Government Data Centers: Meeting Increasing Demands

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

ISBN: 0-309-50721-9, 70 pages, 6 x 9, (2003)

GOVERNMENT DATA CENTERS: Meeting Increasing Demands

Committee on Coping with Increasing
Demands on Government Data Centers

Committee on Geophysical and Environmental Data
Board on Earth Sciences and Resources
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 the federal agencies of the U.S. Global Change Research Program (USGCRP) through the National Aeronautics and Space Administration (NASA) under Contract No. NASW-01008. 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: 0-309-08742-2

Additional copies of this report are available from:

National Academies Press
500 Fifth Street, N.W.
Box 285
Washington, DC 20055
(800) 624-6242
(202) 334-3313 (in the Washington metropolitan area)
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. Bruce M. Alberts 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. Wm. A. Wulf 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. Bruce M. Alberts and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON COPING WITH INCREASING DEMANDS ON GOVERNMENT DATA CENTERS

JEFF DOZIER, *Chair*, University of California, Santa Barbara
ANURAG ACHARYA (through June 2002), Google, Inc., Mountain View, California
LAWRENCE BUJA, National Center for Atmospheric Research, Boulder, Colorado
LEO MARK, Georgia Institute of Technology, Atlanta
JONATHAN OVERPECK, University of Arizona, Tucson
MARY F. WHEELER, University of Texas, Austin
THOMAS R. YENGST, The Aerospace Corporation, Los Angeles, California

NRC Staff

KERI H. MOORE, Study Director
MONICA R. LIPSCOMB, Research Assistant
SHANNON L. RUDDY, Senior Project Assistant

COMMITTEE ON GEOPHYSICAL AND ENVIRONMENTAL DATA

J. BERNARD MINSTER, *Chair,* University of California, San Diego
ROGER C. BALES, University of Arizona, Tucson
MARY ANNE CARROLL, University of Michigan, Ann Arbor
JEFF DOZIER, University of California, Santa Barbara
DAVID GLOVER, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts
MARK J. MCCABE, Georgia Institute of Technology, Atlanta
JOHN M. MELACK, University of California, Santa Barbara
ROY RADNER, New York University, New York
ROBERT J. SERAFIN, National Center for Atmospheric Research, Boulder, Colorado

NRC Staff

ANNE M. LINN, Senior Program Officer
SHANNON L. RUDDY, Senior Project Assistant

BOARD ON EARTH SCIENCES AND RESOURCES

NRC Staff

TERESIA K. WILMORE, Project Assistant
WINFIELD SWANSON, Technical Editor

Acknowledgments

This report has been reviewed by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council'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:

Tom Barclay, Bay Area Research Group, Microsoft Corporation, San Francisco
Francis Bretherton, University of Wisconsin, Madison
James Frew, Bren School of Environmental Sciences and Management, University of California, Santa Barbara
Patricia G. Selinger, Director of Database Integration, IBM Silicon Valley Laboratory, San Jose, California
J. Ronald Wilson, Marine Environmental Data Service (retired), Ontario, Canada

Although the individuals listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions and recommendations nor did they see the final draft of

the report before its release. The review of this report was overseen by Debra Meese, Cold Regions Research and Engineering Laboratory. Appointed by the National Research Council, she 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 committee and the institution.

Preface

As repositories of the nation's environmental data, U.S. government data centers are constantly evolving. The data they collect, disseminate, and archive are critical to assessing the state of the earth and our effect on it. As the data record grows, so does our understanding of the environment. However, because of the increasing amount and complexity of and demand for environmental data, data centers seek technological approaches that would increase their capabilities while maintaining their quality of service.

At the request of the U.S. Global Change Research Program, the National Research Council formed the Committee on Coping with Increasing Demands on Government Data Centers (Appendix A). The committee was charged to hold a workshop to examine the extent to which emerging technologies can help data centers meet user needs and build and maintain the long-term record of environmental change.

The workshop on April 29-30, 2002, at the University of Texas at Austin (Appendix B) was attended by representatives from U.S. government data centers and the global environmental science community, as well as by experts in information technology (IT) from industry and academia (Appendix C). After an introductory plenary session, speakers and participants divided into two working groups: data access and ingest, and data distribution and processing. The group reconvened in plenary session at the end of the afternoon to share the results from their discussion (Appendix D). The following morning a reaction panel with representatives from data centers, the user

community, and the IT industry assessed the conclusions from the first day's deliberations (Appendix D). These discussions and subsequent work by the committee form the basis for this report.

The committee reviews technological approaches that should be given consideration not only by the data center managers and their sponsoring agencies but also by user communities. Some of these approaches are already being implemented at some data centers. However, limitations of budget and time preclude this report being a comprehensive review of individual data center operations.

The committee would like to thank the workshop participants, whose participation and expertise made the event successful. In addition, the committee would like to acknowledge the contributions of Marjory Blumenthal, director, and Jon Eisenberg of the Computer Science and Telecommunications Board; Anne Linn, director of the Committee on Geophysical and Environmental Data; and especially Keri Moore, study director, who worked diligently toward the completion of the project.

Jeff Dozier
Chair

Contents

Executive Summary

Environmental data centers have been successfully acquiring, disseminating, and archiving data for decades, but the increasing volume and number of datasets and more demands from more diverse users are making it difficult for data centers to maintain the record of environmental change. At the request of the United States Global Change Research Program (USGCRP), the National Research Council (NRC) held a workshop on Coping with Increasing Demands on Government Environmental Data Centers. The objectives of the workshop were to consider technological solutions that could enhance the ability of users to find, interpret, and analyze information held in environmental data centers and that could help data centers collect, store, share, manage, and distribute large volumes of data.

The workshop focused on technological approaches that should be given consideration not only by data center managers and their sponsoring agencies but also by various user communities. These solutions could improve both data center operations and the ability of a wide variety of users to obtain data. This report is based on discussions from the workshop and committee deliberations, and the focus areas were identified by workshop participants.

Data ingest into the major data centers appears to be well planned and executed. The process of acquiring environmental data from the centers for research or commercial use, however, continues to be difficult. The workshop considered the following areas where advanced technologies would help data centers' performance:

- improved application of standard translatable formats;
- greater reliance on on-line data storage and network access;
- more sophisticated database technologies;
- expanded metadata management and lineage tracking; and
- greater reliance on nonspecialized, easily available hardware and software solutions.

IMPROVED APPLICATION OF STANDARD TRANSLATABLE FORMATS

Data and metadata formats evolve as the priorities of data producers and users change. Although it is not possible to create a single standard format for data and metadata that meets the needs of every purpose for every dataset and user group, greater uniformity would make it easier for users to query, search, subset, access, and integrate data. In particular, using a standard format, such as XML, for metadata would enable some of these data to be generated automatically, stored in searchable databases, and easily translated among user applications.

Recommendation: With their user communities, data centers should accelerate work toward standardizing and making formats more transparent for data and metadata and thereby improve distribution and interoperability between data centers, between data centers and users, and between users. Metadata formatted in XML would assure that recipients would be able to parse data automatically and feed them directly to their applications.

GREATER RELIANCE ON ON-LINE DATA STORAGE AND NETWORK ACCESS

Providing network access to datasets in an accessible directory hierarchy would ease access to and distribution of data. This approach vastly increases distribution efficiency when subsetting tools are also made available by the data center holding the dataset: users can treat datasets as local files and use subsetting tools to extract only the portions they need, thereby reducing the network bandwidth needed for the acquisition. Network bandwidth is already widely available for retrieval of large volumes of data. However, the use of network bandwidth for data delivery relies extensively on the ability to access data randomly and would require the implementation of suitable database management

and subsetting tools at the data centers. The off-line and near on-line storage techniques (e.g., tape robots) currently used by many data centers can hinder these solutions.

Disk storage is now competitive with tape for long-term, archival-class storage. Over the past decade, disk storage and access have had a greater increase in performance for a given price than any other part of the computing industry, and other technologies for dense storage of information are the subject of much research activity in both industry and academia.

Recommendation: Data centers and their sponsoring agencies should shift the primary storage medium from tape to disk. In addition, data centers and their sponsoring agencies should enable direct random on-line access through networks and provide support for remote queries on databases.

MORE SOPHISTICATED DATABASE TECHNOLOGIES

Files are a reasonable way to organize data when the physical storage medium is tape; however, disk storage permits data to be organized in much more flexible databases. Database techniques structure ordered and related lists of parameters for the application of efficient processing algorithms. The power of database techniques lies in the ability to relate parameters from one dataset to another, thereby reducing processing and storage requirements. Adopting database technologies could significantly improve data center operations because they change the way users search, query, and access data and the way data centers acquire and store data.

Recommendation: Data centers and their sponsoring agencies should implement database technologies. When applicable, these technologies can improve data search and query, access and acquisition, interoperability, and retrieval from storage.

EXPANDED METADATA MANAGEMENT AND LINEAGE TRACKING

As more precise means of measuring and monitoring the environment are developed, the number and volume of the resulting data products increase, and the management of metadata, or data about data, becomes increasingly important. Metadata must be stored, queried, communicated, and maintained just like the data they describe. Data centers have spent considerable effort preserving metadata by routinely documenting information on data lineage, such as the source data, transformation processes, and quality assurance information, of their datasets. Open access to summaries of the dataset assembly processes and lineage has contributed significantly to ensuring user confidence in data product quality.

However, the lack of a definitive universal system for lineage metadata has resulted in incomplete or missing information. The practice of retaining complete data lineage and authenticity information as metadata should be incorporated in the large volumes of scientific data being produced today. In addition, although data centers encourage citation, there is a need for an accepted universal method for citing data products, their origin, or the processing that has been applied to them. Most centers and even some scientific journals have a preferred mode of citation, but dataset citation remains uncommon.

Routine documentation of the original data sources and the subsequent transformation and fusion steps used to develop a processed dataset would be most efficiently carried out by automated tools. Fortunately, database technology and standard formats can be as useful for metadata management as they are for data management. The self-describing approach adopted in the definition of extensible languages, such as XML Schema, is an important step in realizing technologies to support metadata management in government data centers.

Recommendation: To ensure that the greatest use is made of environmental data, (1) data producers should include data lineage and authenticity information in the metadata; (2) data centers should improve management of and access to metadata through standard formats and database technologies; and (3) users should routinely cite the data products they use in their investigations, using agreed upon dataset identifiers. To the greatest extent possible, data centers and data producers should rely on automatic tools for creating and managing metadata.

GREATER RELIANCE ON NONSPECIALIZED, EASILY AVAILABLE HARDWARE AND SOFTWARE SOLUTIONS

Because development and support costs for widely used products are lower, more and more data solutions are likely to be adapted from market-driven and market-proven technologies in an environment of constrained resources. Data centers have dedicated substantial funds toward custom hardware and software development that were the "right answer" two decades ago, but today the data centers should embark on collaborations with industry to apply these proven, easily available technologies. The problems of managing large datasets have begun to receive the attention of the commercial sector, with the result that innovative, easy-to-use methods and tools for data search, retrieval, and analysis are widespread. Moreover, easily available commodity hardware can also be used for data ingest, storage, and distribution.

In addition, the open-source movement (software with its source code made available without any restrictions) has addressed many requirements of the data centers (e.g., Nepster, ModSTER, authentication, lineage tracking). Therefore, a combination of commercial and open-source software minimizes the need for expensive custom development. Finally, while most data centers are managed as centralized organizations, a federated distributed system would formalize current user practices of obtaining some scientific products from colleagues and data projects instead of from data centers and could help reduce infrastructure and management costs for data centers.

Recommendation: Data centers should adopt commodity hardware and commercial and open-source software solutions to the widest extent possible and concentrate their own efforts on problems that are unique to environmental data management. In addition, data centers and user communities should take advantage of federated distributed systems for making data available.

IMPLEMENTATION

To balance the risks of adopting new technologies, smaller-scale prototypes can create a framework for tests with operations and with users. Using demonstration data centers is one means of effectively jump-starting new applications and sharing of new technology.

Recommendation: Data centers and their sponsoring agencies should create independent demonstration data centers aimed at testing applicable technologies and satisfying the data needs of a range of users, including interdisciplinary and nontechnical users. These centers might best prove technological approaches through several participants working in parallel.

Deliberate and appropriate transition to new technology will require planning and testing of technology concepts. In many cases, it is possible to make a gradual transition with periodic migration of datasets and updates to data systems. In other cases, a disruptive transition may be justified. New technologies can help deal with increasing amounts of data, differing data types, changing user communities, and steadily increasing demands of users and data providers. However, in some cases transition will require that software for data ingest, data processing, and data access be rewritten.

Recommendation: Data centers should aggressively adopt newer, more "bleeding edge" technical approaches where there might be significant return on investment. This should be done carefully to minimize the inevitable failures that will occur along the way. Even with the failures, the committee believes the cost savings and improvements for end users will be substantial when compared to the methods practiced today.

The nation's data centers have achieved notable successes. They store huge volumes of data reliably and provide some widely used and trusted products. The challenges posed by the rapidly expanding quantity and diversity of environmental data and increasing user demands can be met in part through technological solutions. The approaches identified in this report would substantially improve users' abilities to search for, find, and retrieve information from data centers. The size of the user community would increase, users' efficacy would improve, and scientific researchers would benefit: all would inevitably improve the information without which policy makers cannot make decisions on climate change.

Although technology can contribute to the solution of important environmental data management problems, human effort is still central to data center operations. Data centers should ensure that the latest technologies are assessed for their relevance and utility. Without question, data centers should not rely solely on technology without continuing to invest in the scientific and human elements of data management and data center operations.

1

About the Data Centers

Investigation of environmental change requires the ability to compare changing conditions through time and between locations. Such comparisons are enabled by access to environmental data stored in government data centers (Table 1.1). These centers have been collecting environmental data for operational and scientific purposes for decades, and, with the lengthening record, the potential usefulness of these data continues to grow (Sidebar 1.1).

TABLE 1.1 U.S. Government Data Centers, Their Sponsoring Agencies, and Their Scientific Specialties

Center	Agency	Specialty
National Data Centers		
Carbon Dioxide Information Analysis Center (CDIAC) <http://cdiac.esd.ornl.gov/home.htm>	DOE	Atmospheric trace gases, global carbon cycle, solar and atmospheric radiation
Center for International Earth Science Information Network (CIESIN) <http://www.ciesin.org>	Columbia University[a]	Agriculture, biodiversity, ecosystems, world resources, population, environmental assessment and health, land use and land cover change

Earth Resources Observation Systems (EROS) Data Center (EDC) <http://edc.usgs.gov>	USGS	Cartographic and land remote-sensing data products
National Earthquake Information Center (NEIC) <http://neic.usgs.gov>	USGS	Earthquake information, seismograms
National Climatic Data Center (NCDC) <http://lwf.ncdc.noaa.gov/oa/ncdc.html>	NOAA	Climate, meteorology, alpine environments, ocean-atmosphere interactions, vegetation, paleoclimatology
National Geophysical Data Center (NGDC) <http://www.ngdc.noaa.gov/ngdc.html>	NOAA	Bathymetry, topography, geomagnetism, habitat, hazards, marine geophysics
National Oceanographic Data Center (NODC) <http://www.nodc.noaa.gov>	NOAA	Physical, chemical, and biological oceanographic data
National Snow and Ice Data Center (NSIDC) <http://nsidc.org>	NOAA	Snow, land ice, sea ice, atmosphere, biosphere, hydrosphere
National Space Science Data Center (NSSDC) <http://nssdc.gsfc.nasa.gov>	NASA	Astronomy, astrophysics, solar and space physics, lunar and planetary science
Distributed Active Archive Centers (DAACs)		
Oak Ridge National Laboratory (ORNL) DAAC http://www-eosdis.ornl.gov	NASA	Terrestrial biogeochemistry, ecosystem dynamics
Socioeconomic Data and Applications Center (SEDAC) <http://sedac.ciesin.org>	NASA	Population and administrative boundaries
Land Processes (EDC) DAAC <http://edcdaac.usgs.gov/landdaac/main.html>	NASA	Land remote sensing imagery, elevation, land cover

NSIDC DAAC <http://nsidc.org/daac>	NASA	Sea ice, snow cover, ice sheet data, brightness, temperature, polar atmosphere
Goddard Space Flight Center (GSFC) DAAC <http://daac.gsfc.nasa.gov/DAAC_DOCS/gdaac_home.html>	NASA	Ocean color, hydrology and precipitation, land biosphere, atmospheric dynamics, and chemistry
Langley Research Center (LaRC) DAAC <http://asd-www.larc.nasa.gov/scar/langley.intro.html>	NASA	Radiation budget, clouds, aerosols, and tropospheric chemistry
Physical Oceanography DAAC (PO.DAAC) <http://podaac.jpl.nasa.gov>	NASA	Atmospheric moisture, climatology, heat flux, ice, ocean wind, sea surface height, temperature
Alaska Synthetic Aperture Radar (SAR) Facility DAAC <http://www.asf.alaska. edu>	NASA	Sea ice, polar processes

NOTE: DOE = Department of Energy; EPA= Environmental Protection Agency; FGDC = Federal Geographic Data Committee; NASA = National Aeronautics and Space Administration; NIH = National Institutes of Health; NOAA = National Oceanic and Atmospheric Administration; NSF = National Science Foundation; USDA = U.S. Department of Agriculture; USGS = U.S. Geological Survey.

[a] The center is supported by contracts from 22 nonfederal and federal (e.g., EPA, NIH, FGDC, USDA, NSF) agencies.

Much of the research on the interactions of natural and human-induced changes in the global environment and the implications for society is coordinated by the United States Global Change Research Program (USGCRP). The USGCRP was established through a presidential initiative in 1989 as a multiagency effort to:

- develop and coordinate a comprehensive and integrated program to increase the effectiveness and usefulness of government-supported global change research;
- address scientific uncertainties about natural and human-induced Earth system changes;

- observe, understand, predict, evaluate, and communicate the societal and environmental implications of global change; and
- provide a sound scientific basis for U.S. policies and resource management (Subcommittee on Global Change Research, 2000).

SIDEBAR 1.1
History of Data Centers

Data centers are permanent facilities that focus on the long-term maintenance, disribution, and archiving of data and data products. There are 13 discipline-based World Data Centers in the United States, including centers for atmospheric trace gases, glaciology, human interactions in the environment, marine geology and geophysics, meteorology, oceanography, paleoclimatology, remotely sensed land data, rockets and satellites, rotation of the Earth, seismology, solar-terrestrial physics, and solid Earth geophysics. In addition to these World Data Centers, federal science agencies maintain nine national data centers, which provide access to an array of publicly available datasets.

Scientists have always collected data, but the creation of data centers for improved archiving and distribution is a relatively recent and evolving activity. For example, U.S. government collection of weather observations began during the War of 1812, although weather records had been maintained in personal "weather diaries" in the United States as long ago as 1644 (Shea, 1987). In 1817 a system of weather observation was established at Weather Bureau land offices, and in 1942 a central Analysis Center was created to prepare and distribute computer weather forecasts, which later became part of the National Meteorological Center (Shea, 1987). The Weather Records Center in Asheville, North Carolina, was created by the Federal Records Act of 1950 (Public Law 754, 81st Congress; CFR § 506[c]), which combined the efforts of the Weather Bureau and the Air Force and Navy Tabulation Units. In 1957 the National Climatic Data Center was established during the International Geophysical Year and now maintains the World Data Center for Meteorology in Asheville. The National Climatic Data Center is the world's largest active archive of weather data (NOAA, 2002).

The National Space Science Data Center was established as part of the National Aeronautics and Space Administration's (NASA's) Goddard Space Flight Center in 1966 and is primarily responsible for the long-term maintenance of space science data (NASA, 2002a). Distributed Active Archive Centers (DAACs) provide

access to the complex multidisciplinary Earth Science Enterprise data from the Earth Observation System Data and Information System (EOSDIS). The DAACs differ from data centers because the focus is on the most scientifically active part of a mission or experiment, rather than on the long-term stewardship of all data. Because a permanent storage facility is not available for NASA Earth Science data, they are transferred to the National Oceanic and Atmospheric Administration or the U.S. Geological Survey 15 years after collection (NRC, 2002). The Langley Research Center (LaRC) DAAC, for example, was created in 1989 and maintains no heritage archives. Other DAACs evolved from data centers in the early 1990s, such as the Goddard Space Flight Center (GSFC) and the Earth Resources Observation System (EROS) Data Center. The Goddard Space Flight Center DAAC maintains records from 1978 on atmospheric science and hydrology. The Land Processes Data Center evolved out of the USGS EROS Data Center, created for long-term data storage in 1972 to archive, process, and distribute Landsat data. These DAACs are among 16 major data archives, data centers, and services that disseminate NASA's Earth Science and Space Science Enterprise data (NRC, 2002).

Most of the data collected through the USGCRP, as well as data for the operational purposes of individual agencies, are housed in environmental data centers.

Since their inception (Sidebar 1.1), however, demands on government data centers for archiving and distributing data have evolved and increased. Data from space missions, process studies, and field experiments continue to flow into the data centers.

The number of datasets and files and the volume of holdings have increased dramatically with the advent of new measurement programs, many of which are space based. For example, the amount of data that the National Oceanic and Atmospheric Administration (NOAA) archives increased from 20 terabytes in 1979 to 760 terabytes in 1999 (NOAA, 2001). Moreover, the use and integration of data across scientific disciplines have increased substantially. In 1979 there were 95,400 requests (accesses) for NOAA's data compared to 4,200,530 in 1999 (NOAA, 2001). Increasing numbers of individuals and organizations outside the research communities seek information for legal matters, decision making, commercial strategies, education, and general curiosity. These users require specialized information and datasets tailored to their

individual applications and place a heavy demand on data center user services. At the same time, many data center budgets have remained flat or declined, making it difficult for data centers to fulfill their missions.

The environmental challenges facing the twenty-first century will place an increasing reliance on the full spectrum of environmental data. These data are critical for understanding how the earth system operates and how to ensure a sustainable future in the face of environmental variability and change. Scientists are interested in issues such as the composition of the atmosphere, changing ecosystems, the way carbon cycles through the environment, the human dimensions of climate change, the variability and change of climate, and the global water cycle. Commercial concerns are prodded to use resources efficiently while minimizing harm to the environment. Policy makers must make decisions on activities that may affect the environment and must determine how best to adapt to environmental changes. Finally, educators work to communicate knowledge to create a more informed populace. Data centers serve all of these user groups, although each requires different products, services, and degrees of assistance.

For example, *information providers* already know what products they want; they will be the least tolerant of barriers to immediate delivery of those products. These users must be offered direct access to standard or custom products via Web services. *Information browsers* are reasonably familiar with a data product domain but not necessarily with the scope or character of holdings in that domain. They may also wish to perform exploratory analyses on the domain to help identify product subsets of interest. *Information seekers* have a constrained notion (e.g., geophysical parameter, region, season, etc.) of what they seek but may be unfamiliar with the corresponding providers and products.

Nine national data centers and eight distributed active archive centers (DAACs) collect, disseminate, and archive environmental data (Table 1.1). Data center holdings vary and include data collected from a variety of measurement platforms—satellite, aircraft, ship, ground—with different temporal and spatial resolutions and degrees of documentation. In addition, each center focuses on specific scientific disciplines, such as oceanography, remote sensing, climatology, or snow and ice. Another variation in data center operations is with the timing of data distribution: some centers deliver data only on request, while others deliver in real time, and others are on a subscription basis.

Government data centers are repositories for the nation's environmental data. Methods of data archiving and stewardship are complemented by strategies for ingesting large volumes of raw data. In addition, data centers perform a valuable service to the scientific

community through data quality control, integration, and value-added activities, such as processing data and developing tools for data analysis and presentation. In many cases, they have been successful in developing a laudable level of customer service and satisfaction.

Increasing amounts of data, differing data types, changing user communities, and steadily increasing demands of users and data providers are precipitating a crisis in the ability of data centers to fulfill their missions. In recognition of this crisis, the centers may have to make trade-offs between maintaining existing holdings and incorporating new holdings, serving more users, or providing quality services.

These challenges prompted the USGCRP to ask the National Research Council (NRC) to host a workshop to examine the extent to which emerging technologies can help data centers meet user needs and maintain the long-term record of environmental change. The Committee on Coping with Increasing Demands on Government Data Centers (Appendix A) was charged to examine

- technological solutions that could enhance the ability of users to find, interpret, and analyze information held in environmental data centers and
- technological solutions that could help data centers collect, store, share, manage, and distribute large volumes of data.

This report results from the requested NRC workshop, which provided a starting point in identifying technological approaches that would build on present data center operations in the areas of data search, retrieval, sharing, and storage. Methods for data ingest appear to have fewer opportunities for technological innovation. This report is not a conclusive technology assessment but a summary and discussion of the challenges and approaches identified at the workshop. Individual data center operations differ, and in many cases, data centers implement new technologies, though to varying degrees. Chapter 2 expounds upon these technological approaches and potential means of implementation. The agenda, participants, and working group conclusions from the workshop are outlined in Appendixes B, C, and D, respectively. Terms and acronyms used in the report are defined in Appendixes E and F.

Finally, over the past decade, many NRC reports have addressed topics that intersect with this workshop's focus. This report is not a comprehensive review of individual data center operations, an important topic addressed by NRC (1997). The issue of community access to data

was the subject of NRC (2001). Finally, NRC (1995) covered the topic of federated distributed data centers.

2

Challenges and Opportunities

Although data management is often viewed as the least glamorous aspect of science, access to well-managed data is critical to the work of many environmental researchers, as well as to an expanding pool of commercial and nontechnical users (NRC, 2001). This chapter reviews technological approaches for data management and storage that could improve the ability of users to search, query, subset, and access data. Consideration and implementation of these approaches have already begun at some data centers but are not yet pervasive. The committee based this chapter on the working group reports presented at the workshop (Appendix D), subsequent discussions, and background information provided to the committee. The committee's expertise and deliberations form the basis of the conclusions and recommendations.

CHALLENGES IN DATA AVAILABILITY AND ACCESS

Data ingest into the major data centers appears to be well planned and well executed. The process of acquiring environmental data for research or commercial use, however, continues to be difficult. Users must first seek out the data they need, which can be time consuming and difficult because there is no comprehensive list of or universal access point to all government data holdings. Although multiple means exist to find data, the chance of missing key datasets is high. In addition,

knowing specifically what to ask for in a data search is not straightforward when query terms and procedures vary from center to center. For users who are less knowledgeable about the datasets they want, searches frequently require help from the centers' customer service representatives. However, NOAA's report to Congress, *The Nation's Environmental Data: Treasures at Risk,* notes that, although requests for NOAA's data increased from about 95,000 in 1979 to over 4 million in 1999, staffing levels decreased from 582 to 321 (NOAA, 2001).

Another challenge for data centers is to deliver only the data that the user needs and requests, neither more nor less. Subsetting is the process of extracting portions of data, such as time slices or spatially defined sections. Subsetting is especially important in large datasets, such as those generated by remote sensing. However, despite consistent user demand, there continues to be a dearth of subsetting tools. Scientific products from the data are also available, but their coverage and diversity are sparse.

Once users have found what they need, they face the challenge of obtaining the data, which can require complex skills. Although frequent users typically become adept at manipulating the infrastructure, access and retrieval methods differ from center to center, so even skilled users may be familiar with only one center's approach. Inexperienced users and investigators using many different data sources require a substantial investment of time to acquire data. Almost without exception, data centers offer multiple methods of retrieving data in their holdings (e.g., file transfer protocol (FTP), which permits users to copy files stored on data center computers, and media order, in which centers copy the data of interest onto compact disk or tape). This provides flexibility but complicates the retrieval process.

Even with the appropriate query term, knowledge of the best access methods, and available subsetting tools, access to data still depends upon the ability of the centers to store data on media that can be retrieved and manipulated easily. Data centers rely too heavily on off-line or near-line (e.g., tape robots) storage. The consequences of this are that retrieval can be slow and that searching and subsetting can be difficult.

For interdisciplinary users, the real challenge arises with integrating disparate datasets, usually obtained from different data centers. Data interoperability remains difficult because standards, formats, and metadata were chosen to optimize the usefulness of a particular dataset, rather than a collection of diverse data. The growth of on-line distributed data archives has prompted many environmental research programs to address their own interoperability needs through data formats and metadata conventions (e.g., Federal Geographic Data Committee, 1998).

However, data exchange between even the most advanced of these communities remains complex and unwieldy.

As more precise means of measuring and monitoring are developed, the number and volume of the resulting data products increase, and the management of metadata, or data about data, becomes increasingly important (Sidebar 2.1). Proper metadata management is essential for government data centers to achieve their missions. Metadata must be stored, queried, communicated, and maintained just like the data they describe. Increasingly, metadata will be a key enabling element for use by communities (e.g., interdisciplinary and nontechnical user groups) that did not originally collect the data.

SIDEBAR 2.1
Metadata

Metadata describe data and data products, allowing users to find, understand, process, and reuse data and data products. Although metadata can require increased storage capacities, they are essential for establishing confidence in the data products by providing information about the history, or lineage, of the data. Metadata in government data centers should include the following types of information:

- data formats (how information is stored within data files);
- data describing how, when, and where raw data were collected;
- descriptions of how raw data were normalized, calibrated, validated, integrated, cleaned, checked for errors, and processed;
- statistics of value distributions, etc., needed for efficient database storage and access of data;
- descriptions of data use, such as how frequently a dataset is used, whether it is subsetted, etc.; and
- data specifically designed to enhance use by interdisciplinary scientists and/or nontechnical users.

In the following sections, the committee describes some steps that would improve data availability and access, including

- improved application of standard translatable formats;

- greater on-line data storage and network access;
- more sophisticated database technologies;
- expanded metadata management and lineage tracking; and
- greater reliance on easily available, nonspecialized hardware and software solutions.

STANDARD TRANSLATABLE FORMATS

Typically, standards for data and metadata management are created by the individuals and organizations collecting the data; community organizations such as professional societies, data centers, and sponsoring government agencies; and international organizations. Formats evolve over time, with new formats introduced and others abandoned as community preferences emerge. This constant evolution results in a bewildering array of standards. Although it is not possible to create a single standard that meets the needs of every dataset and user group, greater uniformity and transparency would make it easier for users to query, search, subset, access, and integrate data.

Formats that can incorporate metadata provide added benefits. Until the early 1990s, data from remote-sensing instruments were stored primarily in binary data files, each unique to a particular sensor. Because of the lack of alternatives and the efficiency of sequential binary data storage, the data had to be stored in files on disk or tape. Metadata, if stored at all, were placed in an accompanying text file. However, in the past decade, computer scientists have devised many self-describing formats for storage of scientific data. These data formats maintain efficient binary storage but allow nonexperts to understand the layout of the data. Two popular formats currently used are netCDF and HDF (network common data form and hierarchical data format, respectively); a version of the latter is a standard used by NASA's Earth Observing System Data and Information System (EOSDIS). In essence, self-describing scientific data formats provide some level of metadata encapsulation with the data.

Databases are intimately tied to metadata as a means of allowing users to search for data products of interest. Most databases are constructed specifically for their applications; custom software is written to extract metadata from multiple sources, including data files, into these databases. As an example, the database behind EOSDIS was fashioned over many years, with new datasets processed and specific metadata entered using custom software. This process is complicated and time consuming, but it leads to providing a mechanism for searching remotely

sensed data. Moving toward standardized data and metadata formats would simplify the search process.

The next step is to generate databases automatically from the metadata. It is possible to use XML Schema to generate database tables automatically from the structure and content of the metadata, as well as to create Web-based forms for database queries. Such query interfaces allow users to formulate restrictions on the data of interest, which are then translated into selection conditions in a query language, such as SQL or XQUERY. This does not relieve sensor operators from generating appropriate metadata for their data, but it eases the search through databases.

Recommendation: With their user communities, data centers should accelerate work toward standardizing and making formats more transparent for data and metadata and thereby improve distribution and interoperability between data centers, between data centers and users, and between users. Metadata formatted in XML would assure that recipients would be able to parse data automatically and feed them directly to their applications.

NETWORK AND ON-LINE RANDOM ACCESS

Providing network file system access would ease obtaining and distributing data. Such a network would allow datasets to be used without the current formal process of copying the data across a network or sending the data physically by tape. The data become available immediately to as many users as want them. This approach can increase distribution efficiency when subsetting tools are also made available: users can treat datasets as local files and use subsetting tools to extract only the portions they need or only a transformation of the data, reducing the network bandwidth needed for the acquisition. Furthermore, once the data have been distributed, authenticity can still be guaranteed by digital signatures supplied by the national data centers. Protocols to compress and expand data automatically when they are transmitted would assist with effective network use.

Network bandwidth[1] is already widely available for retrieval of large volumes of data. However, the dependence on network bandwidth as a solution to the data delivery problem requires the implementation of

[1] Network bandwidth—capacity to move large data files electronically.

suitable database management and subsetting tools at the data centers. Few users will want gigabyte-sized datasets. In addition, network-based solutions rely extensively on the ability to access data randomly. The off-line and near on-line storage techniques (e.g., tape robots) used by many data centers can hinder these solutions. The transfer rates of modern tape systems are on the order of a few megabytes per second; common network transfer rates are 100 times faster. While disk storage capacities continue to increase dramatically, tape capacities and transfer speeds have barely increased during the past five years. In addition, without random access to on-line data, subsetting through a network is unworkable, as users cannot capture slices of the linearly stored datasets. Data that are kept off-line or near on-line cannot be used in database systems. Even databases that direct users to off-line data products must create well-defined delivery timelines. Tape systems at data centers can time-out on user requests, thus requiring a technician to process orders manually.

In 1994 computing experts forecasted that disk storage would become cheap and efficient enough to eliminate the need for off-line storage (Davis et al., 1994). However, in some cases this transition to disk will require that software for data ingest, data processing, and data access be rewritten. As a result, data centers keep most data off-line, thereby reducing the ability of users to search through and retrieve data rapidly. Data centers are moving toward increasing the availability of on-line data; however, only 3 terabytes of NOAA's 76-terabyte digital data archive are on-line (NOAA, 2001), despite the fact that disks to accommodate this amount of data would cost about $100,000 at current prices.

Over the past decade, disk storage and access have had a greater increase in performance for a given price than any other part of the computing industry, and other technologies for dense storage of information are the subject of much research activity in both industry and academia. Price per unit storage has decreased during the past 10 years. Satellite missions of the next decade will generate about 1 petabyte of information per year. As recently as 1995, NASA estimated that today's cost to store a petabyte off-line would approach $100 million, but it is now possible to obtain 1 petabyte of disks for on-line storage for less than $2 million, a very small fraction of the cost of the missions that generate the raw information. Disk storage is now competitive with tape for long-term, archival-class storage.

Recommendation: Data centers and their sponsoring agencies should shift the primary storage medium from tape to disk. In addition,

data centers and their sponsoring agencies should enable direct random on-line access through networks and provide support for remote queries on databases.

DATABASE TECHNOLOGIES

Files are a reasonable way to organize data when the physical storage medium is tape; however, disk storage permits data to be organized in much more flexible databases. Database techniques structure sets of parameters for the application of efficient processing algorithms. Traditionally, a database is composed of a number of interrelated tables containing sets of parameters such as number or text strings. The power of database techniques lies in the ability to relate parameters from one dataset to another, thereby reducing the processing and storage requirements. For example, using a numeric parameter, such as a zip code, to refer to a name, such as a city, makes it easier to store and search the information. Complex databases can have many layers of such associations.

In the early 1990s, the Structured Query Language (SQL) was formalized and is used by most database software. The language provides a standard for the following:

- defining data structures;
- defining indices;
- formulating content-based queries; and
- maintaining data through inserts, deletes, and updates.

Most database software (e.g., Oracle, MySQL, SQL Server) uses SQL as a core language for database interaction. Each has a unique method of optimizing the storage of data on disk or in memory. Capabilities for formulating spatial and temporal database queries are part of the most recent database query languages (e.g., SQL3), and support for indexing data on its spatial and temporal attributes enables efficient query execution. The complexity of the SQL query relates directly to the complexity of the database.

Contemporary database technology permits random access to subsets of data stored on disk. In addition, object-relational databases are now capable of handling large, structured data, such as aerial photographs of the entire United States. For example, since its launch in June of 1998, TerraServer has delivered 108 terabytes of U.S. Geological Survey imagery to 63 million visitors (T. Barclay, Microsoft, personal

communication, 2002). Concurrent requests from multiple users to read data can be supported efficiently without the waiting time typically incurred when many applications are writing to a database simultaneously. However, since database tables are constantly being accessed, they must be stored on-line rather than on tape.

Although databases are commonly used by data centers for metadata management, they are not in widespread use for environmental data. However, application of database technology to environmental data is possible and may be useful for some environmental datasets. For example, Sky Server utilizes database technology to provide public access to Sloan Digital Sky Survey data (Szalay et al., 2001).

Recommendation: Data centers and their sponsoring agencies should implement database technologies. When applicable, these technologies can improve data search and query, access and acquisition, interoperability, and retrieval from storage.

METADATA MANAGEMENT

Data centers have spent considerable effort preserving metadata by routinely documenting information on data lineage, such as the source data, transformation processes, and quality assurance information of their datasets. Open access to summaries of the dataset assembly processes and lineage has contributed significantly to ensuring user confidence in data product quality. For example, in most cases, users interested in data from a particular center can find information on the available data on the center's Web site.

In the past it was sufficient for data producers simply to develop good local data conventions and exercise the discipline necessary to generate the data and metadata in accordance to those conventions. However, the lack of a definitive universal system for lineage metadata can result in incomplete or missing data lineage information. In most cases, it is not possible to re-create data assembly information after the fact; in others it is costly and prone to error. Formatting data and creating metadata robust enough to be discovered and ingested by the emerging national and international data interchange networks would ensure that the data are as useful as possible, especially to other user communities. The practice of retaining complete data lineage information as metadata should be incorporated into the large volumes of scientific data being produced today. This will only be effective if accomplished with the participation and acceptance by the user communities.

Authenticity is another important aspect of data archival. Users often obtain data from the easiest source, some of which may be three or four steps removed from the data centers. At each step, the data may have been processed or reformatted to suit one user's particular purposes. Through neglect or, less likely, malicious intent, data products may become contaminated or altered, endangering their value and use. Consequently, information on authenticity should be included in the metadata.

A related issue, specific to the research community rather than to data centers, involves citing data products in the peer-reviewed literature. The scientific practice of citing past research and methods, necessary for independent verification, has been neglected when citing data supporting an investigation's findings. While this has been discussed for more than 10 years, the various publishing groups have not reached consensus on an accepted universal method for citing data products, their origin, or the processing that has been applied to them or on how to deal with the inherent challenges (e.g., numerous investigators for very large datasets). Most centers and even some scientific journals (e.g., American Geophysical Union journals) have a preferred mode of citation, but dataset citation remains uncommon. Dataset citation helps both data centers and data providers learn what data are being used and how.

Routine documentation of the original data sources and the subsequent transformation and fusion steps used to develop a processed dataset would be most efficiently carried out by automated tools. Many practices in the software engineering field, such as testing, configuration management, and bug tracking, matured only after automated tools were developed to handle the complicated bookkeeping in a systematic manner. Moreover, the generation of structured lineage metadata suitable for ingest into other software presumes the existence of automated documentation tools. However, neither such tools nor recognized semantics to describe data lineage currently exist. Fortunately, database technology and standard formats can be as useful for metadata management as they are for data management. The self-describing approach adopted in the definition of extensible languages such as XML Schema is an important step in realizing technologies to support metadata management in government data centers. This self-describing approach would allow tools developed for data management to be applied to metadata.

The data centers have worked to document data lineage, both by compliance with rich metadata standards (e.g., U.S. Geological Survey, 1995) and by the use of automated metadata tools such as the Science Data Production (SDP) toolkit (National Aeronautics and Space

Administration, 2002b), both of which encourage detailed lineage information. However, a large body of scientific data generated outside of the data centers still lack sufficient metadata information to establish the data's lineage and context. Examples of this are the Cooperative Ocean/Atmosphere Research Data Service (COARDS) and the recent climate and forecast metadata conventions, which use only a single broad "history" attribute to document the dataset's lineage.

Recommendation: To ensure that the greatest use is made of environmental data, (1) data producers should include data lineage and authenticity information in the metadata; (2) data centers should improve management of and access to metadata through standard formats and database technologies; and (3) users should routinely cite the data products they use in their investigations, using agreed upon dataset identifiers. To the greatest extent possible, data centers and data producers should rely on automatic tools for creating and managing metadata.

HARDWARE AND SOFTWARE

Because development and support costs for widely used products are lower, more and more data solutions are likely to be adapted from market-driven and market-proven technologies in an environment of constrained resources. The on-line database, entertainment, and gaming communities are all driving advances in large-scale data management, delivery, and visualization. Many researchers have learned how to construct plain-language database queries using Web search engines (e.g., Google). The data centers should be prepared to embark on collaborations with industry to apply such proven technologies and thereby reduce expensive custom development.

The problems of managing large datasets have begun to receive the attention of the commercial sector, with the result that innovative, easy-to-use methods and tools for data search, retrieval, and analysis are widespread. For example, Google manages billions of individual records, yet searches return nearly instantaneously; digiMine, Inc. processes nearly a terabyte of data nightly (B. Nayfeh, digiMine, Inc., personal communication, 2002); and together America Online and Microsoft's Hotmail handle the email accounts of more than 150 million people (Caslon Analytics, 2002). The challenges facing the data centers are small compared to the load experienced by any of the above enterprises.

Large computational problems can be solved in small pieces by harnessing the power of desktop computing. For example, SETI@home and climateprediction.net use the processing power of millions of desktop computers to solve computationally intensive problems. The Center of Excellence in Space Data and Information Sciences (CESDIS) has constructed computing farms (commonly referred to as Beowulf clusters) to handle and process large datasets (Scyld Computing Company, 1998).

Commodity hardware can also be used for data ingest, storage, and distribution. These computers generally have far smaller capabilities than the scientific computing hardware currently in the data centers. To be useful for scientific applications, the data segments, or granules, have to be broken into smaller units that can be ingested, processed, stored, and served with larger numbers of small processors. Current proprietary operating systems, such as SGi or Sun, to open-source platforms, such as Linux or FreeBSD Commodity solutions, could ease recompiling software on new computing architectures.

In addition, the open-source movement[2] has created the potential for data centers to meet future needs without enormous resource expenditures. Unrelated open-source projects (e.g., the Gnutella project, the XML standard) provide software tools at no cost that in some cases are better than unique proprietary solutions.

Forms of authentication and lineage tracking common in the open-source communities should be adopted for improving metadata management. For example, one common practice in the open-source community is to publish an MD5—message-digest algorithm 5—listing the 32-character signature of the files with any piece of software or data that is distributed. The authoritative source publishes the digest, so that users can check the authenticity of their copies, regardless of where they got them.

In summary, the commercial sector and the open-source movements have created robust software that meets many needs of the data centers. Usage and adaptation of these codes minimizes the need for expensive custom development.

Since each is generally funded by a single agency and deals with a relatively narrow range of scientific disciplines (Table 1.1), data centers tend to be managed as centralized organizations. However, a federation of distributed systems, in which data centers remain the sources of

[2]Open-source movement—software with its source code made available without any restrictions.

authenticated environmental science data but not the only sources capable of distributing data, could help reduce infrastructure and management costs (NRC, 1995). Widely distributed data sources and grid infrastructures reduce resource contention at the data centers and provide a natural backup of earth science data.

For example, Napster provided a global directory of on-line music. Users searching for particular music were redirected to numerous locations where search matches were encountered. Users then chose (based on bandwidth availability between their computers and the source, the authenticity of the source, and the exact characteristics of the music being searched for) where to download the music.

The process is more complex for environmental science data than it was for Napster. In the environmental science community, the analogy would be to *identify* (by whatever means) a desired dataset and then request the dataset by *name* (not parameters) from a Napster analog. This approach would formalize current user practices of obtaining data from colleagues and data projects instead of from data centers. It would strengthen the data centers' partnership with science by increasing the incidence of development of scientifically sound, useful products, reduce data transmission needs, and improve effectiveness and efficiency of the whole system. Multiple copies of products would be available from various sources; the data centers would become authenticators of data and the final archive and would implement production of new scientific products once a design is in hand; and users would have multiple options for retrieving data. Three current projects are attempting to implement this: MODster, NEpster, and the Distributed Oceanographic Data System (Sidebar 2.2).

Recommendation: Data centers should adopt commodity hardware and commercial and open-source software solutions to the widest extent possible and concentrate their own efforts on problems that are unique to environmental data management. In addition, data centers and user communities should take advantage of federated distributed systems for making data available.

IMPLEMENTATION

On the one hand, new "bleeding edge" technical approaches offer ways to reduce costs and significantly improve data center performance. However, it is important to recognize that some new technical approaches may not prove successful and that even those that are

successful may cause disruptions to center operations when implemented. Therefore, the data centers need to be able to test, prioritize, and develop the most promising new approaches at a smaller scale.

SIDEBAR 2.2
Distributed Solutions

Several ongoing environmental science projects are already be-nefiting from easily available nonspecialized solutions. Selected examples are described below.

MODster

Moderate Resolution Imaging Spectroradiometer (MODIS) provides global datasets with data on surface temperature, concentration of chlorophyll, fire occurrence, cloud cover, and others. Instruments on-board several NASA missions gather datasets covering a swath 2,330 kilometers wide, capturing 36 spectral bands of data at three resolution levels every two days for six-year periods. Due to their number and complexity, searching for a specific dataset is not a trivial task. To combat this, the Federation of Earth Science Information Partners is supporting the development of MODster to support the decentralization and distribution of MODIS data and services and to promote sharing of remote-sensing standard products. Organizations within the federation can retrieve standard MODIS data granules (the smallest increment of processed MODIS data that can be ordered, containing data for an area of 2,330 by 2,330 kilometers). The retrieval of these granules will be implemented by the Hypertext Transfer Protocol (HTTP) from a simple inventory server. The system will allow clients to reference MODIS granules by name alone.

SOURCES: National Aeronautics and Space Administration (2002c); Federation of Earth Science Information Partners (2002).

NEpster

The Earth Science Technology Office (ESTO) and the National Polar-Orbiting Operational Environmental Satellite System (NPOESS) Preparatory Project (NPP) support the development of the NPP-ESTO Portal for Science, Technology and Environmental Research (Nepster) to serve the remote-sensing community better. The peer-to-peer architecture of the data archive system is based on the

Napster model, a system developed for sharing music files. In NEpster, several additional features have been added to facilitate the handling of remotely sensed data, specifically (1) a temporary data storage area for sites that do not allow continual access to their servers; (2) an intelligent broker that controls data access in accordance with the distribution policies of each data source; and (3) a comprehensive geographically based query interface to expedite data searches. The NEpster system is made up of two major components: the data notification and entry subsystem, and the query engine. The first phase of NEpster development will focus on accessing and managing real-time data, and the second phase will focus on access to the MODIS Direct Broadcast data archives through the Goddard Space Flight Center DAAC.

SOURCE: National Aeronautics and Space Administration (2002d).

DODS

The Distributed Oceanographic Data System (DODS) is a highly distributed software framework for requesting and transporting data across the Internet, which allows users to control both how their data are distributed and how they can access remote data. As data users prefer to work with software with which they are most familiar, DODS servers make data available regardless of local storage formats. In addition, DODS applications allow users to transform existing data analysis and visualization applications into those able to access remote DODS data. Because DODS data are distributed by the same scientists who develop the data, the DODS protocol and software rely on the user community to use, improve, and extend the system. The current DODS Data Access Protocol (DAP) frames requests and responses using hypertext transfer protocol (HTTP). This data model has already developed a transport protocol, software framework, C++ and Java implementations of the data model and transport protocol, and a set of DODS servers and clients. Users are allowed to access any data on a DODS server via the Internet regardless of native format, without disrupting local functions and access. Although DODS was originally developed for sharing oceanographic data, the design can be applied to other user communities.

SOURCE: University Consortium for Atmospheric Research (2002).

One way to accomplish this is to create independent demonstration data centers, each of which would build small functional prototypes with

small efficient teams that would distribute data from a few substantial datasets that are well documented (such as those from NASA and NOAA). This would be similar to the smaller-scale Sky Server project (Szalay et al., 2001). The costs of implementing demonstration data centers can be minimized by building on work that is already in progress (e.g., Sidebar 2.2). Finally, the demonstration centers would also help the data centers and communities adapt to serving and interacting with a wider range of users.

One possible choice for testing new technologies is Moderate Resolution Imaging Spectroradiometer (MODIS) data. In this example the goals of the demonstration data center could include the following:

1. Define an XML Schema with the standard format definitions for the datasets. Show how the standard format definitions can be used to formulate queries on the data collection.
2. Allow multiple avenues of network access to data already available. Specifically, provide real-time access to all data. Example access protocols include:
    a. FTP browse via a hierarchical tree (sorted by data/time and location).
    b. Network File System (NFS) access via read-only network drives.
    c. Implementation similar to NEpster/MODster (Sidebar 2.2), where multiple sites maintain subsets of the entire MODIS dataset. The participating data center could solicit participation of the MODIS science team and the other sites that have MODIS downlink systems, which have some (if not all) of the data. This might entail acquiring read-only access to datasets at non-data-center locations. The goal would be to leverage the work of researchers seeking to make science data community property.
    d. FTP subscription service, if it is not already provided.
3. Enhance and publish XML-based metadata related to the datasets. This entails adding certain metadata to that already captured by EOSDIS, such as an MD5 signature for authentication. The metadata schema describing the layout of the demonstration data center and a method of providing direct SQL access of the database to users should also be published. The enhanced metadata will allow varied researchers the opportunity to explore the dataset in innovative ways.
4. Utilize database technologies for user queries and searches.
5. Identify and provide limited subsetting tools that run on the host computers. At a minimum, allow users to subset simple spatial grids and temporal intervals. Users would not need direct access to the data storage

computers; perhaps a small application in a language such as Java could accept user subsetting boundaries, subset the data (and accompanying metadata), and deliver the data via FTP.

6. Use commodity-level hardware and software where possible and cost effective.

7. Monitor access statistics of FTP, NFS, and MODster and actively pursue user feedback.

Recommendation: Data centers and their sponsoring agencies should create independent demonstration data centers aimed at testing applicable technologies and satisfying the data needs of a range of users, including interdisciplinary and nontechnical users. These centers might best prove technological approaches through several participants working in parallel.

While the costs of implementing new solutions are likely to be significant, careful strategic planning and phasing in of new solutions could greatly reduce the need to invest substantial new resources in technology. By using opportunities to adopt incremental changes in technology, data centers can spread the costs of hardware and software acquisition over time.

Recommendation: Data centers should aggressively adopt newer, more "bleeding edge" technical approaches where there might be significant return on investment. This should be done carefully to minimize the inevitable failures that will occur along the way. Even with the failures, the committee believes the cost savings and improvements for end users will be substantial when compared to the methods practiced today.

After decades of development and at least one decade of substantial investment, the nation's data centers have achieved successes. They store huge volumes of data reliably and provide some widely used and trusted products. The challenges posed by the rapidly expanding quantity and diversity of environmental data and increasing user demands can be met in part through technological solutions. Although technology can contribute to the solution of important environmental data management problems, human effort is still central to data center operations. Therefore, data centers should ensure that the latest technologies are assessed for relevance and utility but should not rely solely on technology without continuing to invest in the scientific and human elements of data management.

References

Caslon Analytics, 2002, *Caslon Analytics profile: Email, SMS, and IM*, <http://www.caslon.com.au/emailprofile5.htm>.

Davis, F., W. Farrell, J. Gray, C.R. Mechoso, R. Moore, S. Sides, and M. Stonebraker, 1994, *EOSDIS Alternative Architecture—Final Report,* Project Sequoia Team, University of California, Berkeley.

Federal Geographic Data Committee, 1998, Content standard for digital geospatial data, Metadata Ad Hoc Working Group, FGDC-STD-001-1998, Reston, Vir., <http://www.fgdc.gov/standards/documents/standards/metadata/v2_0698.pdf>.

Federation of Earth Science Information Partners, 2002, NewDISS prototyping formulation proposal selections, <http://www.esipfed.org/business/announce.html>.

National Aeronautics and Space Administration, 2002a, About the National Space Science Data Center, <http://nssdc.gsfc.nasa.gov/about/about_nssdc.html>.

National Aeronautics and Space Administration, 2002b, EOS Core System, Science data production toolkit, <http://hdfeos.gsfc.nasa.gov/hdfeos/index.cfm>.

National Aeronautics and Space Administration, 2002c, MODIS data products, <http://eosdatainfo.gsfc.nasa.gov/eosdata/terra/modis/modis_dataprod.html>.

National Aeronautics and Space Administration, 2002d, NEpster description, <http://directreadout.gsfc.nasa.gov/projects/nepster/nepster.htm>.

National Oceanic and Atmospheric Administration, 2002, What is NCDC?, <http://lwf.ncdc.noaa.gov/oa/about/whatisncdc. html#INTRO>.

National Oceanic and Atmospheric Administration, 2001, *The Nation's Environmental Data: Treasures at Risk*, Report to Congress on the status

and challenges for NOAA's environmental data systems, Washington, D.C., 138 pp.

National Research Council, 2001, *Resolving Conflicts Arising from the Privatization of Environmental Data*, National Academy Press, Washington, D.C., 99 pp.

National Research Council, 1998, *Review of NASA's Distributed Active Archive Centers*, National Academy Press, Washington, D.C., 233 pp.

National Research Council, 1997, *The Future of Spatial Data and Society*, National Academy Press, Washington D.C., 67 pp.

National Research Council, 1995, *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources*, National Academy Press, Washington, D.C., 67 pp.

Office of Management and Budget, 1996, Circular A-130, Memorandum for heads of executive departments and establishments, on the management of federal information resources, <http://www.whitehouse.gov/omb/circulars/a130/a130.html>.

Scyld Computing Company, 1998, Beowulf introduction and overview, <http://www.beowulf.org/intro.html>.

Shea, E.L., 1987, A history of NOAA, being a compilation of facts and figures regarding the life and times of the original whole Earth agency, S. Theberge, ed., National Oceanic and Atmospheric Administration, Washington, D.C., 44 pp.

Subcommittee on Global Change Research, 2000, *Our Changing Planet: FY 2001 U.S. Global Change Research Program*, White House Office of Science and Technology Policy, Washington, D.C., 74 pp.

Szalay, A., J. Gray, A. Thakar, T. Kunszt, J. Raddick, C. Stoughton, and J. VandenBerg, 2001, *The SDSS SkyServer—Public Access to the Sloan Digital Sky Server Data*, Microsoft Technical Report MSR-TR-2001-104, Redmond, WA, 23 pp.

University Consortium for Atmospheric Research, 2002, What is DODS?, <http://www.unidata.ucar.edu/packages/dods/home/faq/whatIsDods.shtml>.

U. S. Geological Survey, 1995, Modern average global sea-surface temperature dataset, <http://geochange.er.usgs.gov/pub/magsst/FGDCmeta.html>.

Appendixes

Appendix A

Biographical Sketches of Committee Members

JEFF DOZIER, *Chair*, is professor and founder of the Donald Bren School of Environmental Science and Management at the University of California, Santa Barbara. His research interests are in the fields of snow hydrology, Earth system science, remote sensing, and information systems. In particular, he has pioneered interdisciplinary studies in two areas: one involves the hydrology, hydrochemistry, and remote sensing of mountainous drainage basins; the other is in the integration of environmental science and computer science and technology. Dr. Dozier was the senior project scientist for NASA's Earth Observing System when configuration of the data and information system was being established. He has served on numerous National Research Council committees on data and information technology and is currently also a member of the Committee on Geophysical and Environmental Data.

ANURAG ACHARYA (through June 2002) is a senior software engineer at Google, Inc., where he is responsible for building and serving Google's Web index (several terabytes). Previously he was an assistant professor in the Department of Computer Science at the University of California, Santa Barbara. Dr. Acharya's research interests lie in rapidly evolvable network services, data-intensive computations, the use of active disk architectures for rapidly growing datasets, and the design of high-performance remote sensing and Earth science databases. He received the National Science Foundation Career Award in 1999.

LAWRENCE BUJA is an associate scientist at the National Center for Atmospheric Research, where he is responsible for development and application of the Community Climate System Modeling program. In addition to managing and distributing data from this terabit system, he conducts numerical simulations of future, present-day, and past climate scenarios. Dr. Buja chairs the Data Management Working Group of the University Corporation for Atmospheric Research and is involved in interagency collaborations in both climate modeling and large-scale scientific data management.

LEO MARK is an associate professor in the College of Computing at the Georgia Institute of Technology. His research interests include database system architecture, data models, information interchange, and efficient query processing. Dr. Mark has participated in several committees and working groups on database systems, including a database architecture standardization task group of the American National Standards Institute, which resulted in an international software standard, and a committee on standard data interchange structures for NASA's space science data systems.

JONATHAN OVERPECK is director of the Institute for the Study of Planet Earth and a professor at the University of Arizona. Before joining the faculty, he was director of the World Data Center for Paleoclimatology at the National Oceanic and Atmospheric Administration, where he was instrumental in building a global paleoclimate database. Dr. Overpeck's research focuses on global change dynamics, with a major component aimed at understanding how and why key climate systems vary on timescales longer than seasons and years. In recognition of his interdisciplinary climate research, he received the Department of Commerce Gold Medal and the American Meteorological Society Walter Orr Roberts Award.

MARY F. WHEELER is the Ernest and Virginia Cockrell Chair of Engineering at the University of Texas and director of the Center for Subsurface Modeling in the Texas Institute for Computational and Applied Mathematics. Her research interests include numerical solution of partial differential systems with application to reservoir engineering and contaminant transport in groundwater, bays, and estuaries. Modeling such complex systems requires massive parallel computations as well as expertise in a wide range of disciplines, and Dr. Wheeler holds appointments in the departments of Petroleum and Geosystems

Engineering, Aerospace Engineering and Engineering Mechanics, and Mathematics. She is a member of the National Academy of Engineering.

THOMAS R. YENGST is a research scientist in the Space Sensors Group of the Aerospace Corporation. He has been active in research, development, and implementation of remote-sensing technologies for both government and commercial applications. Having served on the User Working Group of the Langley Distributed Active Archive Center, he is familiar with the problems of managing the disparate datasets and large volumes of information used by the atmospheric science community.

Appendix B

Workshop Agenda

University of Texas, Austin
ACES Building, Room 6.304

April 29, 2002

Plenary Session

8:00 a.m.	Continental Breakfast
8:30	Welcome and Introductions *Jeff Dozier, Committee Chair, University of California, Santa Barbara* *Keri Moore, Study Director, National Research Council*
8:45	Major Challenges to Environmental Data Management—One User's Perspective, *Eugene Clothiaux, Pennsylvania State University*
9:15	Coping with Increasing Demands on Government Data Centers, *John Bates, NOAA*

9:45	Myth Connections, *Richard McGinnis, NASA*
10:15	Lessons Learned from EOSDIS, *Bruce Barkstrom, NASA*
10:45	Break
11:15	The Emerging Infrastructure for Environmental Information Management, *Jim Frew, University of California, Santa Barbara*
11:45	Data Ingest: A Case Study, Building a Global Environmental Database from Many Sources, *Sydney Levitus, NOAA*
12:15 p.m.	Data Distribution and Processing, *Joel Saltz, Ohio State University*
12:45	Lunch
2:00	Working Group Discussions
4:00	Break
5:00	Breakout Group Presentations
6:00	Recess

April 30, 2002

Plenary Session

8:00 a.m.	Continental Breakfast
8:30	Recap and Reiterate Objectives, *Jeff Dozier, Committee Chair*
8:45	Reaction Panel

- What is your reaction to the first day's deliberations?
- Are we on the right track?

- What technologies can be resource multipliers (near versus long term)?
- If there's one you'd apply in the short term, which would it be?
- What have we missed?
- What's the coolest thing you've heard so far?

Tom Barclay, Microsoft
Kelly Redmond, Western Regional Climate Center
Vanessa Griffin, NASA
Rob Mairs, NOAA
Basem Nayfeh, digiMine, Inc.

9:45	Discussion
10:15	Break
10:45	Revisit and Develop Themes and Conclusions From Yesterday
1:00 p.m.	Lunch
2:00	Workshop Adjourns

Appendix C

Workshop Speakers and Participants

Nabil Adam, Rutgers University, Newark, New Jersey
Tom Barclay, Microsoft Corporation, San Francisco, California
Bruce Barkstrom, NASA Langley Research Center, Hampton, Virginia
John Bates, National Climatic Data Center, NOAA, Asheville, North Carolina
David Clark, National Geophysical Data Center, NOAA, Boulder, Colorado
Eugene Clothiaux, The Pennsylvania State University, University Park
James Frew, University of California, Santa Barbara
Vanessa Griffin, Earth Science Data and Information Systems Project, NASA, Greenbelt, Maryland
Robert Grossman, University of Illinois, Chicago
Sydney Levitus, World Data Center for Oceanography, NOAA, Silver Spring, Maryland
Martha Maiden, NASA Headquarters, Washington, D.C.
Robert Mairs, NOAA/NESDIS, Silver Spring, Maryland
Richard McGinnis, NASA Headquarters, Washington, D.C.
Basem Nayfeh, digiMine, Inc., Bellevue, Washington
Connie Nelin, IBM, Austin, Texas
Silvia Nittel, University of Maine, Orono
Kelly Redmond, Western Regional Climate Center, Reno, Nevada
Joel Saltz, Ohio State University, Columbus
Hanan Samet, University of Maryland, College Park
August Shumbera, National Climatic Data Center, NOAA, Asheville, North Carolina

Appendix D

Workshop Discussions

WORKING GROUP REPORTS

Workshop participants divided into two working groups: (1) data access and ingest and (2) data distribution and processing. The questions posed to the working groups (given in italics) and their conclusions (indented) are listed below.

Data Access and Ingest Working Group

The working group on data access and ingest assessed the ways users access data and the ways that data centers collect data.

What is good and bad about the way users access data?

Subsetting capabilities should be improved, so that users obtain only the data they want.

Users do not always know of opportunities or "windows" for easier access to data. For example, potential users should be alerted before a data center transfers data to tape for storage or if the data are available elsewhere in a format that is more easily used. Data centers should track the diverse access

opportunities and improve the way users are informed about these opportunities.

Some users purposefully retrieve more data than they require, either because of uncertainty that the data will always be available or because it is often easier than retrieving subsets. This practice unnecessarily strains the network. In addition, hoarding data can waste users' storage resources and result in datasets that are not kept up-to-date.

Data collected by individual researchers are not available to the community in a timely manner and are lost when the researcher retires.

Duplication of effort in data management has many benefits and some drawbacks. Duplication can lead to new ideas, better metadata, increased access options, and greater data security. On the other hand, duplication can make tracking the data lineage more complicated and can be a waste of resources.

The user community is broader and more diverse than the community for which it was originally planned. Facilitating the access and understanding of data by interdisciplinary and non-technical users should be a priority for the data centers.

What kind of infrastructure/technology would make it easier for users to access and exploit data? What search tools would be useful for isolating the requested data and obtaining them in a useable format? Is a common format (e.g., HDF) the right answer, or are there better formats for archiving, storage, and transmission?

Better dataset visualization tools would ease user access.

Using translatable structured formats would be a logical way to allow both independence and interoperability. The working group noted that XML, which was developed for the World Wide Web, might be a good starting point for standardizing metadata formats.

Libraries might be a key new player in the digital world as archival entities for global climate change data. Libraries

have a long tradition of preserving and indexing information, and many are expanding their scope as providers of information science services. University libraries could cache and copy datasets and enable users to access relevant information at other libraries.

How can the use and refinement of data be tracked? How can pedigree effectively be made a part of the data? How can the quality control and pedigree of data products best be assured?

It is not enough simply to document the data. Obtaining historical perspective from data requires the ability to query the entire sequence of data use and refinement. Some technological solutions exist, but it is still an active area of research.

What are the greatest problems getting data into data centers?

Shortage of ingest staff and difficulty with maintaining state-of-the-art hardware and software are challenges.

File transfer protocol (FTP) is not always an effective means of transferring data, especially if data volume rates are high.

Computing or creating metadata is the most time-consuming and labor-intensive part of data processing and may create a bottleneck. Some metadata could be computed and stored automatically when data are processed. However, before this can happen, it has to be determined which and how metadata should be stored. In addition, software that can extract and store the metadata must be developed. Other metadata cannot be automatically computed and stored but must be identified, created, and entered by human experts. Even in those cases, software should be developed to aid the human expert.

The practice of retaining versions of data at each stage of processing places heavy demands on storage space. Users should be able to reproduce different versions from archived raw data; however, hardware changes make it difficult, if not impossible, to do so.

Overall, the working group concluded that the main limitation in data ingest is not technology but the human expertise and time for building knowledge into datasets. In addition, although small datasets, such as those resulting from observing stations, are often quite useful, they are time consuming to maintain. Finally, better coordination and communication among agencies, data producers, data archivists, and producers of metadata would improve data ingest.

Data Distribution and Processing Working Group

The group on data distribution and processing examined the different data processing strategies at the NOAA, DOE, and NASA data centers. The participants were asked to consider how data can be accessed efficiently by increasingly diverse users once data become part of the national archive.

How do we handle both data- and compute-intensive processing? What about reprocessing? What about supply-driven versus demand-driven processing? Is it advantageous to do all processing on demand?

> Data- and compute-intensive processing are typically handled separately, so it is not necessary to address both simultaneously. Reprocessing demands vary by data type and depends on the information needs. Reprocessing is done when there are new data or models and thus a chance to improve the usefulness of the data.
>
> A data center's decision to adopt a supply-driven or a demand-driven data processing model reflects the scientific and economic needs of its user base. Supply-driven data processing results in high data quality and availability at the cost of having to carry out continuous high-volume data reprocessing. Under a demand-driven model, only those data that a user requests are processed, resulting in lower processing and storage costs.

Can this processing be distributed to commodity-level computing equipment?

Although commodity hardware and software are easily applied to generic processing, widespread adoption of commodity solutions is hindered by data center needs for high bandwidth, fast processing speeds, and specialized error handling capabilities.

What technologies could make data distribution more efficient? Are there efficient and globally applicable subsetting tools that would dramatically decrease distribution costs while simultaneously simplifying exploitation of data?

Few users need all of the data in a database. Rather, they want to extract only the data relevant to their application (i.e., subsetting). Consequently, data subsetting is a necessary component of efficient data distribution. Although subsetting tools, such as all-disk storage and database technology, are available, the lack of familiarity within the file-centric science community has hindered their implementation.

How can access or resource restrictions be managed?

Restrictions can be managed by limiting consumption via charges based on resource usage, such as media use, consulting time, or data volume. Such charges must be in accordance with U.S. data policy (i.e., OMB Circular A-130 [OMB, 1996]).

REACTION PANEL SUMMARY

On the second day of the workshop, a panel of workshop participants representing data center managers, agency sponsors, data users, and the information technology industry was asked to react to the previous day's working group presentations. Panelists were asked the questions given in italics below; responses are summarized with each question.

What is your reaction to the first day's deliberations? Are we on the right track?

The panelists noted the following:

- Improvements for user access are needed (especially for interdisciplinary and nontechnical users).
- Data centers should focus on tools for finding data and for decision making.
- Technology is but one of the challenges facing data centers.
- Some of the technological challenges facing data centers have already been addressed in the information technology fields for other applications.
- Humans are the limiting factor in adopting and adapting to new technological capabilities.

What technologies can be resource multipliers (near versus long term)? If there's one you'd apply in the short term, which would it be?

Techniques for tracking, searching, and sharing metadata, especially through standardizing the use of a format such as XML Schema, would be a substantial benefit for data search and access. In addition, on-line datasets and databases and market-driven technologies have great potential applications in data center operations.

What have we missed?

Data centers and their sponsoring agencies must still consider the interaction between people and technology, rather than simply focusing on technology. One panelist suggested that it would be useful to improve the way available resources are promoted and communicated to users. Data centers should define the true metric of their performance carefully: better user services, decreased costs, increased number of users?

Appendix E

Glossary

DATABASE SCHEMA A definition of the table structures and data types used to store data in a database. All the data in the database adhere to the database schema.

EXTENSIBLE MARKUP LANGUAGE (XML) Provides its users with capabilities for defining sets of tags that can be used to mark up documents with information about the meaning of the data contained in the documents.

FILE TRANSFER PROTOCOL (FTP) DOWNLOAD A user is provided access to data stored on the data center computers, allowing copying of files.

INTERFACE Means for users to interact with computing and database systems.

INTEROPERABILITY The ability of a system or a product to work with other systems or products without special effort on the part of the customer.

LINEAGE Metadata describing where reused structures and types of data came from.

METADATA Data about data. Includes data about how, when, where and by whom a set of scientific data was named, structured, represented, collected, calibrated, stored, processes, exchanged, etc. Also, data about data distribution, data statistics, data usage, etc.

NETWORK BANDWIDTH The capacity to move large data files electronically.

OPEN-SOURCE Software with its source code made available without any restrictions on its redistribution or reuse.

PARSE To process a data stream and recognize the individual data in the stream by using the schema for the data stream.

PETABYTE A measure of memory or storage capacity and is 2 to the 50th power bytes or, in decimal, approximately 1,000 terabytes.

QUERY Language entered by a user of a search engine or database to find data or information.

RANDOM ACCESS To access individual data directly, as on a CD, computer disk, or computer memory, rather than scanning through data to find individual data, as on a tape.

SCHEMA A definition of the structures and representation types of a collection of data.

SUBSETTING The process of computing and making available a subset of the data in a data stream that satisfies a set of specific conditions. This subset may be delivered immediately or may be stored and subsequently delivered when requested by a user.

TERABYTE A measure of computer storage capacity and is 2 to the 40th power or approximately 1,000 billion bytes (i.e., a thousand gigabytes).

XML schema A definition of the structures and representation types of data in a collection of XML documents. All the XML documents in the collection defined by the XML schema adhere to the XML schema.

XML Schema The language in which XML schemas can be defined. Also used to denote the self-describing XML schema that defines all possible XML schemas, including itself.

Appendix F

Acronyms

CDIAC	Carbon Dioxide Information Analysis Center
CESDIS	Center of Excellence in Space Data and Information Sciences
CIESIN	Center for International Earth Science Information Network
COARDS	Cooperative Ocean/Atmosphere Research Data Service
DAAC	Distributed Active Archive Center
DAP	Data Access Protocol
DODS	Distributed Oceanographic Data System
EDC	EROS Data Center
EOSDIS	Earth Observing System Data and Information System
EROS	Earth Resources Observation Systems
ESTO	Earth Science Technology Office
FTP	file transfer protocol
GSFC	Goddard Space Flight Center
HDF	hierarchical data format
HTTP	hypertext transfer protocol
IT	information technology
LaRC	Langley Research Center
MD5	message-digest algorithm 5
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NCDC	National Climatic Data Center
NEIC	National Earthquake Information Center

NEpster	NPP-ESTO Portal for Science, Technology, and Environmental Research
NetCDF	Network Common Data Form
NFS	network file system
NGDC	National Geophysical Data Center
NOAA	National Oceanic and Atmospheric Administration
NODC	National Oceanographic Data Center
NPOESS	National Polar-Orbiting Operational Environmental Satellite System
NPP	NPOESS Preparatory Project
NRC	National Research Council
NSIDC	National Snow and Ice Data Center
NSSDC	National Space Science Data Center
ORNL	Oak Ridge National Laboratory
SAR	synthetic aperture radar
SDP	science data production
SEDAC	Socioeconomic Data and Applications Center
SQL	Structured Query Language
USGCRP	United States Global Change Research Program
XML	eXtensible Markup Language