and made readily available to users. The most recent versions of the data will obviously be in high demand, but older versions might also be of interest due to the public controversy that has surrounded some of the revisions to the data, as well as the fact that some prominent publications (including international assessments) have referenced the older data sets. Thus, it would be prudent to provide some degree of access to at least the last few versions of the data. Detailed descriptions of all revisions should be made available in the metadata for each data set, and ideally the metadata for older versions will contain information about more recent versions of the data and also the publications that have used the data. Given the high level of public interest in the data, it would also be advisable to create a standing advisory group to provide ongoing advice on the level of archiving and access needed for MSU data and other similar types of data.

(one-third of a nautical mile) Defense Meteorological Satellite Program (DMSP) satellite imagery, which is archived at the National Snow and Ice Data Center for use in analyzing snow cover and is only available in hard copy format. However, it is imperative to make certain that the data could still be retrieved if needed, and implementation of this guideline at NOAA would likely require some modification to existing policies across the agency.

THE IMPORTANCE OF STAKEHOLDER INVOLVEMENT

Since the benefit to society is the ultimate rationale for federally supported data archives, the decision to archive or continue archiving a data set or a general category of data should be driven by the societal benefits that the data provides or could provide. However, there is a vexing problem associated with this imperative: while the societal benefits of environmental data are both ubiquitous and diverse, estimating the present and especially the future value of any particular environmental data set or data type is extremely difficult. This difficulty complicates all data management decisions, especially the decision of what data not to archive. While data managers can use a variety of strategies to help inform their archiving decisions, the most effective and essential practice is to actively engage a broad range of stakeholders on an ongoing basis. These stakeholders include the current and potential future users of environmental data, data managers from other program elements at NOAA, and data managers at other federal agencies, universities, and international entities. The following section focuses on the usefulness and importance of broad stakeholder involvement in data archiving decisions.

Guideline: It is essential to solicit user input when making decisions on whether to archive or continue archiving a data set.

There are many aspects of data management that benefit when users are engaged in a substantial, active, and ongoing manner. In regard to data archiving decisions, which are inherently complex, feedback from users is critical to help data stewards evaluate the current uses and potential future societal benefits associated with a particular data set or general category of data. While such evaluations must necessarily be qualitative in nature, even a relative assessment of the current or likely future value of data is critical for three main purposes: (1) to ensure that all important data are preserved; (2) to identify data sets that could potentially be reprocessed or regenerated on demand; and (3) to inform decisions about the appropriate level of service for each data set. For example, advice from users can help data managers decide on the appropriate resolution of data to include in the archive for observations taken at very high temporal, spatial, and/or spectral resolutions, such as radar or satellite data. The participation of expert users is also helpful for identifying obsolete or redundant data and for identifying important or irreplaceable at-risk data that are currently not in the archive.

Even when user input is solicited, deciding what environmental data to archive is difficult because the user communities are large and diverse and each data set has a broad range of potential future applications. The

archiving requirements for global change science applications are particularly difficult to predict because the field is young and rapidly evolving. These considerations reemphasize the need for a broad, formal, and ongoing process to inform data archiving decisions. In addition to engaging "pure" users, who often focus on a particular application or type of data, the advice of data scientists both inside and outside of NOAA who critically evaluate data, integrate data from different sources, and develop value-added products should also be weighed carefully in any decision to archive or stop archiving a particular data set. In addition to being aware of data quality, applicability, and limitations, data stewards should be aware of the past uses for the data as well as new research, data sources, and processing algorithms that might influence the future usefulness of the data. Data usage metrics are a valuable tool for assessing the current status, potential future developments, and level of service required for different data sets, but communication with data providers is also essential. NOAA might wish to consider implementing a decision-making process similar to the one described in Box 5-3.

Guideline: Because the decision to stop archiving is normally irrevocable, extra attention to community engagement is needed before final disposal of any data.

One of the most important practical functions that users can serve is to identify data sets that could be considered for disposal or for reduced levels of service when all foreseeable uses of the data have been exhausted or when more current or comprehensive information is available elsewhere. In the case of decisions to stop archiving data, which are irrevocable, it is essential to notify and actively engage the broadest possible spectrum and number of users before final disposal of the data. Currently, NOAA data managers only notify known stakeholders (that is, historically active users and, sometimes, data managers at other agencies) about the potential loss of data. With the continuing expansion of multidisciplinary analysis and the increase of new potential users that will result from improved data discovery and integration capabilities (see Chapter 6), a better enterprise-level notification system should be established to advertise data management decisions, especially decisions to dispose of data but also decisions to significantly reduce levels of service. The decision to stop archiving a data set should at least be communicated to any current users and could also be posted in the federal register at least one year prior to the scheduled disposal date. This would give all users, including other federal agencies, nongovernmental organizations, universities, and international groups, an opportunity to comment on the decision or take over archiving responsibilities.

BOX 5-3
Decision-making Process for Archiving Data at EROS

The records disposition schedule process at the U.S. Geological Survey's (USGS's) Earth Resources Observation and Science (EROS) data center[a] is an example of a formal, enterprise-wide policy for data disposal designed to preserve essential data while maximizing cost-effective disposal of data that are no longer needed. The three objectives of the program are to:

1. Preserve records that have permanent value;
2. Destroy records of temporary value as soon as they have served the purpose for which they were created; and
3. Remove old records from office space and filing equipment to storage facilities, thereby improving use of files and reducing maintenance costs.

To accomplish these objectives, USGS developed mission-specific records schedules for geological, geographical, and water data; a schedule for their biology records program is currently being planned.[b] These records disposition schedules provide mandatory instructions on what to do with records (and non-record materials) that are no longer needed for current business and also the authority to dispose of recurring or nonrecurring records. There is an online tool to facilitate this decision-making process.[c] Overall, this process provides an efficient, cost-effective, and transparent method for making data archiving decisions.

---

[a]http://edc.usgs.gov/.
[b]http://www.usgs.gov/usgs-manual/schedule/432-1-s1/howtouse.html.
[c]http://eros.usgs.gov/government/RAT/tool.php.

Guideline: NOAA should establish close partnerships with other national and international data holding institutions and engage these institutions as part of the archiving process. It is important to have clear agreement on which partner has what archival responsibility.

All environmental data collected or generated using governmental resources or that fall under NOAA's mission need to be considered for archiving. However, many different federal agencies, state and local agencies, universities, international organizations, and other groups are involved in the collection and management of environmental data. For example, some of the international organizations responsible for collecting and distributing environmental data relevant to NOAA's mission include the Group on Earth Observations (GEO), Committee on Earth

Observation Satellites (CEOS), International Council for Science (ICSU), World Meteorological Organization (WMO), International Oceanographic Commission (IOC), and International Hydrographic Organization (IHO). NOAA should coordinate with all of its domestic and international partners to establish common standards and protocols for the broad spectrum of environmental data that falls under its mission, to ensure that all important data and metadata are archived and made available to users, and to maximize the overall efficiency and cost-effectiveness of government archiving activities.

The importance of using agreed-upon standards and protocols to facilitate and promote data stewardship was discussed in Chapter 4, and Chapter 6 includes additional discussion of the benefits of interagency and international coordination for data discovery, access, and integration. Before these higher-level benefits of interagency and international coordination can be realized, however, a formal process should be developed and implemented to determine which agency is responsible for archiving each data set. Since archiving and access are so closely related, ideally this process would be part of a full evaluation of roles and responsibilities throughout the data management life cycle. Coordination is essential to improve cost-effectiveness, to reduce duplication and redundancies, to identify at-risk data streams, and to make sure all critical environmental data are archived.

The 1989 and 1992 memorandums of understanding (MOUs) between NOAA and NASA, discussed previously in Chapter 2 and Box 3-1, illustrate both the benefits and the pitfalls of interagency agreements. Originally, these MOUs called for careful coordination between the two agencies and suggested that NOAA would be willing to provide the long-term archive for Earth Observation System (EOS) data generated by NASA satellites. However, NOAA has subsequently assumed archiving responsibilities for only a small fraction of EOS data (specifically, the Moderate Resolution Imaging Spectroradiometer [MODIS] Level 1b data), with others not being archived. This decision was presumably driven by lack of resources and has been justified by citing a lack of relevance to NOAA's operational mission, even though much of the EOS data is critical for global change research. Moreover, a number of program elements within NASA were not made aware of NOAA's decision for some time. Fortunately, NASA has subsequently found a way to take on the archiving responsibilities for the remaining EOS data sets through several of its distributed data archives, but it is unclear how discoverable these data will be to regular NOAA archive users or what further inefficiencies this unfortunate lack of coordination might engender. In circumstances such as these, where specific archiving responsibilities were not agreed to at

the beginning of the data collection campaign, closer ongoing collaboration is needed to make certain that all environmental data are properly archived. If interagency agreements continue to prove ineffective, other mechanisms for improving coordination, such as oversight at the CCSP, Office of Management and Budget, and/or Office of Science and Technology Policy levels, should also be considered.

6

Data Discovery, Access, and Integration

PRINCIPLE #8: An effective data archive should provide for discovery, access, and integration.

NOAA's environmental data should be made easily discoverable and readily accessible to a broad range of users, and its data management system should be designed to allow the integrated exploitation of data from multiple sources within and outside the agency to answer environmental questions and support NOAA's mission.

The utility of any data archive is ultimately defined by the ability of users to make use of the data it contains. Improving access to the environmental data in NOAA's archives is, justifiably, the main focus of many current data management activities. Simply making data available upon request, however, is only the first step toward a data management system that promotes extensive utilization of different kinds of environmental data to address a broad range of societal issues. An equally important but often neglected element of effective data management is the ability of the access system to support and promote the discovery of data that might be brought to bear on specific questions or problems, ideally without the user having prior knowledge of the contents of the archive. A third essential functional element of an effective environmental data management system is its ability to facilitate the integration of multiple data sets from different sources to address increasingly inter- and multidisciplinary environmental topics. These three crosscutting functions are the main

focus of this chapter, but as a base for those discussions, we first explore the characteristics of NOAA's data users.

USER CONSIDERATIONS

One of the overarching principles introduced in Chapter 3 is that "data management activities should be user-driven." The previous two chapters have noted the importance of user input for data stewardship and data archiving decisions. User input is also critical for driving the design, implementation, and maintenance of data access capabilities, particularly the crosscutting capabilities of data discovery and data integration. This section explores some general characteristics of NOAA's user community that are important to consider when improving the data management system to promote extended or enhanced capabilities for data discovery, data access, and data integration.

Guideline: NOAA's data access systems need to account for diverse user needs and capabilities.

Effective data access should be tailored to the needs and capabilities of different kinds of users. For example, the technical capabilities and scientific sophistication of NOAA's user base ranges from elementary school children looking for temperature data near their house for a school project, to inexperienced users seeking highly specific weather and climate information for legal or business purposes, to experienced modelers looking for a large volume of well-calibrated data to test a physical parameterization, to multidisciplinary scientists seeking to project the combined effect of multiple environmental stresses on a particular ecosystem. Users may also prefer data in a particular format: ASCII data to import into a spreadsheet, JPEG images, or GIS-compatible formats, to name but a few. Access systems should strive to serve as many of these levels of scientific skill, technical sophistication, and packaging preferences as possible. One approach that has been successfully employed, which NOAA may wish to consider adopting, is to create several different access pathways for different user classifications, starting from the same access portal.[1] A good example is the Archiving, Validation and Interpretation of Satellite Oceanographic data (AVISO) Web site.[2]

[1]A web portal is a term, often used interchangeably with gateway, for a World Wide Web site whose purpose is to be a major starting point for users when they connect to the Web (Alexandrou, 2007).

[2]http://www.aviso.oceanobs.com/.

Guideline: The needs of the user community should be evaluated on a continuing basis to improve the effectiveness of data access tools.

As noted in Chapters 3 and 4, user surveys and other feedback mechanisms are among the most important tools available for evaluating and improving the data in an archive, as well as for ensuring that important environmental data are preserved. User feedback can also improve many aspects of data access because it provides important insights about the effectiveness of the various discovery, search, and access tools and protocols, as well as how these processes could be improved. Since the needs of the user community are constantly evolving, this feedback needs to be collected on an ongoing basis and incorporated into the ongoing, enterprise-wide data management planning process. Feedback mechanisms should also be imbedded within the system implementation to provide an avenue for immediate suggestions and comments on the effectiveness of the data management system. This feedback is invaluable because it allows NOAA to better understand the capabilities, needs, and characteristics of its customers and to design their data access tools and user interfaces to match these requirements.

As with many aspects of data archiving and access, obtaining user feedback is fraught with many practical challenges. For example, users do not always know what data or data access tools they would find most useful. Detailed surveys, crafted by sociologists rather than data scientists, would be extremely useful in characterizing NOAA's current and potential users. However, the Paperwork Reduction Act of 1995 makes it difficult for federal agencies to conduct formal user polls or surveys.[3] NOAA will thus need to be creative to ensure that user needs are being met and to develop strategies for improving data access capabilities. For example, even lower-level feedback such as Web-based tools that anonymously capture the activities and search patterns of individual customers would allow NOAA to obtain a better understanding of its customers and their needs. Likewise, usage metrics can be used to capture the net activity of the data user population; NOAA should consider collecting a set of standardized metrics enterprise-wide to help expose data discovery, access, and integration gaps.

To be most effective, the feedback collected from a broad range of users should be accompanied by more formal, focused feedback from external advisory groups or user panels. These focused groups, which should include scientific experts as well as representatives from other important user groups, can be tasked to take into account the broad user metrics and feedback and provide specific, balanced advice to data

[3] 4 U.S.C. 3501 et seq.

managers. Focused stakeholder panels can also provide advice on data archiving requirements, as discussed in Chapter 5. Additional methods for incorporating user feedback to improve data discovery, access, and integration are described in the sections that follow.

DISCOVERY

Guideline: Environmental data should be easily discoverable by a broad range of users. In particular, data discovery should not require any specific knowledge about the data or how they are managed.

Discoverability is an essential characteristic of a data management system that promotes the full exploitation of archived data. NOAA's current data access systems tend to be structured mainly to meet the needs of the scientific community. For example, since scientific users tend to be technologically sophisticated, with extensive knowledge of both the different data types that are available and where different data sets can be found, the structure of NOAA's access systems tends to emphasize efficient access to contiguous or closely related data volumes. The onus is placed on users to find the appropriate portals for access to particular data sets, which can create problems even for experienced, technically proficient users (see Box 6-1).

The majority of NOAA's current data access systems are not optimized for nonexpert users such as students or lawyers who may need data as simple as the temperature at a specific location or time, or who may need more extensive but still user-friendly data sets, or who may not even know what data they need in order to address a particular question or problem. These nonexpert users almost certainly do not know that NOAA archives Geostationary Operational Environmental Satellite (GOES) and NEXt-generation RADar (NEXRAD) data but not CloudSat data, or what data products are produced from the Advanced Very High Resolution Radiometer (AVHRR) and WindSat sensors. However, potential users seeking environmental data should not need to know anything about mission boundaries, organizational structures, or observational campaigns in order to locate the data they require for their application.

Since current data access systems at NOAA meet the needs of many present users, they should be continued. However, efforts to link, consolidate, or expand access to potentially related data sets would improve discoverability and thus yield additional societal benefits. A good starting point would be to create a complete, publicly available inventory of NOAA's data holdings (or, better still, all federal data holdings). The

BOX 6-1
Data Discovery Can Pose a Challenge Even for Experts

To illustrate the need for improved data discovery in NOAA's current data management scheme, one of the members of this committee attempted to obtain pan evaporation data for the nearest station to Norman, Oklahoma, using NOAA Web pages. Pan evaporation data would be of interest to a variety of different users, but this measurement is taken at just a small fraction of NOAA's cooperative observer sites, so this search, while limited in sample size, represents a realistic test of discovery and search capabilities. The committee member, an experienced meteorologist with extensive knowledge of the meteorological observational network and high computer literacy, spent 30 minutes searching for the relevant data through various NOAA Web portals. The committee member found that the steps one needed to take to find this particular data set were not obvious, and he encountered several apparent dead-end links on the National Climatic Data Center's (NCDC's) Web site. One promising pathway, "Find a Station," linked to "NNDC [NOAA National Data Center] Climate Data Online," which provided an option to choose a city. A number of variables, including pan evaporation, were listed when Norman was selected. However, after choosing pan evaporation and the period 1970–2006, the query came back indicating that no data were available for that location. Moreover, there did not appear to be an option at any time during the process to list sites with specific types of data for specific periods so that a substitute site could be chosen. Attempts to access such data through other portals were also unsuccessful. It should be noted, however, that when climatologists who frequently use the NCDC Web site were consulted for advice, they were able to describe a successful procedure for obtaining the data. While the committee member may have eventually discovered this procedure, the inability to successfully locate the intended data in a reasonable amount of time suggests that significant improvements in data discoverability are possible.

current maze of data access entry points should also be better organized and connected as part of an enterprise-wide effort to enhance discoverability for a broad range of users who only want to keep track of a single portal to obtain environmental data. The success of preliminary efforts to create problem-specific Web portals, such as the National Integrated Drought Information System (NIDIS),[4] suggests that further investments in application-specific Web portals would increase the discoverability and utilization of other data that are already in NOAA's archives. Investments in data discovery are among the highest-yielding investments that could be made to existing data access infrastructure at NOAA.

[4]http://wwa.colorado.edu/themes/current_research/water_and_climate_products.html.

Guideline: Search tools and other discovery-enhancing features could be improved at many environmental data access points. To promote data discovery, the structure of the data archive should allow searches to be intuitive, geographically tailored, thematic, and integrated across data set types.

The ubiquity of Web-based search engines has made data discovery a less daunting problem than it once was, but simple keyword searches (for instance, Google) still have limitations. Text-based searches limit searching by space and time, and typically they require the person performing the search to have some knowledge of the relevant discipline-specific terminology. NOAA should explore methods to make their data more readily searchable. These methods might include the creation and implementation of more detailed metadata and the development of formal, machine-understandable, contextual knowledge-encoding mechanisms (or ontologies). Because data are typically integrated (and sub-setted) across time and space, it is especially important to have detailed metadata describing the spatial and temporal attributes of the data. Formal thematic vocabularies, taxonomies, and ontologies can facilitate cross-disciplinary data exploration and automated reasoning. This formal semantic approach, coupled with consistently structured data, can aid sophisticated data mining or advanced search techniques.

While making good use of external search mechanisms, NOAA should also strive to improve the visibility and effectiveness of its own data access mechanisms, especially on the World Wide Web. To facilitate broad interdisciplinary discovery, these search and access mechanisms should be interoperable and linked across NOAA, where possible. Formal user-driven interface and software design techniques should be used to improve their effectiveness and intuitiveness. For example, user cases or scenarios (remember the AVISO example noted in the previous section) could be developed to make sure that the tools are designed to meet specific user needs, while formal usability testing can help ensure that they are intuitive and functional. In addition, users may want other information related to a particular data set such as data quality attributes, activity logs, and user feedback. This sort of information can be very valuable to a user assessing the appropriateness of a data set for a specific application, as discussed in additional detail in the next guideline.

Guideline: Data discovery would be improved by the use of expanded metadata.

The importance of metadata, or the additional information about a data set that allows the data to be independently understood and used,

was discussed in several previous chapters. The close relationship between metadata and data stewardship and the importance of metadata standards were discussed in Chapter 4, and the role played by metadata in more effective data archiving was noted in Chapter 5. Well-organized, standardized, and managed metadata benefits both users (through improved discovery, access, and integration) and data managers (by encouraging efficient stewardship and archiving). As discussed earlier in this chapter, metadata can also enable machine-to-machine and application interoperability. An additional, especially important application of metadata is to promote the discovery, access, and integration of different environmental data streams for a variety of users.

The spectrum of useful metadata is large and diverse. Metadata are discipline dependent and invariably expand as a data set matures, so they should always be managed in flexible systems to accommodate their expansion and improvement. Some discovery metadata are most effective if they are internally part of the data file structure (for example, netCDF formatted data), while other metadata typically reside in separate files or in databases associated with the data archive. The best systems tie all data and metadata together so that users can assess their options.

NOAA is progressing toward, and in many cases meeting the requirements for, providing standard metadata to the current recommended national and international levels. Although this is a successful trend and an accomplishment, it should not be viewed as a final end point. Many user benefits could be realized by extending metadata management and provision beyond the minimum requirements. What follows is a nonexhaustive list of metadata that is meant to be illustrative. It begins with some very basic metadata and ends with examples of ancillary metadata that could be useful for placing data in the correct context, fostering efficient discovery, and providing improved user experiences.

- A distinctive dataset title
- A general description or summary of the data set
- Definition of parameters, variables, and physical units
- Observation or grid location, period of record, spatial and temporal resolutions, and update frequency
- Description of methods and procedures on how the data were derived, including calibrations, error estimates, processing, etc.
- Relevant information about the data set size, format, where it is located (this could be at multiple sites)
- The types of data services available (GIS, OPeNDAP, FTP, etc.)
- Reference authority and background information: who created the data, and for what purpose; associated publications and technical reports

- Directives, where possible, about appropriate and inappropriate usage
- Historical lineage about the data set version and what caused the evolution, including algorithm changes and reprocessing rationale
- Connections to related data sets either at higher or lower processing levels, within the same discipline, and across disciplines using relationships (semantic metadata) that can logically map between different terminologies and enable broad automated discovery
- Pointers to useful tools for analyzing the data, such as scripts for reading the data or applications to manipulate and display the data
- Online environments to capture and share user logs

The importance of user log information was also noted in Chapter 4, but some ideas are worth reiterating here because this information could be considered as another form of extended metadata. To the maximum extent possible, NOAA is encouraged to create and maintain open access to user feedback logs and make them part of the metadata associated with each data set. In addition to being valuable to the experts/stewards responsible for data set maintenance, assessment, and improvement, feedback logs would provide valuable information to users about the quality, utility, and appropriate applications for the data. User feedback can also provide information about the current and potential future uses of each data set and the relationship between the data and data sets at other archives. NOAA personnel could also monitor the logs for feedback and error analysis, and could take advantage of user comments and suggestions to improve the user-friendliness of data access protocols. Capturing user feedback in logs and incorporating these logs into standard metadata would thus yield benefits in data stewardship, discovery, access, and integration.

A critical concern for NOAA is assuring that at the end of the data discovery exercise, a user has been made aware of all relevant and helpful data offered by the agency. This is not an easy task. Nevertheless, the use of standardized metadata and deployment of systems that can collect, share, catalog, and publish metadata from the many diverse NOAA centers of data and National Data Centers should make it possible to more effectively support data discovery. A metric that quantifies users' perceived success in discovering what they need would also be a powerful indicator that, over time, could be used to further improve NOAA's data access systems.

BARRIERS TO ACCESS

Guideline: Environmental data should be made available to users in a timely manner and should be accessible with as few barriers as possible.

Reducing the barriers on data access and use is essential in advancing scientific and societal goals. Despite the huge volume of data accessed at NOAA every month (see Figure 2-1, for example), millions of other potential users have never requested or gained access to data from a NOAA data archive. Many of these users probably do not even know about the broad range of environmental data that might be available to address their needs, while others may have encountered obstacles while trying to access NOAA data. Enhancing data discovery, providing easier data access, and promoting more extensive usage of archived data are all steps that would significantly increase the realized value of NOAA's data holdings.

The most challenging and difficult barriers are not always technical in nature; rather, they involve the socioeconomic, legal, institutional, and political dimensions of data access. These barriers individually and collectively limit full utilization of NOAA's data assets and impede progress on many fronts, including education and research, creation of new knowledge, as well as commercial and societal applications. In this section, we focus on users who are aware of the data they could potentially use and examine the various barriers that might prevent them from easily accessing the data that interests them. NOAA should address these barriers, which may be further classified into administrative, technical, and systematic barriers, and remove or at least reduce them wherever possible.

Administrative Barriers

Authorization and Security: More sophisticated access control procedures should be implemented using "smart" authorization or authentication methods. For example, in some cases cost-free access to NOAA data is available to academic users for education and research purposes. However, simple approaches, such as restricting access to certain users based on their Internet protocol (IP) addresses, may inadvertently exclude qualified users. In general, any registration or login requirement will discourage data usage. On the other hand, there are some data for which access restrictions are clearly needed, such as the location of rare in situ specimens or data with national security implications. Some access constraints may also be needed to protect data archives from unreasonable data

requests (either intentional or inadvertent) that would result in denial of service to other users. In general, however, our view is that access controls should be kept to a minimum.

Proprietary: Unavoidably, some of the data sets NOAA will want to archive are proprietary in nature, particularly data derived from international and/or commercial sources, and NOAA will not be able to provide full, free, and open access to all of them. Therefore, there should be provisions in the data management system for incorporating such proprietary data, as well as guidelines on the use of such data. In general, these proprietary barriers to access should be kept to a minimum and should exist only where they are absolutely necessary.

International Policy Issues: World Meteorological Organization (WMO) Resolution 40 ("WMO policy and practice for the exchange of meteorological and related data and products including guidelines on relationships in commercial meteorological activities"[5]) expressly states, as a fundamental principle, that WMO is committed to free and unrestricted exchange of meteorological and related data and products for education and research purposes. However, the resolution also provides guidance restricting reexport of some data for commercial use. Such a policy, while not draconian, does place certain barriers on the broad use of NOAA data holdings.

Full and Open Access: Since the advent of the Internet, full and open access to all types of digital information, including environmental data, has been the subject of considerable discussion among a variety of stakeholders and interests, including educators, researchers, librarians, publishers, sponsoring agencies, commercial enterprises, and government officials. Many previous reports (NRC, 1995c; NRC, 1997; NRC, 1999; NRC, 2001; etc.) have discussed the benefits of full and open exchange of data, and there are already a number of disciplines with large data compilations that are freely and openly available online, such as the Sloan Digital Sky Survey, the National Virtual Observatory, and the National Institute of Health's GenBank. NRC (1997) offers the following recommendation regarding the importance of full and open access:

> *The value of data lies in their use. Full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from publicly funded research. The public-good interests in the full and open access to and use of scientific data need to be balanced against the legitimate concerns for the protection of national security, individual privacy, and intellectual property.*

[5]http://www.nws.noaa.gov/im/wmor40.htm.

In the view of this committee, and in accordance with applicable U.S. law and policies,[6] full and open access to data should be a fundamental tenet of all US federal agencies, including NOAA. Consistent with other definitions (WMO Resolution 40 and NRC, 1997 are two examples among many), "full and open" should be taken to mean nondiscriminatory, without restrictions, and without additional charges beyond the cost of reproduction and delivery of the data and products themselves. The provision of full and unrestricted (or minimally restricted) access can greatly enhance the utility of NOAA data holdings: it advances not only scientific discovery but also goals in the realms of commerce and policy making, among other benefits to society. Full and open access policies also have the ability to catalyze the development of new methods of analysis and integration, to result in increased collaborations, to reduce inefficiencies, and to accelerate the production of new knowledge.

Technical Barriers

Digital Divide and Bandwidth Limitations: In the simplest terms, digital divide refers to the gap between those with effective access to information technology and those without such access. This division encompasses both physical access to technology hardware, such as computers and networks, and, more broadly, the skills and resources that allow for its use. The digital divide limits or excludes certain segments of society, especially underrepresented groups, from benefiting from NOAA's data efforts. For the broadest and most effective use of NOAA data, steps are needed to reduce the gap and to devise solutions for overcoming such hurdles.

As a specific example, not all users or potential users have access to high-speed Internet connections. To address this barrier to access, NOAA and its partners could consider approaches like progressive disclosure, in which a small amount of data is initially provided and the user can request progressively more data as their network resources permit. One simple way to achieve this would be with "sub-setting/decimation" routines, which allow users to access and download only the data they need (such as a certain spatial area, time period, and/or resolution). In fact these routines should be considered for widespread use because they promote more efficient data access, especially for large data sets.

Standardized Protocols and Conventions: The lack of standardized protocols and conventions can be a formidable barrier in the interoperability and integration of diverse data sets from different disciplines. For

[6]For example, White House Office of Management and Budget (OMB) Circular A-130, as revised December 2000.

example, Internet usage grew explosively when the hyper-text transfer protocol (http) became the foundation of the Web in the 1990s. Similarly, the use of common protocols and conventions for data access and formats can lead to more effective and broader utilization of NOAA's environmental data. As an example, many teachers in the K-12 community routinely use spreadsheet applications such as Excel, but these same professionals may be less familiar with or reluctant to use the more specialized scientific application tools favored by educators and researchers in higher education. To facilitate all users, NOAA data systems should provide access via standard protocols recognized by significant user communities.

Latency: NOAA needs to recognize that data latency is not just an operational consideration. For example, certain research communities need near real-time access to data in order to initiate the collection of correlative data during extreme events (severe weather, volcanic eruptions, earthquakes, etc.). In some cases data acquisition systems have variable collection frequencies that need to be adjusted during such events. In general, environmental data should be made available to users as soon as possible. If there are situations in which data need to be withheld (for homeland security reasons, for instance), there should be a clear rationale for the withholding period, as well as a clear mechanism for qualified users to apply for the ability to obtain the data if needed for a certain time-critical application.

Systemic Barriers

Competency of Users: Data systems that are not tailored to individual levels of user sophistication can pose a significant barrier to the utilization of data holdings in those systems. NOAA should provide multiple levels of entry to data portals at a variety of levels of user sophistication, both in terms of customer knowledge and skills (scientists, K-12 educators, the general public, decision makers, and applied or multidisciplinary users) and in technical capabilities. A multi-portal system allows for efficient access by experts, carefully explained and supervised access for beginners, and, possibly, several levels in between. For example, extensive use of discipline-specific terminology and acronym-laden data set names in data access systems should be avoided in portals designed for a wide audience.

Disability Considerations: In 1998, the U.S. Congress amended the 1973 Rehabilitation Act[7] to require federal agencies to make their electronic and information technology accessible to people with disabilities. Inaccessible technology poses barriers for certain individuals to obtain and use

[7]29 USC Sec. 793.

information quickly and easily. Section 508 was enacted to eliminate such barriers in information technology, to make available new opportunities for people with disabilities, and to encourage the development of technologies that will help achieve these goals. As a federal agency, NOAA is required to comply with this regulation and to eliminate such barriers to the use of its data information technology systems.

Incomplete Metadata: As discussed previously in this chapter and in earlier chapters, full and complete metadata are essential to maximizing data discoverability, accessibility, and usage. Thus, incomplete metadata can be a barrier to complete and effective data access. For example, data sets that lack adequate metadata can be hard to locate, difficult to use, and difficult to integrate with other data sets. This can lead to inefficiencies and, ultimately, unrealized societal benefits. NOAA should reduce such barriers by including the appropriate metadata required for the most effective use of its data holdings.

INTEGRATION

Guideline: NOAA's data access system should be designed to allow the integrated exploitation of data from multiple sources within and outside of NOAA to answer environmental questions and support NOAA's mission.

Most environmental problems cannot be addressed using just a single piece or type of information. Thus, in addition to promoting the discovery and access of environmental data, NOAA's data management system should allow for and facilitate the integration of data and other information. This integration may entail the sequential or simultaneous access of original data from a variety of different platforms, such as satellites, ground sensors, and buoys, or it may be facilitated by Web-based portals that assemble related informational products from different data centers at a single location or specified time.

While simple in concept, data integration, like many other aspects of data access, is notoriously difficult to implement. For example, many of the questions currently being addressed by environmental stakeholders require data that are housed at other agencies and international groups. Cooperation with these entities is essential for ensuring that user needs are met (yet another manifestation of the overarching principle that "effective interagency and international cooperation is needed"). The ability to integrate data sets—either to determine new relationships or to create a new, value-added data set—is often critically dependent on the extended metadata discussed previously. Users with multidisciplinary skills can

BOX 6-2
Educational Data Access and Integration

This example raises several important issues about data access and integration. TERC, an education research and development organization, has developed an educational resource called the Earth Exploration Toolbook.[a] One of the chapters (activities) in the toolbook, "Investigating the Precipitation-Streamflow Relationship,"[b] shows teachers how to obtain precipitation data from NCDC and stream-flow data from the U.S. Geological Survey, import these data into an Excel spreadsheet, and then create graphs and other analysis tools to investigate how precipitation events impact stream flow. The chapter uses climate data from Boston and stream-flow data from the Sudbury River as an example, but it shows how data from any geographic region could be obtained. What makes this exercise so valuable to students is the ability to explore the data and thus gain insight into stream-flow dynamics, such as the variation in stream flow over the course of the year and the temporal correlation between rainfall and stream flow.

According to TERC staff, educators at all levels find the chapter useful, but teachers and students always want to access data for their town or region (education, like politics, is inherently local). In terms of accessibility, the data should be in a format that can be readily used with the tools that the teachers and students have available. The data also need to have adequate quality control to guarantee a reasonably accurate result, but not so highly processed that the teachers and students do not have the ability to manipulate the data. Finally, this example illustrates the societal benefits that can be realized through data discovery, access, and integration; however, the fact that the NCDC and USGS Web sites must be accessed through two different portals also points to the improvements that could be made in interagency coordination for data management.

---

[a]http://www.terc.edu/work/147.html.
[b]http://serc.carleton.edu/eet/module_discharge/index.html.

help data managers identify related data sets that could be included at a given data portal. Box 6-2 illustrates the value of integrated data access capabilities, as well as some of the difficulties involved in providing these capabilities.

Guideline: Practical considerations require a distributed data system architecture and access infrastructure for environmental data.

Ideally, all environmentally relevant data would be easily discoverable and readily accessible from a single portal or access point that facilitates the integration of multiple data sources and meets the needs of all

users. In reality, the diversity, complexity, and volume of environmental data, coupled with the wide range of user communities and applications of these data, necessitate having a tightly integrated data management strategy that is more broadly distributed. It is thus not surprising that a large number of different data access points have been implemented across NOAA, as well as at other agencies, each catering to a different user base, type of data, level of expertise, or application. For example, the Comprehensive Large-Array [data] Stewardship Systems (CLASS) Web interface provides direct access to large-array satellite data, while each of NOAA's three National Data Centers and many NOAA centers of data have portals designed to provide access to popular subsets of data by a broader user community. For example, the academic community sometimes uses the National Geophysical Data Center's Defense Meteorological Satellite Program data for special studies such as auroral analysis.

As shown, for example, in Figure 2-1, many users avail themselves of NOAA's current data access systems. Therefore, these systems should be preserved. There are also a number of prototype projects in various stages of development that enhance data discovery, access, and integration for specific applications or user groups (the National Integrated Drought Information System, or NIDIS, is one such project; see Box 6-3). However, all aspects of data access would be improved if there were an enterprise-wide effort to increase the linkages between different portals, to improve and expand the search functions offered at individual portals, and to design new portals that focus on new applications. The nature and justification for these steps are described in more detail in the next guideline, and the next chapter describes how a system-of-systems approach to data management could provide the distributed access infrastructure demanded by NOAA's diverse data holdings and user groups.

Guideline: A distributed data access infrastructure can and should support improved data discovery and seamless integration.

In general, all aspects of data access would be dramatically improved if an enterprise-wide effort were made to increase the linkages between different portals and to improve and expand the search functions offered at individual portals. There have also been some important preliminary efforts to consolidate access to a variety of data sources at portals focused on a particular interdisciplinary problem, such as NIDIS (Box 6-3). These problem-specific portals represent a particularly effective way to improve the relevance of NOAA data for a wide variety of environmental problems and decisions and should be considered a ripe area for further system improvements. However, a truly integrated archive requires a systematic approach that begins with the conceptual design of the data management

system itself. As discussed in Chapter 7, many of these goals could be achieved using the Global Earth Observation Integrated Data Environment (GEO-IDE) or a similar approach.

BOX 6-3
Integrated Information Resources for Drought Assessment on All Time Scales

When assessing and addressing issues related to drought, it is critical to have a diverse assemblage of information that bridges multiple temporal and spatial scales. Of greatest immediacy for decision makers is an evaluation of recent and current conditions on time scales ranging from daily to seasonal or even interannual. In addition to specific, drought-related indexes such as the U.S. Drought Monitor and U.S. Forest Service Fire Danger Rating, other variables of interest may include precipitation, snow pack water content, surface air temperature, humidity, evapotranspiration, soil moisture, stream flow, groundwater levels, reservoir levels, and foliage or range conditions. Because absolute values and anomalies of all variables are important, all these data may need to be accessible in tabular, graphic map, and narrative formats. Conditions specific to fish and wildlife populations or knowledge of certain climatological states (El Niño-Southern Oscillation, or ENSO, for instance) are also valuable assessment tools. Observational metadata and other ancillary information such as water restrictions, passing flow requirements in regional rivers, water usage statistics, or census figures can also be critical for making effective decisions. It is also important to place present conditions in historic perspective, which might require both instrumental records and paleoclimatic or proxy records to estimate past periods of drought, as well as the relationship between regional precipitation and other indexes.

Once the present situation is well understood, the next step is to attempt to gain some insight as to whether drought conditions might worsen or ameliorate. Thus, forecasts for atmospheric and associated surface conditions need to be accessible to the decision maker, again on daily to seasonal time steps. These forecasts can be variable specific (precipitation) or for an integrated assemblage of information (drought). For those looking well beyond the present (years and decades ahead), model projections would need to be available at appropriate spatial and temporal scales. As with current conditions, outlooks need to be provided to decision makers in formats that permit rather straightforward interpretation. This includes sufficient information regarding probabilities and confidence limits.

One-stop shopping is a useful goal as long as the data and information are discoverable and understandable. With respect to drought, work is under way through the National Integrated Drought Information System (NIDIS) program to develop a Web portal that will provide a link to connect the data, scientists, and decision makers.[a] It is hoped that this portal, as it develops, will serve as a model for other meteorological and climatological phenomena.

[a]http://wwa.colorado.edu/resources/nidis/.

7

Integrated Data Management at NOAA

PRINCIPLE #9: Effective data management requires a formal, ongoing planning process.

NOAA should establish and codify an enterprise-wide data management plan that explicitly incorporates all the principles in this report and is in alignment with NOAA's mission, vision, and goal statements. The plan should be flexible and needs to include a formal, ongoing planning process—developed and integrated across NOAA with substantial user involvement and coordination with other agencies—to ensure that the system keeps pace with changes in observing systems, models, data storage and access technologies, scientific understanding, and user needs, as well as available resources.

NOAA has responsibility for a vast and growing collection of environmental data sets, some of which are critical for supporting environmental decisions, while others may be of limited value. Perhaps the biggest deficiency in NOAA's current data management enterprise is the lack of a formal decision-making process for assigning specific archiving, stewardship, and access responsibilities for individual data sets. An additional, related concern is the absence of a clear, enterprise-wide framework to connect its many disparate data management activities. NOAA's current Strategic Plan mentions data management in its crosscutting priorities section, but it does not explicitly acknowledge the data management functions needed to support NOAA's mission or the challenges associated with meeting its ever-expanding data management responsibilities.

Thus far NOAA has been able to rely on ad hoc groups and the dedication of its personnel to provide reliable archiving and stewardship of essential data and to develop data access capabilities for specific user communities. For example, data managers at individual NOAA National Data Centers and centers of data have assumed many of the decision-making responsibilities for individual data sets. However, the recent and expected future increases in data volumes and complexity necessitate a highly coordinated, more defined process for managing the nation's environmental data, as well as a comprehensive, user-centric framework to guide the ongoing development of NOAA's data management activities.

The principles and guidelines in this report provide not only guidance for determining which data sets to retain and provide access to, but also a foundation on which to build a comprehensive data management plan. The next section presents one possible framework for data management at NOAA—based on a "system-of-systems" concept—along with some specific steps that describe how this framework could be implemented to take advantage of and extend existing data archiving, stewardship, and access capabilities. However, it should be noted that the principles and guidelines in previous chapters would remain applicable under any number of alternative data management frameworks. Also included at the end of this chapter are some concluding thoughts regarding the need for, and potential benefits of, a comprehensive, integrated, inclusive, and ongoing data management planning process across NOAA.

VISION AND FRAMEWORK

NOAA has already established a vision, illustrated in Figure 7-1, for an integrated, enterprise-wide data management system. The fundamental attribute of this system is the ability to link disparate data sources with the interdisciplinary applications in support of NOAA's mission. However, this is an extremely difficult and complex task, made all the more challenging by the volume, complexity, and diversity of NOAA's many data sets, the myriad needs of its diverse users, the constraints of limited resources, and the underlying need to preserve and support current data management and user support activities. It should also be noted that the connection between data and mission objectives ultimately depends on scientific analysis and interpretation of data, not just hardware and software. Thus, the challenge is to define an effective framework that fills in the details of the broad and amorphous "integrated data management system" at the center of Figure 7-1.

In addition to fulfilling the basic vision of connecting environmental data to NOAA's mission goals, the integrated data management system should be designed to fully address the main functional elements (stew-

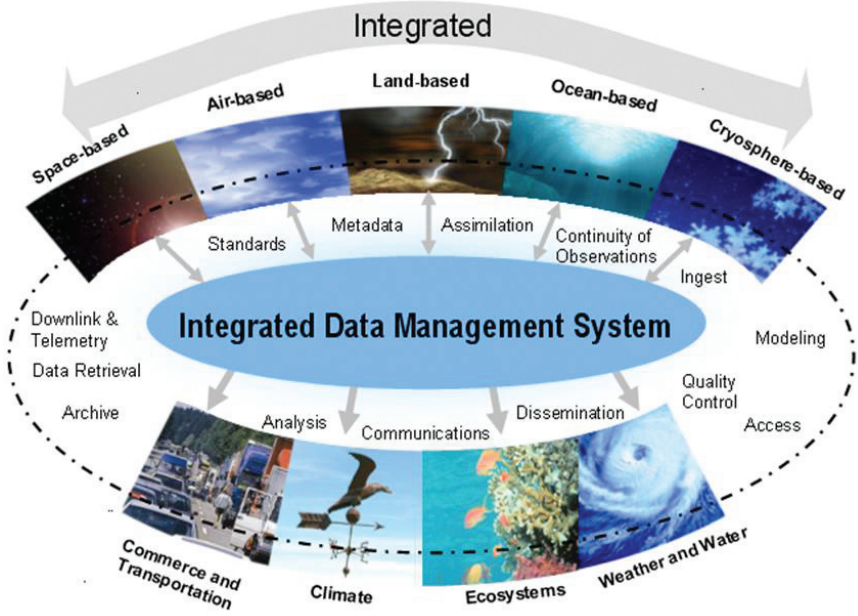

FIGURE 7-1 NOAA's Observation System: Target Architecture (SOURCE: Adang, 2006.)

ardship, archiving, and access) described in Chapters 4, 5, and 6, as well as the overarching data management principles discussed in Chapter 3. In addition, the system will also need to be designed with sufficient flexibility, scalability, and adaptability to ensure that user needs and legislative and administrative mandates all continue to be met, despite growing data volumes and rapidly evolving user requirements. The elements of a continuously evolving and expanding data system include:

- stewardship of past, present, and future data, the value of which is likely to change over time;
- scalability to accommodate the demands of both existing and future user communities;
- flexibility to incorporate new data sources as they become available;
- adaptability to take advantage of transformational changes in data system technology;

- extensibility to incorporate smooth and nondisruptive upgrades for hardware, software, and data; and
- an optimal and cost-effective combination of technologies that maximally leverages off-the-shelf components with specialized components developed in-house to meet specific NOAA needs.

A number of practical considerations also need to be taken into account when designing and implementing a comprehensive, integrated, enterprise-wide end-to-end environmental data management system. For example, such a system requires persistent (or sustained) facilities, hardware, software, support staff, and, as discussed in Chapter 3, a dedicated, ongoing funding process that supports these infrastructural components. A more structured and inclusive process for user input is necessary, and acceptance and "buy in" by data providers as well as users will be necessary to make sure that the data management system meets both current and future user needs. Finally, NOAA needs a strategy for prioritizing its data management activities to focus resources on the most important management issues and thus maximize cost-effectiveness. Collectively, these practical considerations demand an organizational structure for data management that is reliable, flexible, adaptable, scalable, cost-effective, and responsive to user needs; the nature of this organizational structure is the main focus of the next section.

A System-of-Systems for Data Management

The framework for NOAA's data management system should ultimately be driven by NOAA's mission objectives, with the system design and components based on existing infrastructure and plans, available resources, and, most importantly, the data archiving, data stewardship, and data access services required by its evolving set of user communities. An important lesson learned from the development of data management systems at other agencies (NASA's Earth Observing System Data and Information System [EOSDIS], among others) is that a combination of top-down and bottom-up approaches is necessary to iterate to an effective data management solution, and NOAA would also be well advised to build upon the legacy of many years of successful support to its primary user communities. A federated system, unified by a common vision from the top and building upon small successful discipline- and instrument-focused systems, is a cost-effective design that embraces exactly this approach.

In light of these considerations, it is advisable for NOAA to continue moving toward a federated but integrated "system-of-systems" to manage its environmental data. Such an approach would leverage the many

successful data management activities that NOAA is currently engaged in, facilitate the inclusion of users and other stakeholders in the planning process, and provide the top-level support and strategic guidance required to ensure adequate coordination and support while improving integration and cost-effectiveness. This approach would also mirror and support the evolution toward a system-of-systems concept in the observational network (for example, the Global Earth Observing System of Systems, or GEOSS), and it is highly compatible with the Global Earth Observation Integrated Data Environment (GEO-IDE) concept discussed in Chapter 2.

A schematic diagram of an integrated data management system that illustrates and defines some of the components of the integrated system-of-systems concept is depicted in Figure 7-2. This framework for NOAA's data management enterprise, which combines the legacy discipline-specific National Data Centers, centers of data, and data portals with a central data archive and interconnected multi-portal data access, provides the fundamental mechanisms required to connect diverse communities of users with NOAA's broad spectrum of environmental data. The central data archive and other data archives would focus on data storage, data integrity, standards and formats, emerging technologies, and the associated research and development, while NOAA's three National Data Centers and centers of data would focus on data acquisition, data stewardship, quality control, and data access. It should be noted that the data centers would still maintain their existing data portals; however, as indicated by the interconnected multi-portal access block at the center of Figure 7-2, these portals should be better connected—with improved linkages to related data sets at other centers—to facilitate better utilization and integration of existing data. Seamlessly interconnected, multi-portal access is the most essential component of the framework illustrated in Figure 7-2 because it "connects the stovepipes" present in current data management activities and thus allows improved data discovery, integration, and interoperability.

There are several motivations for recommending that NOAA build upon existing decentralized, discipline-specific systems rather than replace them. NOAA has a proven track record of providing fundamental data records and supporting the data archive and access needs of its existing user communities. It is also apparent from past experiences with archive system development at both NOAA and NASA that predominantly top-down, centralized, technology-driven system designs have deficiencies and are limiting, largely because they are inflexible and fail to recognize and accommodate the extremely wide variety of data characteristics and user needs. These considerations strongly suggest that NOAA's data management vision should be grounded on a bottom-up approach

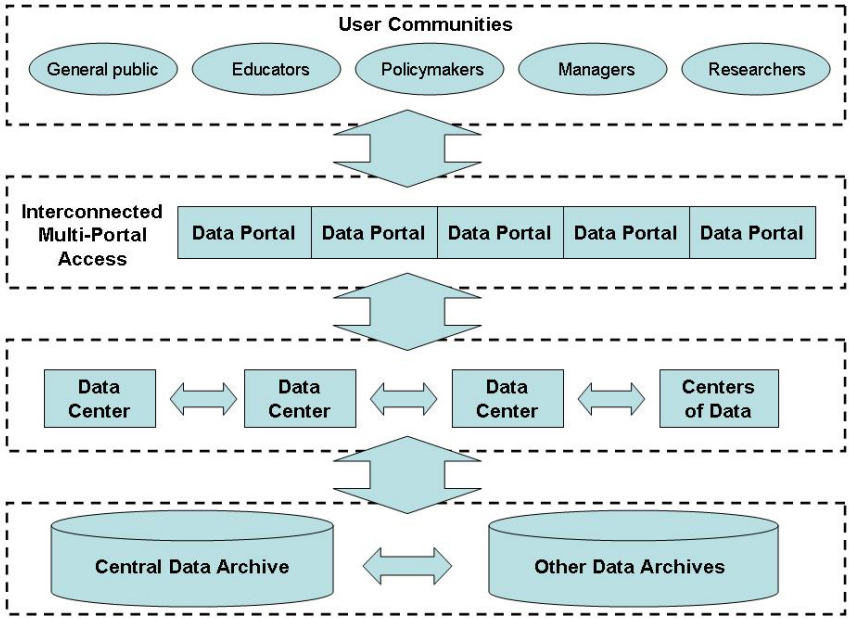

FIGURE 7-2 Elements of an integrated data management system that connects users to environmental data via an interconnected multi-portal access framework. The "interconnected multi-portal access" includes search engine and automated data mining capabilities, as well as access to multidisciplinary advisory support. The National Data Centers and centers of data are regarded as "disciplinary" data centers that are responsible for both user support and data stewardship. The central data archive and other data archives are responsible for central data storage and preservation, as well as the research and development of emerging technologies, standards, and protocols. Archiving responsibilities for individual data sets may be assigned to any two of the centers in the lower two dashed boxes to guarantee survivability. The double-headed arrows indicate generalized connections between different elements of the system, illustrating the flow of data as well as data support services. Links to other data systems are not shown.

that leverages existing archive, stewardship, and access capabilities and facilities, as well as direct and ongoing input from different user communities. However, it is equally clear that enterprise-wide coordination in the planning, implementation, and continual refinements of the eventual system will be needed to ensure adequate interoperability and to make sure that the needs of users and stakeholders are met.

Implementation

The system-of-systems framework for integrated data management at NOAA is not radically different from either the GEO-IDE data management plan or the broader GEOSS data architecture, both of which are currently still in the development phase, although it does represent a major departure from the current implementation of data systems within NOAA. The existing legacy systems at NOAA, with decentralized, discipline-specific support, have actually been working quite well despite limited resources, especially in supporting discipline-specific questions from a range of users. However, they could and should be expanded and linked more effectively to meet the challenges associated with more complicated, multidisciplinary questions, to increase the realized value of collected data sets, and to make the overall system more cost-effective.

GEO-IDE, which was discussed in detail in Chapter 2, provides an excellent starting point from which to develop an integrated data management system-of-systems at NOAA. Indeed, the draft concept of operations and implementation plan (NOAA, 2006b, 2006c) describes GEO-IDE as an enterprise-wide system that stresses improved integration and accessibility, as well as the use of standards, user input, and a federated approach to link diverse data streams (see Figure 2-3). In particular, the implementation plan calls for NOAA to:

- continue operating existing systems, but gradually evolve them to be more interoperable, reliable, and scalable;
- create cross-organizational teams to find solutions to focused problems specified by higher management and to implement agreed-upon adoption of solutions;
- define and implement a process for adopting information management standards, and ensure that legacy programs migrate to these standards and new programs use them from the start;
- develop, publish, and maintain a NOAA Guide to Integrated Information Management; and
- invest in education to develop necessary new leadership, management, and technical skills for the integrated data environment.

In addition, the GEO-IDE implementation plan calls for NOAA's Line Offices, across the whole organization, to participate in well-ordered, standards-based data infrastructure to improve efficiency and promote easy data discovery and access. These and other concepts in the GEO-IDE planning documents are consistent with the principles and guidelines offered in this report, and they provide an excellent foundation for planning additional improvements.

However, as discussed in Chapter 2, GEO-IDE remains very much in the planning phase, with only limited resources dedicated to its further development and almost none for its eventual implementation. In addition, the GEO-IDE planning documents will need to be expanded to more completely address several key issues, such as data stewardship responsibilities, interagency cooperation, and user input, in order to conform to all the principles and guidelines offered in this report. Specific examples of missing elements include: an inventory of existing data and a structured process to review and determine the status of each environmental data set at key points during its life cycle; a clear mechanism or mechanisms for soliciting and incorporating user input into data archiving and data access decisions; a prioritization process for applying limited resources to various data management activities; and mechanisms to improve data discoverability and accessibility. If these and other changes are made in order to bring the GEO-IDE concept into better alignment with the principles and guidelines in this report, and sufficient resources can be made available for its implementation, the resulting data management system will yield significant benefits. Some examples of how an integrated, coordinated, and user-driven data management system at NOAA would make these benefits possible are given in Boxes 7-1 and 7-2.

NOAA is also encouraged to continue exploring new technologies and alternative management structures in an ongoing effort to improve capacity, reliability, and cost-effectiveness. Data integration and exploitation through advances in information technology (such as computing power, networking capabilities, and data storage volumes) have the potential to accelerate scientific progress over the next several decades in much the same way that modeling and simulation studies have over the previous two decades. NOAA will need not only to plan for the near- and intermediate-term requirements of these technological advances, which will define their initial system but also to anticipate as yet undefined future possibilities and corresponding user needs. NOAA also needs to establish a significant priority, within the current development and future planning, for improving data discovery and integration capabilities. A good starting point would be creating a complete, searchable inventory of existing data holdings, together with the development of more intuitive portals that provide more direct discovery and access of all environment data by different user groups, including interdisciplinary scientists, K-12 professionals, and the general public.

Finally, it should be reiterated that the specific framework for data management depicted in this section is intended to assist NOAA in the planning, funding, and implementation of a comprehensive data management system that is compatible with the principles and guidelines offered in previous chapters. The proposed system-of-systems architec-

BOX 7-1
Addressing Interdisciplinary Problems

To illustrate how an integrated data management system needs to work in order to address inherently interdisciplinary problems, consider a researcher who wishes to study the decline in a particular fish species in the eastern Pacific Ocean. Even if the researcher were aware of all the data they might need to study this problem, it would currently be quite difficult for them to easily find and obtain the most appropriate fish catch data, population model results, *in situ* sea surface temperature and salinity measurements, satellite-derived ocean color data, winds and currents from reanalyses, and other potentially important variables such as cloud cover and aerosol concentrations. A truly integrated data management system would facilitate the researchers' ability to identify and obtain all potentially relevant data sets, including those that they might not have thought of as initially being relevant.

Expanded metadata standards and mappings among discipline-specific standards are critical to this integration and would be a good starting point for any planned coordination or integration activities within NOAA. Logical presentation of the information during discovery is also important. In this particular case, the researcher might need assistance in selecting the appropriate satellite data; for example, would orbital swath, gridded multiday averages, or seasonal mean ocean color, be most appropriate? This example illustrates both how advances in our understanding of the Earth System often require the merging and analysis of disparate data sets and how an integrated data management vision can be used to answer complex interdisciplinary questions.

BOX 7-2
Addressing Novel Applied Problems

Derelict fishing gear and other marine debris pose increasing hazards to commercial and recreational navigation, entanglement of endangered and protected species, and wasteful "ghost fishing." To mitigate the potential damage, NOAA has funded several efforts that use near real-time data from ocean transport models, satellites, piloted and autonomous aircraft, and drifting buoys to guide ships to interdict the debris before it reaches land. The capacity to easily integrate the data from the diverse suite of platforms has proven critical in the efforts of the program managers to allocate resources in an efficient and effective manner. These efforts illustrate the need for coordination and flexibility in designing a data management system that can be applied toward novel applied problems.

ture provides avenues for the complementary top-down and bottom-up development approaches necessary to ensure that the system is integrated and coordinated across the agency, is cost-effective, and is sufficiently flexible and adaptable to respond effectively to increasing data volumes and evolving user requirements. Should NOAA decide that an alternative vision or approach is necessary, however, the principles and guidelines in this report are general enough that they may be applied to alternative data management frameworks. Of course, for any significant progress to be realized, NOAA needs to receive secure funding to establish and maintain these important data management activities.

CONCLUDING THOUGHTS

NOAA has the legacy systems, basic organizational structure, several successful prototype user support systems, and infrastructure plans (namely, GEO-IDE and the Comprehensive Large-Array [data] Stewardship System, or CLASS) to meet many of its diverse user needs. The discipline-specific NOAA National Data Centers and centers of data already provide excellent user support and continue to develop new methods and practices in response to user needs. However, the expanding demands and diversity of users, coupled with an explosive increase in data volumes, represent a formidable data management challenge. Although technological improvement will help, the resources currently allocated for data management across the agency may not be sufficient to ensure that NOAA continues to meet its basic data archiving and access requirements, and further resources may be needed to improve the discoverability, accessibility, and integration of different data sets needed to address important interdisciplinary problems.

In light of these considerations, some modifications to NOAA's current plans and resource allocations will be needed to address all of the key issues involved in providing the focused user support demanded by NOAA's mission. The principles and guidelines in this report are intended to provide a foundation for the design, implementation, and operation of an integrated, forward-looking data management system that meets or exceeds current user requirements and responds to new requirements as they emerge. The three main functional requirements of such a system—namely, data stewardship, data archiving, and data access—were explored in Chapters 4, 5, and 6, respectively. Chapter 3 provided five overarching principles that address the motivation, financial resources, interagency and international coordination, metadata requirements, and responsiveness to user needs that are necessary for truly effective data management. Finally, in this chapter we described an integrated, adaptable, user-driven "system-of-systems" that is compatible with these

principles and guidelines; that leverages NOAA's existing data archiving and access infrastructure and expertise; that emphasizes reliability, cost-effectiveness, and broad stakeholder involvement in data management decisions; and that maximizes the societal benefit of existing and planned observational capabilities.

Understanding and predicting environmental change will remain critically important activities throughout the 21st century. These activities depend on historical data and reliable projections of the future evolution of the Earth System, both of which rely on accurate descriptions of current environmental conditions. Hence, it is critical to archive and provide access to the environmental data needed to describe, understand, and predict changes in the Earth System and the impacts of these changes on both natural and human systems. NOAA and its partners are clearly dedicated to continue providing excellent service for their various user communities and to continue meeting their legal mandates and requirements. However, these goals can be achieved only if NOAA's data management system enables and promotes the discovery, access, and integration of a wide variety of environmental data by a broad range of users. With such abilities, users are empowered to reap the many societal benefits of environmental data and to advance our understanding of the Earth's environment. NOAA and its partners will, we hope, find the guidance offered in this report useful as they continue to serve as effective stewards of important national assets.

References

Adang, T. 2006. Global Earth Observation Integrated Data Environment (GEO-IDE). Presentation to the 2006 CLASS Users Workshop, August 7.

Alexandrou, M. 2007. Technology Definitions. Available online at http://www.mariosalexandrou.com/definition/index.asp (accessed October 16, 2007).

ASCE (American Society of Civil Engineers). 2006. Minimum Design Loads for Buildings and Other Structures. ASCE Standard No. 7-05. Reston, VA: ASCE Publications.

Asrar, G. 1998. Statement of Dr. Ghassem R. Asrar, Administrator National Aeronautics and Space Administration, before the Subcommittee on Space and Aeronautics Committee on Science, House of Representatives, September 10, 1998. Available online at http://icesat.gsfc.nasa.gov/press_release/DrGhassemRAsrarStatement.html (accessed October 16, 2007).

Butler, D. 2007. "Agencies Join Forces to Share Data." *Nature* 446: 354.

CCSDS (Consultative Committee for Space Data Systems). 2002. Reference Model for an Open Archival Information System (OAIS). Washington, DC: CCSDS Secretariat, National Atmospheric and Space Administration.

CEOS (Committee on Earth Observation Satellites). 2006. Satellite Observation of the Climate System, The Committee on Earth Observation Satellites (CEOS) response to the Global Climate Observing System (GCOS) Implementation Plan (IP). Available online at http://www.ceos.org/pages/CEOS%20Response%20to%20the%20GCOS%20IP.pdf (accessed August 27, 2007).

Christy, J. R., R. W. Spencer, and W. D. Braswell. 2000. "MSU Tropospheric Temperatures: Dataset Construction and Radiosonde Comparisons." *Journal of Atmospheric and Oceanic Technology* 17(9): 1153–1170.

Christy, J. R., R. W. Spencer, W. B. Norris, W. D. Braswell, and D. E. Parker. 2003. "Error Estimates of Version 5.0 of MSU–AMSU Bulk Atmospheric Temperatures." *Journal of Atmospheric and Oceanic Technology* 20(5): 613–629.

CCSP (Climate Change Science Program). 2003. Vision for the Program and Highlights of the Scientific Strategic Plan, A Report by the Climate Change Science Program and the Subcommittee on Global Change Research. Washington, DC: Climate Change Science Program, Subcommittee on Global Change Research.

CCSP. 2006. Temperature Trends in the Lower Atmosphere, Steps for Understanding and Reconciling Differences. Synthesis and Assessment Product 1.1. T. R. Karl, S. J. Hassol, C. D. Miller, and W. L. Murray, eds. Washington, DC: Climate Change Science Program, Subcommittee on Global Change Research.

Consultative Committee for Space Data Systems. 2002. Recommendation for Space Data System Standards: Reference Model for an Open Archival Information System.

Fetterer, F., and K. Knowles. 2004. Sea Ice Index. Boulder, CO: National Snow and Ice Data Center. Available online at http://nsidc.org/data/seaice_index/ (accessed April 13, 2005).

Group on Earth Observations. 2005. Global Earth Observation System of Systems (GEOSS), 10-Year Implementation Plan. Reference Document GEO 1000R / ESA SP-1284. Noordwijk, The Netherlands: European Space Agency Publications Division, ESTEC.

Hrastar, J. 2003: National Aeronautics and Space Administration (NASA) Earth Observing System Data Information System (EOSDIS) Case Study. Greenbelt, MD: National Aeronautics and Space Administration.

ICSU (International Council for Science). 2004. Report of the CSPR Assessment Panel on Scientific Data and Information. Paris, France: International Council for Science.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph. 1996. "The NMC/NCAR 40-Year Reanalysis Project." *Bulletin of the American Meteorological Society* 77: 437–471.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S-K Yang, J. J. Hnilo, M. Fiorino, and G. L. Potter. 2002. "NCEP-DEO AMIP-II Reanalysis (R-2)." *Bulletin of the American Meteorological Society* 83(11): 1631–1643.

Lautenbacher, C. 2006. Linking Earth Observations to Societal Benefits: The Data Management Connection. Presentation at the American Meteorological Society Annual Meeting, January 5, 2006, Atlanta, GA.

Lazo J., and L. Chestnut. 2002. Economic Value of Current and Improved Weather Forecasts in the U.S. Household Sector. Silver Spring, MD: National Oceanic and Atmospheric Administration. 213 pp. Available online at http://www.economics.noaa.gov/library/documents/benefits_of_weather_and_climate_forecasts/econ_value-weather_forecasts-households.pdf (accessed August 27, 2007).

Maiden, M. 2006. EOSDIS Evolution in Support of Measurement Needs/Science. Presentation to the MODIS Science Team, January 4, 2006, Baltimore, MD.

Maslanik, J., T. Agnew, M. Drinkwater, W. Emery, C. Fowler, R. Kwok, and A. Liu. 1998. Summary of Ice-Motion Mapping Using Passive Microwave Data. NSIDC Special Report 8. Boulder, CO: National Snow and Ice Data Center.

Millard, K., S. Newcombe, R. Vaughan, C. Arleton, D. Sauzade, R. Kalaydijan, E. Buisman, P. Salz, B. Cahill, and J. Monso. 1998. *Establishing the Value of Environmental Data Products: A Guide to Present Practice in Coastal Zone Management.* Howbery Park, Wallingford, UK: H.R. Wallingford, Ltd.

Mock, D. 2001. Implementing the NOAA Center of Data Concept at the Climate Diagnostics Center. Presented at the American Meteorological Society Annual Meeting, November 15–18, 2001, Atlanta, GA.

NASA-NOAA, 2007: Impacts of NPOESS Nunn-McCurdy Certification on joint NASA-NOAA climate goals. Draft white paper, January 8.

NOAA (National Oceanic and Atmospheric Administration). 2001. The Nation's Environmental Data: Treasures at Risk. Washington, DC: National Oceanic and Atmospheric Administration.

NOAA. 2003. NOAA Environmental Data: Foundation for Earth Observations and Data Management System. Washington, DC: National Oceanic and Atmospheric Administration.

NOAA. 2005. Report to Congress on Data and Information Management. Washington, DC: National Oceanic and Atmospheric Administration.

NOAA. 2006a. Integrating the Pieces: Assessment of NOAA's Environmental Data and Information Management. Washington, DC: National Oceanic and Atmospheric Administration.

NOAA. 2006b. Global Earth Observation Integrated Data Environment (GEO-IDE) Draft Implementation Plan. Washington, DC: National Oceanic and Atmospheric Administration.

NOAA. 2006c. Global Earth Observation Integrated Data Environment (GEO-IDE) Concept of Operations. Version 3.3. Washington, DC: National Oceanic and Atmospheric Administration.

NOAA. 2006d. CLASS Draft Request for Information or Solicitation for Planning Purposes, September 12, 2006.

NRC (National Research Council). 1994. *Review of the World Center-A for Meteorology and the National Climatic Data Center.* Washington, DC: National Academy Press.

NRC. 1995a. *Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government: Working Papers*. Washington, DC: National Academy Press.

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

NRC. 1995c. *On the Full and Open Exchange of Scientific Data.* Washington, DC: National Academy Press.

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

NRC. 1998. *Review of NASA's Distributed Active Archive Centers.* Washington, DC: National Academy Press.

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

NRC. 2000. *From Research to Operations in Weather Satellites and Numerical Weather Prediction: Crossing the Valley of Death*. Washington, DC: National Academy Press.

NRC. 2001. *Resolving Conflicts Arising from the Privatization of Environmental Data*. Washington, DC: National Academy Press.

NRC. 2003a. *Government Data Centers: Meeting Increasing Demands*. Washington, DC: The National Academies Press.

NRC. 2003b. *Review of NOAA's National Geophysical Data Center*. Washington, DC: The National Academies Press.

NRC. 2003c. *Satellite Observations of the Earth's Environment: Accelerating the Transition from Research to Operations.* Washington, DC: The National Academies Press.

NRC. 2004a. *Climate Data Records from Environmental Satellites*. Washington, DC: The National Academies Press.

NRC. 2004b. *Utilization of Operational Environmental Satellite Data: Ensuring Readiness for 2010 and Beyond*. Washington, DC: The National Academies Press.

NRC. 2006. *Earth Science and Applications from Space: Urgent Needs and Opportunities to Serve the Nation*. Washington, DC: The National Academies Press.

NRC. 2007. *Earth Science and Applications from Space: National Imperatives for the Next Decade and Beyond*. Washington, DC: The National Academies Press.

NSB (National Science Board). 2005. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. NSB-05-40. Arlington, VA: National Science Foundation.

NSF (National Science Foundation). 2003. Final Report from the Workshop on Research Challenges in Digital Archiving and Long-Term Preservation. Arlington, VA: National Science Foundation.

Parsons, M., and R. Duerr. 2005. "Designating User Communities for Scientific Data: Challenges and Solutions." *Data Science Journal* 4: 31–38.

PREMIS Working Group. 2005. Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group. Dublin, OH, and Mountain View, CA: Online Computer Library Center and Research Library Group.

Schubert, S., D. Dee, S. Uppala, J. Woolen, J. Bates, and S. Worley. 2006. Report of the Workshop on "The Development of Improved Observational Data Sets for Reanalysis: Lessons Learned and Future Directions." College Park, MD, September 28–29, 2005. Greenbelt, MD: National Aeronautics and Space Administration Global Modeling and Assimilation Office.

Spencer, R. W., J. R. Christy, W. D. Braswell, and W. B. Norris. 2006. "Estimation of Tropospheric Temperature Trends from MSU Channels 2 and 4." *Journal of Atmospheric and Oceanic Technology* 23(3): 417–423.

Stroeve, J., and J. Smith. 2001. Comparison of Near Real-Time DMSP SSM/I Daily Polar Gridded Products and SSM/I Polar Gridded Products. NSIDC Special Report 10. Boulder, CO: National Snow and Ice Data Center.

Stroeve, J., X. Li, and J. Maslanik. 1997. An Intercomparison of DMSP F11- and F13-derived Sea Ice Products. NSIDC Special Report 5. Boulder, CO: National Snow and Ice Data Center.

USGCRP. 1999. Global Change Science Requirements for Long-Term Archiving. Report of the Workshop, October 28–30, 1998. Washington, DC: USGCRP Program Office.

Appendix A

Statement of Task

The ad hoc committee charged to conduct this study will assist NOAA as it develops plans to meet its data archiving and data access requirements. The committee will first produce a letter report that includes a preliminary list of principles and guidelines that NOAA can use to begin planning specific archiving strategies for the data streams it currently collects. This preliminary set of principles and guidelines for data archiving will be refined and expanded in a final report that will also address the extent to which a wide variety of data sets should be made available. The final committee report will also include specific examples of how these principles and guidelines could be applied to existing and planned data streams across NOAA.

In summary, the committee will:

1. Consider the existing and planned observational and derived data streams collected by NOAA, along with current data management procedures, in the context of existing NOAA legal requirements.
2. Consider the relative costs of preserving and providing access to certain types of derived data products versus regenerating these data from archived first-stream input.
3. Develop a list of principles and guidelines regarding the types of data that should be archived indefinitely (for at least 75 years) and the types of data that could be stored for shorter durations under budgetary constraints.
4. Develop a list of principles and guidelines on how best to provide access to different variables, data sets and derived products, illustrated with a limited set of examples and case studies.

Appendix B

Biographical Sketches of Committee Members

David A. Robinson (*Chair*) is the Chairman of the Geography Department at Rutgers University. He has expertise in the collection and archiving of accurate climatic data and is interested in climate change (particularly state and regional climate issues), hemispheric and regional snow cover dynamics, interactions of snow cover with other climate elements, and the dynamics of solar and terrestrial radiative fluxes at and close to the surface of the Earth. Robinson is the author or coauthor of approximately 130 articles, more than half of which have been published in peer-reviewed journals or as book chapters. He also is the State Climatologist for New Jersey. Robinson has served on several National Research Council (NRC) committees and chaired the NRC's Committee on Climate Data Records from Operational Satellites: Development of a NOAA Satellite Data Utilization Plan. He received his Ph.D. from Columbia University.

David C. Bader is the Director of the Program for Climate Model Diagnosis and Intercomparison, a scientist at the Lawrence Livermore National Laboratory, and Chief Scientist for the Department of Energy's (DOE's) Climate Change Prediction Program (CCPP). He previously held the positions of project manager, research scientist, and senior research scientist at Pacific Northwest National Laboratory. Bader has also been a visiting scientist at the National Center for Atmospheric Research, has had an intergovernmental program act assignment as program manager for CCPP, and has directed the DOE's Scientific Discovery through Advanced Computing program and DOE's Computer Hardware, Advanced Math-

ematics and Model Physics program. Bader received his Ph.D. in Atmospheric Science in 1985 from Colorado State University.

Donald W. Burgess is a research fellow at the Cooperative Institute for Mesoscale Meteorological Studies at the University of Oklahoma. He has served as the Chief of the Warning Research and Development Division of the National Severe Storms Laboratory. He has also served as Acting Director and Chief of Operations at the National Weather Service (NWS) Radar Operations Center, and Chief of the NWS Radar Training Branch. Burgess's research interests lie in the areas of severe weather and on techniques for improving warnings of weather hazards, particularly techniques using Doppler radar to detect and warn of tornadoes. In 1979, he was a cowinner of the Department of Commerce Silver Medal, and in 2003, he was a cowinner of the Department of Commerce Bronze Medal. He is a fellow of the American Meteorological Society. Burgess received his M.S. from the University of Oklahoma in 1974.

Kenneth E. Eis is Deputy Director of Colorado State University's Cooperative Institute for Research in the Atmosphere (CIRA). He is in charge of CIRA's infrastructure and Earth Station and data holdings, and he oversees collaborative research with CIRA collaborators working within NOAA's Earth System Research Laboratory (ESRL). In addition to this ESRL-oriented work, Eis has helped direct the research requirements for Colorado State University's Center for Geosciences Atmospheric Research Phase II-IV. Eis is also the Director of the CloudSat Data Processing Center (DPC). The DPC is responsible for the ingestion of all CloudSat data and the production of science products and their distribution to the science community. Prior to his current job and retirement from the U.S. Air Force, Lieutenant Colonel Eis was commander of the Environmental Technical Applications Center (USAFETAC), now called the Air Force Combat Climatology Center (AFCCC). At USAFETAC/CC he was responsible for all Air Force climate support and data holdings. He later served as Director of Environmental Sciences at HQ Air Weather Service and Chief Staff Meteorologist for the Air Force Systems Command. In these jobs he developed aerospace requirements in support of the Air Weather Service future systems as well as heading the analysis and mitigation efforts of Air Force advanced weapons systems environmental limitations. He also managed weather-related launch support requirements at Vandenberg and Patrick Air Force Bases.

Sara J. Graves is the Director of the Information Technology and Systems Center and University, Board of Trustees University Professor, and professor of Computer Science at the University of Alabama in Huntsville. She

is also the Director of the Information Technology Research Center at the National Space Science and Technology Center. Graves currently serves on the U.S. National Committee for the Committee on Data for Science and Technology (CODATA) of the International Council for Science and was elected to the CODATA Executive Committee. She is a member of NOAA's Data Archiving and Access Requirements Working Group and a member of the Advisory Committee for Cyberinfrastructure to the Director of the National Science Foundation. Graves is on the editorial board of the Statistical Analysis and Data Mining Journal and on the Emerging Technologies Panel for the Defense Intelligence Agency's Missile and Space Intelligence Command. She has served as a member of the NASA headquarters Earth System Science and Applications Advisory Committee (ESSAAC) and Chair of the ESSAAC Subcommittee on Information Systems and Services. Graves directs research and development in large-scale distributed information systems, data mining and knowledge discovery, sensor networking, semantics and ontologies, high-performance computing, and geoinformatics. She received her Ph.D. in Computer Science from the University of Alabama in Huntsville.

Ernest G. Hildner was the Director of NOAA's Space Environment Center (SEC)—the nation's official source of information about space weather storms—from 1986 to 2005. Under his direction, SEC conducted research and consulted on space weather instrument development for NOAA, NASA, and the Air Force. Hildner is a solar physicist who has worked for the High Altitude Observatory, National Center for Atmospheric Research (NCAR), and NASA's Marshall Space Flight Center as Chief of its Solar Physics Branch. He was an experimental scientist for both the Skylab and the Solar Maximum Missions during the 1970s. Hildner has published dozens of papers in coronal and interplanetary physics and is a coholder of a patent for a variable-magnification X-ray telescope. In addition to his administrative responsibilities with NOAA, among them serving as NOAA's Program Manager for Space Weather, Hildner cochaired the Committee on Space Weather for the National Space Weather Program, was a member of the advisory committees for the National Solar Observatory and NCAR High Altitude Observatory, and served on review panels for NASA and Defense Department projects. In December 2003 he received the Department of Commerce Gold Medal for advancing the nation's space weather services through the conception, funding, and development of the first-operational Solar X-ray Imager. He has twice received the Presidential Rank Award for Senior Executive Service Managers.

Kenneth E. Kunkel is the Director of the Center for Atmospheric Science at the Illinois State Water Survey, where he has held a variety of positions since 1988, including Director of the Midwestern Regional Climate Center and Director of the Office of Applied Climatology. From 1982 to 1988, he served as the New Mexico State Climatologist and had a research and teaching appointment as an associate professor at New Mexico State University. Kunkel also studied atmospheric optical phenomena as a research meteorologist with the Atmospheric Sciences Laboratory at White Sands Missile Range. His recent research has focused on climate variability, extremes, and change. He has managed several projects designed to expand, improve, and analyze surface climate data sets, with a specific emphasis on the early cooperative observer network data. He has published numerous articles in peer-reviewed scientific journals and has written three book chapters. Kunkel received his Ph.D. in 1978 in meteorology from the University of Wisconsin-Madison.

Mark A. Parsons is a Senior Associate Scientist at the National Snow and Ice Data Center at the Cooperative Institute for Research in Environmental Sciences of the University of Colorado. He is also the program manager for the International Polar Year Data Management and the Frozen Ground Data Center at the World Data Center for Glaciology. Parsons has over 15 years of data management experience, including appointments as an environmental scientist at AScI Corporation and Coe-Truman Technologies, as well as research associate and research assistant at the University of Edinburgh and Cornell University, respectively. He earned his B.Sc. in Natural Resources and Communications from Cornell University in 1988.

Mohan K. Ramamurthy is the Director of The University Corporation for Atmospheric Research's (UCAR) Unidata Program and is a scientist in the National Center for Atmospheric Research's Mesoscale and Microscale Meteorology Division. Unidata provides a broad array of data for use in geosciences education and research. In addition to providing data, Unidata also develops software for data access, processing, management, analysis, and visualization and provides support to a diverse community of 160-plus institutions vested in the common goal of sharing data. As a scientist, Ramamurthy studies weather processes and prediction, including mesoscale phenomena such as gravity waves, precipitation band, hurricanes, and ensemble forecasting. His other research interests include information technology, interactive multimedia instruction and learning, and end-to-end data services. Ramamurthy joined UCAR in 2003 after spending nearly 16 years on the faculty in the Department of Atmospheric Sciences at the University of Illinois at Urbana-Champaign. He earned his

Ph.D. from the University of Oklahoma, where his doctoral research dealt with the four-dimensional assimilation of data and modeling of disturbances associated with monsoons.

Deborah K. Smith is a senior scientist of geology and geophysics at the Woods Hole Oceanographic Institution. Her research focuses on large-scale plate tectonics, as well as the dynamics of submarine and subaerial rift zones. She routinely goes to sea and collects bathymetry, side-scan sonar, gravity, magnetic, and photo-imagery data. She has strong interests in data quality and preservation and has organized workshops about these topics. Smith also has interests in education and outreach and has written for popular magazines. She has designed a Web site permitting schoolchildren and the public to participate in a virtual research expedition and one highlighting the day-to-day lives of research scientists. Smith has recently served on two advisory committees: the RIDGE 2000 Executive Committee and the U.S. Science Advisory Committee. She earned her Ph.D. in Earth Sciences from Scripps Institution of Oceanography at the University of California in San Diego.

John R. G. Townshend holds a joint appointment as professor in the Institute for Advanced Computing Studies and the Department of Geography at the University of Maryland. He is also a member of the Department of Geography's Laboratory for Global Remote Sensing Studies. Townshend's research focuses on the use of remote sensing and advanced computing methods for improvements in the characterization of regional and global land cover. He has been a member of NASA's Moderate Resolution Imaging Spectroradiometer science team, and he is a principal investigator on the Landsat Pathfinder Project for monitoring the Earth's tropical moist forests. Townshend has also been chairman of the Joint Scientific and Technical Committee of the Global Climate Observing System. His previous NRC service includes membership on the Committee on Geophysical and Environmental Data and the Board on Earth Sciences and Resources. He served as a member of the NRC Committee for Review of the Science Implementation Plan of the NASA Office of Earth Science. Townshend earned his Ph.D. in Geography (Geomorphology) in 1971 from University College in London.

Paul D. Try is a Senior Vice President and Program Manager at Science and Technology Corporation (STC) and Former Director of the WMO/WCRP International Global Energy and Water Cycle Experiment Project Office. Try has expertise in meteorological *in situ* and remote sensors (satellite and radar), as well as data collection, processing, exchange, and archival activities. His recent STC management activities include meteoro-

logical satellite processing and application support activities for NOAA's National Environmental Satellite Data and Information Service and management of several research support efforts at laboratories of NOAA's Office of Atmospheric and Oceanic Research. Prior to joining STC, he served in the U.S. Air Force Air Weather Service, where his responsibilities included oversight of the Automated Weather Distribution System, and in the Office of the Secretary of Defense, with oversight of all Department of Defense research and development in environmental sciences. Try is a fellow of the American Meteorological Society (AMS) and was president of the AMS from 1996 to 1997. He received his Ph.D. in atmospheric sciences from the University of Washington.

Steven J. Worley is the manager of the Data Support Section of the Computational and Information Systems Laboratory at the National Center for Atmospheric Research (NCAR), where he has also worked as a programmer IV and a programmer III. Before his work at NCAR, Worley worked at Texas A&M University as a research assistant and a research associate. The activities he is involved with include lead for the U.S. data management data center for The Observing [system] Research and Predictability Experiment (THORPEX) of the World Weather Research Program, the THORPEX Interactive Grand Global Ensemble (TIGGE), the Users Working Group advisory panel for the Jet Propulsion Laboratory at NASA, and chair on data archiving for the Integrated Ocean Observing System Data Management and Communications Steering Team and Expert Team. Worley is a member of the American Geophysical Union and the American Meteorological Society. He received his M.S. in Oceanography from the University of Alaska-Fairbanks in 1977.

Xubin Zeng is a professor at the University of Arizona in Tucson. Zeng's research in the past 20 years, documented in more than 80 peer-reviewed publications, has covered atmospheric turbulence (theory, parameterization, its interaction with clouds and radiation, and large-eddy simulations), mesoscale modeling of atmospheric flow over complex terrain, chaos theory and its applications to the atmosphere, global land-atmosphere interactions, ocean-atmosphere interactions, sea ice–atmosphere interactions, monsoon dynamics, remote sensing, and most recently, nonlinear dynamics of vegetation. In the past 10 years, he has focused on the land-atmosphere-ocean-sea-ice interface processes of the Earth's climate system by integrating global modeling with remote sensing and *in situ* data. Zeng has acted as a bridge linking the remote sensing and field experiment community to the weather and climate modeling community. Numerous groups worldwide (among them the National Center for Atmospheric Research, the National Centers for Environmental Predic-

tion, and the European Center for Medium-range Weather Forecasts) have used his models, algorithms, value-added datasets, and other research products. He has extensive experience with most satellite land products and some experience with satellite atmosphere and ocean products. Zeng earned his Ph.D. in atmospheric sciences from Colorado State University in 1992.

Appendix C

NOAA Data, Information, and Products

GUIDANCE FOR DEVELOPING PRINCIPLES AND GUIDELINES FOR NOAA DATA ARCHIVE AND ACCESS[1]

NOAA archives and provides access to a wide variety of data and products. These activities are based upon numerous legislative mandates. A few of many examples include:

- The National Climate Program Act of 1978 (15 USC CH29 PL 95-357), which calls for ". . . management & active dissemination of climatological data. . . ."
- Magnuson-Stevens Fishery Conservation and Management Act (Public Law 94-265), which states that "The collection of reliable data is essential to the effective conservation, management, and scientific understanding of the fishery resources of the United States."

NOAA must continue to archive and provide access to all data as required by law. This is a fundamental principle that NOAA will strictly adhere to, so this NRC study can be of most value to NOAA by focusing on the scientific and societal value of NOAA's data and not the legislative mandates.

NOAA's data, information, and products have previously been

[1]Submitted by NOAA's National Environmental Satellite, Data, and Information Service to the NRC Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA.

grouped into seven broad categories.[2] Five of the seven categories are relevant to archived data and this NRC study. They appear in bold text in the following list: **(1) Original Data; (2) Synthesized Products; (3) Interpreted Products; (4) Hydrometeorological, Hazardous Chemical Spill, and Space Weather Warnings, Forecasts, and Advisories;** (5) Natural Resource Plans; **(6) Experimental Products;** and (7) Corporate and General Information. These seven categories are described in more detail below.

Original Data are data in their most basic useful form. These are data from individual times and locations that have not been summarized or processed to higher levels of analysis. While these data are often derived from other direct measurements (for example, spectral signatures from a chemical analyzer, electronic signals from current meters), they represent properties of the environment. These data can be disseminated in both real time and retrospectively. Examples of original data include buoy data, survey data (for instance, living marine resource and hydrographic surveys), biological and chemical properties, weather observations, and satellite data.

Synthesized Products are those that have been developed through analysis of original data. This category includes analysis through statistical methods; model interpolations, extrapolations, and simulations; and combinations of multiple sets of original data. While some scientific evaluation and judgment are needed, the methods of analysis are well documented and relatively routine. Examples of synthesized products include summaries of fisheries landings statistics, weather statistics, model outputs, data display through Geographical Information System techniques, and satellite-derived maps.

Interpreted Products are those that have been developed through interpretation of original data and synthesized products. In many cases, this information incorporates additional contextual and/or normative data, standards, or information that puts original data and synthesized products into larger spatial, temporal, or issue-oriented contexts. This information is derived through scientific interpretation, evaluation, and judgment. Examples of interpreted products include journal articles, scientific papers, technical reports, and production of and contributions to integrated assessments. (Note, however, that this does not include the data and products used to make these interpretations.) These data are not applicable to the NRC study.

Hydrometeorological, Hazardous Chemical Spill, and Space Weather Warnings, Forecasts, and Advisories are time-critical interpretations of

[2]National Oceanic and Atmospheric Administration Information Quality Guidelines, 2002; available online at http://www.noaanews.noaa.gov/stories/iq.htm.

original data and synthesized products, prepared under tight time constraints and covering relatively short, discrete time periods. As such, these warnings, forecasts, and advisories represent the best possible information in given circumstances. They are derived through scientific interpretation, evaluation, and judgment. Some products in this category, such as weather forecasts, are routinely prepared. Other products, such as tornado warnings, hazardous chemical spill trajectories, and solar flare alerts, are of an urgent nature and are prepared for unique circumstances.

Natural Resource Plans are information products that are prescribed by law and have content, structure, and public review processes (where applicable) that are based upon published standards (for example, statutory or regulatory guidelines which are not applicable to the NRC study). These plans are a composite of several types of information (scientific, management, stakeholder input, policy) from a variety of internal and external sources. Examples of natural resource plans include fishery, protected resource, and sanctuary management plans and regulations, and natural resource restoration plans.

Experimental Products are products that are experimental (in the sense that their quality has not yet been fully determined) in nature, or are products that are based in part on experimental capabilities or algorithms. Experimental products fall into two classes. They are either (1) disseminated for experimental use, evaluation or feedback, or (2) used in cases where, in the view of qualified scientists who are operating in an urgent situation in which the timely flow of vital information is crucial to human health, safety, or the environment, the danger to human health, safety, or the environment will be lessened if every tool available is used. Examples of experimental products include imagery or data from non-NOAA sources, algorithms currently being tested and evaluated, experimental climate forecasts, and satellite imagery processed with developmental algorithms for urgent needs (such as wildfire detection).

Corporate and General Information includes all nonscientific, nonfinancial, and nonstatistical information. Examples include program and organizational descriptions, brochures, pamphlets, education and outreach materials, newsletters, and other general descriptions of NOAA operations and capabilities. These data are not applicable to the NRC study.

Appendix D

Terms of Reference for NOAA's Data Archiving and Access Requirements (DAAR) Working Group

Provide scientific advice and broad direction to NOAA regarding the wide range of data, information, and products that NOAA should archive and how NOAA can best provide access to this information. The Data Archiving and Access Requirements (DAAR) Working Group will evaluate data archiving and access requirements from all of NOAA's observing systems and computational models, as well as non-NOAA information as specified below_

DAAR will:

1. Prioritize data sets and products that NOAA should archive. These data sets can be reported from NOAA observing systems and other observing systems operated by Federal and non-Federal agencies, and organizations.
2. Make specific recommendations with respect to data access.
3. Provide advice on an ongoing basis as data and IT technology evolve and change.
4. Report at a minimum of once per year to the NOAA Science Advisory Board.
5. Establish and dissolve subcommittees, as appropriate, that address specific issues related to data archive and data access.

Initially, the DAAR will:

- Use external documents as an aid in the DAAR decision process, e.g.,
    1) NRC letter report outlining principles and guidelines as to what NOAA should archive.
    2) Longer NRC report completed in FY07 which would include principles and guidelines regarding data and information archive and access,
    3) Government Internal Review of NOAA's Comprehensive Large Array-Data Stewardship System (CLASS).
- Evaluate CLASS data archive/access requirements as an initial DAAR task as CLASS is expected to evolve into the preferred architecture for all NOAA data archive and access systems.

Appendix E

Acronyms and Initialisms

AVHRR	Advanced Very High Resolution Radiometer
AFCCC	Air Force Combat Climatology Center
ASCE	American Society of Civil Engineers
AVISO	Archiving, Validation and Interpretation of Satellite Oceanographic data
BASC	Board on Atmospheric Sciences and Climate
CCSP	Climate Change Science Program
CDO	Climate Data Online
CEOS	Committee on Earth Observation Satellites
CLASS	Comprehensive Large-Array [data] Stewardship System
CIRA	Cooperative Institute for Research in the Atmosphere
DAAR	Data Archiving and Access Requirements Working Group
DMC	Data Management Committee
DMAC	Data Management and Communications subsystem
DMSP	Defense Meteorological Satellite Program
DOE	Department of Energy
DAAC	Distributed Active Archive Center
EOS	Earth Observing System
EOSDIS	Earth Observing System Data and Information System
EROS	Earth Resources Observation and Science

ESRL	Earth System Research Laboratory
ESSAAC	Earth System Science and Applications Advisory Committee
ESRL	Earth Systems Research Laboratory
ENSO	El Niño-Southern Oscillation
ECMWF	European Center for Medium-range Weather Forecasts
FGDC	Federal Geographic Data Committee
FTP	File Transfer Protocol
GIS	Geographic Information System
GOES	Geostationary Operational Environmental Satellite
GCOS	Global Climate Observing System
GEO-IDE	Global Earth Observation Integrated Data Environment
GEOSS	Global Earth Observing System of Systems
GEO	Group on Earth Observations
IOOS	Integrated Ocean Observing System
ICSU	International Council for Science
IHO	International Hydrographic Organization
IOC	International Oceanographic Commission
IP	Internet protocol
MOU	Memorandum of Understanding
MSU	Microwave Sounding Unit
MODIS	Moderate Resolution Imaging Spectroradiometer
MODAPS	MODIS Adaptive Processing System
NAO	NOAA Administrative Order
NARA	National Archives and Records Administration
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NCDC	National Climatic Data Center
NESDIS	National Environmental Satellite, Data, and Information Service
NGDC	National Geophysical Data Center
NIDIS	National Integrated Drought Information System
NOAA	National Oceanic and Atmospheric Administration
NODC	National Oceanographic Data Center
NOMADS	National Operations Model Archive Distribution System
NPOESS	National Polar Orbiting Environmental Satellite System
NSF	National Science Foundation
NSIDC	National Snow and Ice Data Center

NWS	National Weather Service
NEXRAD	NEXt-generation RADar
NOSA	NOAA Observing Systems Architecture
NOSC	NOAA Observing Systems Council
NPP	NPOESS Preparatory Project
OAIS	Open Archival Information System Reference Model
OMB	Office of Management and Budget
OPeNDAP	Open-source Project for a Network Data Access Protocol
PREMIS	PREservation Metadata: Implementation Strategies
STC	Science and Technology Corporation
SEC	Space Environment Center
USAFETAC	United States Air Force Environmental Technical Applications Center
USGS	United States Geological Survey
WMO	World Meteorological Organization