Solving the Global Change Puzzle: A U.S. Strategy for Managing Data and Information. (1991)

Pages
62

Size
5 x 9

ISBN
0309296838

Committee on Geophysical Data; Commission on Geosciences, Environment, and Resources; National Research Council

Solving the Global Change Puzzle

A U.S. Strategy for Managing Data and Information

A Report by the
Committee on Geophysical Data
Commission on Geosciences, Environment, and Resources
National Research Council

National Academy Press
Washington, D.C., 1991

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 report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of 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. Frank Press 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. Robert M. White 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. Stuart Bondurant is acting 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. Frank Press and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

Support for this project was provided by the Department of Agriculture, the Department of Defense (U.S. Navy), the Department of Energy, the National Aeronautics and Space Administration, the National Oceanic and Atmospheric Administration, the National Science Foundation, and the U.S. Geological Survey.

Copies of this report are available from:
Committee on Geophysical Data
National Academy of Sciences
2101 Constitution Avenue, N.W.
Washington, DC 20418

Printed in the United States of America

Committee on Geophysical Data

Chairman:
Ferris Webster, University of Delaware

Committee Members:
Shelton S. Alexander, Pennsylvania State University
Peter Cornillon, University of Rhode Island
Murray Dryer, Space Environment Laboratory, NOAA
Joan Feynman, Jet Propulsion Laboratory
Glenn R. Flierl, Massachusetts Institute of Technology
Robert E. Gold, Johns Hopkins University
Dorothy K. Hall, NASA Goddard Space Flight Center
William J. Hinze, Purdue University
Roy L. Jenne, National Center for Atmospheric Research
Kenneth C. Jezek, Ohio State University
Robert L. McPherron, University of California, Los Angeles
Richard J. Reed, University of Washington

Liaison Member to NRC Committee on Global Change:
Francis P. Bretherton, University of Wisconsin, Madison

NRC Staff:
Pembroke J. Hart, *Staff Officer*
Lorraine W. Wolf, *Staff Officer*
Charlene E. Anderson, *Administrative Secretary*

Board on Earth Sciences and Resources

Commission on Geosciences, Environment, and Resources

Contents

Preface

The U.S. government is giving unprecedented priority to the creation and implementation of the U.S. Global Change Research Program. The plans for this program, as first described in the federal Committee on Earth Sciences report entitled *Our Changing Planet: The FY 1990 Research Plan* (Committee on Earth Sciences, 1989a), recognize the critical importance of data management.

In developing program plans, the involved federal agencies created a group to look at data and information management needs—the Interagency Working Group on Data Management for Global Change (IWG). The IWG's work has addressed operational, technical, and policy questions of data management, broadly and in detail. In 1988, it turned to the Committee on Geophysical Data (CGD) for guidance and advice. The CGD membership represents all relevant geophysical disciplines. Its members are nominated by National Research Council (NRC) boards and committees dealing with geophysical areas. In addition, the CGD has a formal link to the NRC's Committee on Global Change. Members of the CGD have met regularly with members of the IWG, and the two groups have had representatives at each other's meetings.

In response to the IWG's request, the CGD addressed the basic principles of data management for the U.S. Global Change Research Program. In so doing, it took into account the extensive work of the IWG (Committee on Earth Sciences, 1989a, Interagency Working Group on Data Management for Global Change, 1990), the recommendations in the report of the National Aeronautics and Space Administration (NASA)-sponsored Earth System Sciences Committee (NASA Advisory Council, 1988), the findings of the NRC Committee on Global Change (1989, 1990a), and the work of the International Geosphere-Biosphere Programme of the International Council of Scientific Unions (unpublished report, 1988).

Summary

Successful management of data and information is critical to the success of the U.S. Global Change Research Program (USGCRP). Present data management for scientific research is barely adequate. The USGCRP requires a new system of data and information management to meet its needs. By "system" we mean the interconnected process of gathering, processing, evaluating, archiving, and distributing of data, products, and information.

The existing environmental data management components have served reasonably well. However, many components necessary to meet the data and information challenge of the USGCRP are either inadequate or wholly lacking. New initiatives, such as the National Aeronautics and Space Administration's Earth Observing System, will generate an unprecedented amount of data for global change research. Existing components must be improved, restructured, or replaced. New components must be created, with existing institutions serving as models where appropriate. This report describes data and information management needs, reviews the status of existing components, gives a vision of how a global change data and information system might evolve, and proposes a strategy for future actions. That strategy is to build on the successes of the current data centers' infrastructure to achieve a data and information system in support of the USGCRP.

A global change data and information system should be based on simple principles that result from an analysis of existing data center activities. Successful data systems or centers are those that combine data management with scientific use of the data. Users contribute to the development of the data system and provide ongoing feedback. A successful system involves the scientific community at all stages of development and operation.

This report concludes that scientific community support for and participation in a global change data and information system are

critical. Without active scientific support, any data and information system is unlikely to meet the needs of the USGCRP.

The key points made in this report and possible actions to implement them are as follows:

1. Investment in data and information management should be visibly driven by and accountable to the scientific objectives of the USGCRP. As actions to implement this strategy, the system should:

- Link data management with specific science projects.
- Identify responsibilities of individual researchers.
- Provide incentives for scientists producing validated datasets to submit them to archives. (Journal publication of datasets may prove effective.)
- Create integrated product teams involving a partnership between active scientists and data centers, responsible for all aspects of relevant data collection, quality assurance, product generation, product accuracy assessment, and product distribution.
- Provide for visiting scientists and postdoctoral positions at each of the major data centers on a continuing basis.

2. Data and information management must be adequately supported to ensure the success of the USGCRP. Existing national data centers are not adequately supported to service present users, much less the expanded user community and data needs that the global program will involve. Steps that might be taken are:

- Allocate, at a high administrative level, resources for data and information management as a substantial fraction of overall program cost.
- Provide direct support for the system. The first objective of the USGCRP is scientific understanding. Attempts to

support a data and information management system through user fees are likely to prove ineffective and counterproductive.

- Scale the elements of a global change data and information system to the resources available and required to carry out its functions. Technical approaches need to be well matched in capability and cost to the uses of the system.

- Seek a balance between the resources allocated to a given activity and the overall needs and priorities of the total program.

3. The existing system of national data centers and academic and project data units provides a useful starting point for a distributed data and information management system. However, the system needs to be improved to:

- Provide adequate support to serve the expanded user community and data needs that the global program will involve.

- Cover all the disciplines necessary to address the entire range of global change issues.

- Strengthen the links between these various data and information activities into an effective national system.

4. The global change data and information system should provide for an evolving data management system and an expanding user base. Chapter 5 of this report presents a vision of how such an evolving system might look. The system should:

- Minimize impediments to data access.

- Identify key elements of the present data system that can be implemented today but that will not preclude evolution

in data types, algorithms, and products. Those elements should address the following functions cited in Chapter 5:

- Accept the responsibility for stewardship of global change datasets.
- Be driven by and accountable to scientists involved in global change research.
- Support developmental experiments with prototypes that help define functionality, technical and organizational approaches, and resource requirements.
- Require all primary data gatherers to supply relevant global change data to the appropriate archival center on a regular basis.

5. The system that is created should ensure that scientists can use the datasets effectively both at present and in future decades. To achieve this result, the system should:

- Preserve the data.
- Maintain a long-term capability to permit a scientist to locate global change datasets of interest.
- Compile and preserve dataset documentation (also called "metadata") together with original datasets.

6. The preparation of products and analyses is a needed and valuable function that should be an integral part of an effective data and information management system. Many of the existing components do not carry out the additional value-added analysis and information functions necessary to address global change. A global change data and information system should:

- Prepare products and analyses both by the primary data gatherers and by scientists at data centers or at individual institutions.

- Ensure that validated products and their documentation go back into the data and information system.

1. Introduction

Purpose of this Report

This report describes a strategy to manage data and information to meet the needs of the U.S. Global Change Research Program (USGCRP). It responds to a request from the Interagency Working Group on Data Management for Global Change (IWG). Thus, an important part of the audience for this report consists of people in those federal agencies who will manage the USGCRP. This report is also addressed to global change research scientists. Their involvement and support in developing a data and information management system for global change research are critical.

The report is organized into four main chapters. Chapter 2, Information for Global Change, begins with a set of priority requirements for a global change data and information management system and discusses some of the functions such a system should have. Chapter 3, The Present System, reviews the current status of data management. Chapter 4, A Data Management Strategy for Global Change, gives the committee's views on developing a system to meet the requirements. Chapter 5, A Vision of a National Information System for Global Change, presents one possible approach to creating a global change data and information management system.

The report focuses on strategy, not on technical issues or implementation. It complements other documents that describe specific program and implementation plans for a data and information management system to support global change research.

The strategy described here is to build a national global change data and information management system on existing components, strengthened and supplemented to create a system that works to meet the needs of global change research.

Global Change Research

The United States initiated the USGCRP to elucidate the causes and effects of natural and human-induced global change. The fiscal year 1992 budget is estimated to be $1,185 million and further growth is foreseen (Committee on Earth Sciences, 1989b, 1990; Committee on Earth and Environmental Sciences, 1991). This program will generate unprecedented amounts of environmental information.

An important aspect of the program is the acquisition and analysis of large amounts of information. This information will allow researchers to address the four core questions of the program (NRC, Committee on Global Change, 1989):

- What forces initiate global change?
- How does the earth system respond to changes in forcing functions?
- How has the earth's environment changed in the past?
- How well can global change be predicted?

It is beyond the scope of this report to list all the data types that will have to be managed in a national global change research program; there are thousands. However, the science priorities defined for the USGCRP give some idea of the wide range of data which must be managed:

- Climate and hydrologic systems
- Biogeochemical dynamics
- Ecological systems and dynamics
- Earth system history
- Human interactions
- Solid earth processes
- Solar influences

Recently, the USGCRP adopted four high-priority Integrating Themes for Fiscal Year 1992 (Committee on Earth and Environmental Sciences, 1991):

- Climate modelling and prediction
  To develop an improved predictive capability of the earth as a coupled system with enhanced regional resolution, with initial priority given to the climate system.

- Global water and energy cycles
  To improve our understanding of the water and energy cycles by focusing on the role of clouds, the role of the oceans, the role of terrestrial ecosystems, and changes in sea level.

- Global carbon cycle
  To improve our understanding of the carbon cycle by quantifying the natural and anthropogenic terrestrial and oceanic sources and sinks of key carbon compounds.

- Ecological systems and population dynamics
  To improve the capacity to assess the effects of global change at regional scales. Specifically, to improve understanding of the responses of intensively managed and natural oceanic and terrestrial ecosystems to global change.

To meet these research needs of the USGCRP, data from land, ocean, and atmosphere will have to be managed. They may include biological, chemical, physical, geological, hydrological, sociological, economic, and demographic datasets. Proxy data, data from the fossil record, and extraterrestrial solar measurements will have to be managed.

Heightened scientific attention to potential global warming—climate change—from increasing atmospheric CO_2 concentrations and other greenhouse gases has focused public awareness on this and other global issues. Global research issues often have common and demanding properties:

- They have larger spatial and longer temporal scales than usually addressed by traditional research.
- They are truly global in scale and scope (e.g., problems related to climate, atmospheric chemistry, and global carbon and nitrogen cycles).
- They can be local or regional issues with important feedbacks to global systems (e.g., deforestation).
- They require multidisciplinary approaches to analysis and understanding.
- They require national and international understanding and agreement before solutions can be realized.
- They are driven by technology and population.

The Need for Data and Data Management

The USGCRP will make unprecedented demands for the assembly and dissemination of large volumes of diverse and interdisciplinary data and information. Measurements acquired on regional and world-wide scales must be merged with other, often dissimilar, data, to produce analyses and products. That process will identify global change, evaluate its impact on human activities, and set a course of action to mitigate harmful effects. Important subsets of these data must be evaluated for quality and documented, distributed, and archived. Contemporary and future researchers must be able to acquire and use these data in their analyses of global change phenomena. If the USGCRP is to be successful, a strong data management system must exist to support it.

Existing U.S. environmental data management units have served their purposes reasonably well. However, many components necessary to meet the data and information challenges of global change, are inadequate or wholly lacking. They must be improved, restructured, or replaced. New components must be created, with some

existing institutions serving as models. This report describes needs, points to components that work well or are inadequate, and proposes a strategy to build on success and to achieve an information system to support the USGCRP.

The strategy in this document is based on an analysis of current activities and a forecast of future scientific needs. Successful data systems and centers combine data management with scientific use. Successful data centers not only work *with* scientists but have their active support. Users support the development of the data system and provide feedback. This leads to the conclusion that a successful system must involve the scientific user community at all stages of development and operation. Without that support and involvement, the data system is unlikely to meet the needs of the program.

The efficient acquisition, quality assurance, documentation, distribution, and preservation of relevant datasets of all types are crucial to the success of the USGCRP. To meet this need, the existing system must evolve. There must be a new way of handling research data for global issues. The NRC Committee on Geophysical Data's report entitled *Geophysical Data: Policy Issues* (1988) dealt primarily with geophysical data, but its conclusion applies to all disciplines concerned with global change research:

> The quantity of geophysical data obtained...has increased dramatically in the past few decades. Collected often at enormous expense, these data represent a national resource that must be managed carefully to ensure that they are preserved and available when needed. However, because of a substantial increase in the amount and complexity of geophysical data being collected and in the demands for them, *the management policies and procedures that have been developed are no longer adequate* (emphasis added).

2. Information for Global Change

Global change studies are interdisciplinary and the issues associated with such studies are often long term. Data and information will be needed, not as an end in themselves, but as a means for gaining understanding of the global system.

Many requests to the data management system will be for derived products such as analyses or edited data collections in association with descriptive text or graphical material, rather than raw observational data. Thus, the system must provide such products (here described as information) as well as data. It must keep the raw data as well, in order to provide the material for future reanalyses. To fulfill the role of information management, it must play an active role in the generation, acquisition, quality control, dissemination, and retention of value-added products.

Priority Requirements

For most global change studies, regional and global data and information will be required. No one nation, agency, or institution will be able to gather the appropriate data without cooperation from other nations, agencies, and institutions. For the United States this means that any one agency will often need the cooperation of others to produce datasets for global change. Moreover, the U.S. Global Change Research Program (USGCRP) will require international cooperation and data exchange. A policy will be needed that ensures unrestricted exchange of global change data across national boundaries.

The principal scientific activities having data and information system requirements are as follows:

1. Long-term measurements and derivation of products. Long-term measurements of the global environment will form a core part of

the USGCRP. Products and information derived from the data should play a principal role in judging the state of global change. Measurement teams should be responsible for end-to-end integration from initial measurements to final products. The team should include data managers. The process should be forward-looking, with an emphasis on obtaining reliable measurements. It should include a review process that takes into account the need to document long-term global change. Examples of such activities are the long-term studies of surface temperature.

2. Process studies. A process study may have its own data management system. The relevance of such a project comes from the final step of codifying the understanding through new algorithms or new measurement techniques.

Additional funding outside the project will be needed to guarantee, with a scientifically knowledgeable representative of the archival system, the useful transfer and preservation of the project data. An example of a process study is the Tropical Ocean-Global Atmosphere experiment.

3. Model assimilation. Model assimilation involves the use of physical laws as represented in a numerical model to reduce a set of disparate but related data to a uniform representation with known levels of credibility.

The information content of the derived product comes in part from the data and in part from the nature of the numerical model that is used. The distinguishing characteristic in this process is the integration of disparate datasets from multiple data sources and frequently the need for those data to be available in near real time. Product archiving is the final data management activity. A critical factor is documentation of the model output.

4. Non-global-change data. A number of existing data sources and holdings are not obviously relevant to present high-priority global change science activities. Yet they are clearly a valuable resource with major potential value in the future. To increase the chances for wise stewardship, we need to set in place disciplinary panels to advise data

center managers on data acquisition and retention policies and priorities.

Data needed for broad global change research run the gamut from site-specific to truly global datasets. Disciplines involved include earth, atmosphere, and ocean sciences. Many datasets will serve multiple disciplines and will need multiple talents to assemble. Time series data of all types are used prominently in global change research to detect past and present trends. Proxy data may serve when direct measurements are impossible.

The need for precision is important, as some of the predicted changes will be small and will take place over long periods of time. Often datasets will be enormous, due to the resolution and spatial scale needed to address global issues. A common attribute of many of the datasets is that they must be accessible and usable by future generations. Analyzing global change is a long-term process.

The term "data archaeology" is sometimes used to refer to research done using old direct or proxy data that have been unearthed in archives. As distinct from data archaeology, which refers to numerical data recorded by humans, the natural environment may retain information (often called "proxy data") about global processes, including climate. Such information may be recorded in the geological record (glacier ice, for example). The natural record may be analyzed to infer and measure characteristics of past climate and to decipher the importance of decadal-scale present global change. Aspects of this activity affecting development of a global change data management system include:

- Preservation of large volumes of material (e.g., sediment cores and ice cores).

- Development of algorithms that construct measured, ordered sequences of physical and chemical variables (isotopes, dust content, for example) to a time series of environmental parameters (surface temperature).

- Development of criteria for subjecting samples to destructive analysis.

The global change data and information management system will be created with limited resources. The high levels of quality assurance, documentation, and long-term archiving described herein cannot practically apply to all research data now collected. Priorities must be set. Research data in the system must be reviewed. This review will have to be done by central coordinating committees and data centers working closely with the global change research community.

Functions of a Data and Information Management System

The prime function of any data and information management system is the stewardship of the data and information, with all its ramifications. The data and information needed for global change research are costly to acquire. Their safekeeping must not be left to chance.

Data system functions may vary depending on programmatic goals, attributes of the thematic data or programmatic issues, and researcher needs. Most components of a data management system will have these elements.

The functions of the system include, but may not be limited to, the following:

- Providing a programmatic focus for data management.
  A data management system component should focus the flow of information necessary to conduct global change research.

- Data identification and acquisition.
  The system should take an active role with scientists in identifying datasets useful for global change research.

- Standardization of procedures.
  Standards for quality assurance, documentation, and distribution should be similar among system components.

- Data quality assurance.
  High standards of data quality assurance must be employed to maximize the application of data to answer global change questions.

- Data preservation.
  Long-term stewardship of research data must be assured.

- Data documentation.
  Datasets must be completely documented (the documentation is often termed "metadata") to ensure their utility for present and future researchers.

- Selective data retrieval.
  It must be possible to retrieve selectively data relevant to a user's needs.

- Data distribution.
  Data must be easily accessible by the world research community with as few restrictions (including cost) as possible.

- Derived data products.
  The system must be able to integrate data within and across disciplines to create derived data products for use by the research community and policymakers.

3. The Present System

For more than a century earth scientists have made national and international arrangements for the management and interchange of data. These arrangements have evolved in patterns driven by the needs of the user community and can serve as a base to meet global change research needs. The present system has many strengths and should be more fully exploited for global change research purposes before new data management elements are created.

In recent years the need for effective environmental sciences data management has increased tremendously because of scientific requirements and technological advances in observing capabilities. However, the data management system has not kept pace with these advances, largely because data management for secondary users has low priority in the funding and execution of scientific research. The data management system is no longer adequate for current scientific activities, let alone for the unprecedented challenges posed by a global change research program. The following sections summarize some major issues to be addressed if the U.S. Global Change Research Program (USGCRP) is to build effectively on the present system of data management.

Will and Commitment

A major challenge facing the USGCRP is the development of the collective will and commitment to managing data and information properly. The critical problems of setting up a data and information system are often perceived as being technical. They are not; they are policy problems. There are indeed some important technical problems to be overcome to create the system. However, these are not as serious as the policy issues.

The financial and moral support of data management is inadequate. Federal agencies should support data management with adequate funds. The scientific community should support data management by being part of its development and operation. The enthusiasm of scientists for the data management part of research programs is tepid. As a result, federal agencies do not feel the pressure to act; budgets suffer. The funding allocated by the agencies for data storage and dissemination generally has been the weakest part of their research budgets.

Factors contributing to budget inadequacies include the frequent need to cover cost overruns in other parts of a project, unrealistically low estimates of data management costs in order to sell a project, or ignorance or apathy about resources needed for data archiving and preservation. Most scientists would rather see budgets directed to their perceived "real research" than to data management. Thus, federal managers who fund data and information management inadequately are not often called to account. Data management is "everyone's second priority." If a system for the USGCRP is to succeed, there is a need to show that data management is valuable.

Because of the lack of enthusiasm and of a policy for handling scientific data, data facilities in the federal government have suffered a general decline of funding relative to other research endeavors. The result has often been decreased efficiency of data processing and decreased data availability.

Existing centers offer potentially unique combinations of long-term continuity, general accessibility, data management expertise, and user orientation required to broaden the use of the data. The data management system for global change must build on existing national centers. Thus, the present underfunding of data centers, in particular, and data management, in general, must be remedied.

Data Centers

The current U.S. national data management system involves relatively large national data centers, specialty data centers not necessarily charged with long-term preservation, and many large collections at academic centers. The holdings of most of these centers

are available to users at large. In addition, there are many project data centers, at federal and academic laboratories, where often the data are generally available to only a small number of users associated with the project.

The system has a number of strengths. Data, in general, are not willfully destroyed or discarded. The variety of holding sites means that data are held where there is some measure of local expertise on the particular data class. A fraction of the holdings are exercised by scientific users and reworked into quality-controlled, multisource datasets that are returned to data centers. There is a widespread determination to create an effective system for global change research. The Interagency Working Group on Data Management for Global Change (IWG), created in 1987, is an ad hoc voluntary group with senior representatives from the federal agencies involved with data management for global change. The purpose of the IWG is to coordinate interagency data and information management. Currently the group is developing a coherent interagency plan for handling global change research data.

The system also has weaknesses. They include the lack of links, both managerial and operational, among the many components. This means that it is not possible systematically to find the holdings of one data center by calling another and that a project data center may not transfer all its expertise, documentation, and data to a national data center before terminating. There is a lack of interagency agreements and policy regarding the preservation and enhancement of data. A scientist has difficulty securing funding for reworking data in a manner that would improve their quality and usefulness for more than a group with narrowly defined scientific interest. The funding at national data centers has generally decreased in relative terms, while the numbers of datasets and users have increased. There is difficulty in funding "technology conversions" at national data centers, where data held on, say, a thousand decaying low-density tapes could be copied to ten new media such as optical disks or helical-scan magnetic tapes. Science users have little say or control over national data center operations. The data centers sometimes have difficulty obtaining usable data and documentation from scientists after a reasonable period of privileged use.

Data centers exist for many disciplines. For example, within the National Oceanic and Atmospheric Administration (NOAA), a set of environmental data centers is operated with the mandate for national support. The National Environmental Satellite, Data and Information Service (NESDIS) operates three national data centers: the National Climatic Data Center (NCDC) in Asheville, North Carolina; the National Oceanographic Data Center (NODC) in Washington, D.C.; and the National Geophysical Data Center (NGDC) and the associated National Snow and Ice Data Center (NSIDC) in Boulder, Colorado.

Other federal agencies have also established data center functions relevant to their particular missions. Some examples are National Aeronautics and Space Administration's (NASA) National Space Science Data Center (NSSDC) in Greenbelt, Maryland; the U.S. Geological Survey's Earth Resources Observation System (EROS) data center in Sioux Falls, South Dakota, and the NASA's Ocean Data System (NODS) in Pasadena, California. There are other smaller data centers operated or supported by the Departments of Interior, Energy, Defense, and Agriculture. The National Center for Atmospheric Research (NCAR) in Boulder, Colorado, maintains a Data Support Group. This list of centers is not exhaustive. Though they are not always identified as such, these centers are, in effect, de facto national centers for their scientific specialties.

In the geosciences an international network of World Data Centers (WDCs) has been in operation for more than 30 years. The WDC system, in which U.S. data centers play a major role, is an excellent foundation for international exchange of global change research data. However, the WDC system does not yet include data from several disciplines (e.g., biological data, socioeconomic data, atmospheric chemistry data) that are critical to understanding global change.

Recently, some data centers have been establishing stronger links with the research community. One such link worth noting is that between the NODC and the Scripps Institution of Oceanography. The Joint Environmental Data Analysis Center (JEDA) combines scientific use and quality control of oceanographic datasets at Scripps with archiving and distribution at NODC.

With the possible exception of the NOAA/NESDIS centers, the national centers have loose and informal linkages. Rapid computer

connections between data centers or between centers and researchers are sometimes minimal or nonexistent. Except for the NOAA/NESDIS centers, there is no formal coordination among the managers of data centers. The tendency for the holdings of the various centers to have evolved in disciplinary patterns without strong interdisciplinary links is therefore not surprising. However, an interdisciplinary approach will be important in meeting the goals of the USGCRP.

The centers have their data in a variety of formats. Many are high-resolution digital values in computer files, photographic images, handmade drawings, numerical tables, analog recordings, microfilm, etc. For example, seismic and meteorological data are quite different in character, especially in view of the meteorologists' extensive use of objectively analyzed gridded fields at regular time intervals.

The centers tend to serve a relatively narrow clientele that is generally familiar and comfortable with the major data types and formats of the discipline. Some important fields that will be involved in global change research do not have established data centers. For example, there are no recognized national data centers for ecological, biological, and geological data. These gaps must be filled if an effective global change data management system is to be established.

Effectiveness of the Centers

We begin with two comments:

1. The national data centers are staffed by dedicated data management professionals who have been limited by decades of second-priority budget status.

2. Members of the Committee on Geophysical Data have not had the opportunity to visit recently all of the centers. (More site visits and reviews are planned, at the request of the operating agencies.) The comments that follow are based on the records of many visits in earlier years, on presentations by data center personnel, on documentation provided by the centers, or on personal experiences in working with the data centers.

The national data centers are essential. However, it is the committee's opinion that the existing structure of the centers will not serve all the purposes of the interdisciplinary USGCRP. In addition to

discipline-oriented data services, there is the need for centers to provide issue-oriented information services. Existing centers or new information analysis centers must provide information and data to meet the needs of interdisciplinary global change research issues and not continue simply to treat data on a discipline basis.

Data center funding has been a significant problem. The impacts of funding limitations have been particularly severe at the data centers operated by NOAA.

Some centers are technologically behind—some perhaps by decades. As a result of a lack of the technical means to cope with the increasing size of recent datasets, some satellite and hard-copy data have been effectively lost during the past several years.

Some centers are perceived as doing a better job than are others of providing data and information for research purposes. Two data centers subjectively viewed as particularly successful by the science community are the Carbon Dioxide Information Analysis Center (CDIAC) of the Department of Energy in Oak Ridge, Tennessee, and the NCAR's Data Support Group in Boulder, Colorado.

Both the CDIAC and the NCAR Data Support Group have a strong linkage between data center activities and the scientific community. Dataset users have been involved in the development of the centers and provide ongoing feedback. Both centers operate under special conditions which are not applicable to many national data centers. Though both have limited value for global change research, we cite them as examples that could be used in creating a new system.

The CDIAC at Oak Ridge National Laboratory sees itself as an issue-oriented information analysis center. In this role it contrasts with the discipline-oriented data centers. The scientific issue it supports is research into environmental consequences, such as the greenhouse effect, that are potentially related to carbon dioxide. CDIAC is funded by the Department of Energy. It is located at a national laboratory, actively conducts research into many CO_2-related issues, and is run by a staff that includes data managers and scientists.

Unlike the national data centers, CDIAC's data holdings are not voluminous. They are chosen to be those considered most essential to the pursuit of CO_2 research goals. Datasets and information distributed by CDIAC go through extensive quality assurance procedures, often with the help of researchers in the field. The datasets are documented

with information needed to interpret the data by someone unfamiliar with its generation. Typical documentation describes the limitations of the data and comes bound with reprints of pertinent journal articles and reports.

CDIAC takes an active role in developing needed data and information. It fills a useful niche but cannot serve as the only model for a data center. It does not operate under the constraints of most of the national data centers, nor does it charge for any of its services. It does not comprehensively archive datasets in a given discipline; rather, it works with the user community to develop, to assure the quality of, and to document only those datasets selected as the most important in addressing its issue-oriented mission.

The NCAR's Data Support Group has served as a model of a successful disciplinary data system. This group is considerably smaller (four to six full-time personnel) than most data centers. It has benefitted—perhaps uniquely—from the computational facilities at NCAR, from the expertise and dedication of its small staff, and from its support by the NCAR and university community of atmospheric scientists. This group has been successful not only in distributing datasets to a broad range of users, but also in providing, usually informally, documentation and information on data availability. The NCAR holdings include a wide variety of raw and instantaneous station reports, time-averaged (e.g., monthly) station data, satellite-derived fields, operational analyses from various centers, value-added analyses (e.g., hemispheric surface temperature grids spanning a century or more), ocean surface datasets, and output from global climate models.

From the USGCRP perspective, NCAR must be viewed strictly as a disciplinary center. It has, to a large extent, been free to choose its holdings—although these choices often respond to user needs. The NCAR's Data Support Group functions in an environment surrounded by users. The group has kept its approach simple. It uses high-tech solutions only when necessary. These factors, which have contributed to the success of the NCAR system, can be regarded as luxuries that may not be available to all the system components of data management for global change. Nevertheless, the NCAR example shows that a data management system can work. It may provide useful input to the design of a much more complex system for global change.

Finding Data and Information

Not all significant public-domain data are in national repositories. This situation has developed for a variety of reasons: economic policy, research efficiency, and the personalities of principal investigators. Moreover, the secondary user is often unaware of the range of data available through the data centers and may be unaware even of the centers' existence.

Many scientists face major obstacles in finding out what pertinent data are available; in some situations, obtaining the data may be a practical impossibility. In other cases, inadequate dataset documentation makes personal contact with the primary user imperative.

There is a need for all centers holding data acquired through federal funding to provide well-documented information about the extent of their holdings and the accessibility of the data. Furthermore, data interchange between agencies will become a major issue as global change programs require data from ever-wider sources. The lack of interoperability between data directories is a serious deficiency of the current system. It must be addressed if the USGCRP is to draw on existing and future data holdings.

Users must have an easy method to access an on-line system to search for specific datasets of interest. In addition, a directory with information about data centers must be available.

A national data catalog system, the Global Change Master Directory, sponsored by the IWG and based on a similar NASA system, is under development. Nearly 2,000 datasets are already described in the system. In addition, a user who is searching the national master catalog can be automatically transferred to search more detailed catalogs and inventories at individual data centers.

User Participation Issues

Data Submission

The scientific research community does not participate uniformly in data management. This creates a many-sided problem of

accessibility. Valuable datasets remain in the custody of individual research groups or even individual investigators, with the consequent acceptance by them of the data-handling task. Some of these datasets are widely known and accessible; others are not.

There are many reasons for the failure of individual scientists to provide data to the data centers. Among these are a desire for exclusive access to data, the reluctance to divert time and effort away from research in order to clean up a messy research dataset, and unawareness of the existence of appropriate repositories or of the importance of depositing data in such centers.

How long recently collected data should remain under the exclusive control of the scientist(s) who collected the data is an issue being addressed by the federal agencies involved in the global change program and is discussed further in the next chapter of this report.

In general, the global change research community is not sufficiently aware of the importance of ensuring the availability of environmental data. Until this changes, potentially valuable datasets will continue to be lost as primary users retire, relocate, or move on to other projects.

Quality Assurance

Many data centers provide little or no quality assurance of their data holdings. The assumption is that the individual researcher has applied adequate attention to assuring the accuracy of the data. This is often not the case.

Scientific use of data is the best road to its quality assurance. Many data centers have little or no in-house scientific research and therefore are not in a good position to provide effective quality assurance. Sometimes the data management system unwittingly corrupts datasets.

The working scientist sometimes treats data supplied by a data center with ambivalence. While healthy skepticism by the working scientist is a desirable part of the data exchange process, the present system suffers unacceptably regarding dataset credibility. Data management policies need to include more specific arrangements for data validation so that data preserved for the future are of adequate

quality. Defining the meaning of adequate quality and establishing standards are, fundamentally, responsibilities of the scientific community. However, applying these standards is a joint responsibility of scientists and data system managers.

Documentation

Data are frequently separated from information about the data. This is an unfortunate by-product of the explosion of digital data and techniques for handling such data over recent decades. Documentation can permit the user to judge the reliability or value of a data product for a particular application. The same is true for original data in terms of calibration, quality control flags, and station histories. Such documentation should therefore be an inseparable part of the data. Examples of useful documentation include information about the algorithms used for a derived product, quality control procedures, comparisons with independent measurements, and reviews of the dataset by outside experts.

Dataset Evaluation

For global change research purposes, few criteria have been developed to guide the evaluation, retention, and purging of datasets. These activities involve assigning priorities and establishing thresholds of importance for archiving and retention. Evaluations of individual datasets are important for determining which existing ones may be most useful for global change research. The data centers should play important roles in this function.

By monitoring distribution and obtaining feedback from users of datasets, data center personnel should be able to compile a consensus of the user community on the quality of a particular dataset. At the very least, the feedback obtained by the data centers can serve as input to groups of experts assigned to make recommendations about dataset retention or purging. The feedback obtained by the data center also can benefit the scientists who originally provided the data if the data compilation is an ongoing effort. Although this activity

represents a key step in interfacing present and future data for studies of global change, it is lacking at many data centers.

Data centers can play such a role most effectively if scientific expertise exists at the centers. With few exceptions the data centers do not now have adequate scientific expertise to perform this function.

4. A Data Management Strategy for Global Change

The U.S. Global Change Research Program (USGCRP) requires a strategy to meet its data and information needs. This chapter discusses a strategy and makes recommendations to achieve a new approach.

Scientific Involvement

This report, as others before it (e.g., NRC, Committee on Data Management and Computation, 1982, the "CODMAC report"), describes the lack of scientific involvement as being a problem in data management. Scientists may be involved in fundamental ways. One scientist may "make" data (by collecting, controlling quality, and processing), while another may "use" data from the first scientist and/or from an operational source. In the process the investigator may find a problem with the data.

In a free, competitive marketplace, the choice of one supplier's goods over another's is the mechanism by which users reward good quality, efficiency, and low cost. But a scientific data activity cannot be judged by such economic yardsticks. In view of the limits of science budgets, any costs significantly higher than the cost of reproduction are self-defeating. A mechanism by which users can be assured of quality, efficiency, and low cost must be found.

Involving a sufficiently large sample of knowledgeable users in the funding priorities for data activities may be the new mechanism. The process might require that data activities be accompanied by a 3-year proposal which is reviewed by data management peers and by scientists who supply and request data from that activity. A standing oversight committee, akin to those that review academic departments every few years, could advise on longer-term plans and data acquisition

or deletion decisions. Members of the scientific advisory groups would work as advocates for the system with their scientific colleagues.

Some scientists are knowledgeable about certain data types and their applications. Data activities should collaborate with these individuals to exercise the data, to reorganize and document them, and to control their quality.

Any successful business understands its customers' needs, knows its product line, and chooses its location carefully. By analogy, a data and information management system should follow some of the same principles. For example, an oceanographic data center would benefit from being physically adjacent to an oceanographic research activity. In spite of the increasing electronic transfer of information, proximity is important for understanding needs and sharing experiences. Data center personnel must know how scientists work with the data and why they choose certain datasets over others. Visiting scientist programs at data centers, and visiting data center personnel programs at research institutions, should be a part of the activity.

Scientific participation and authoritative oversight may be the key to creating and maintaining an effective global change information system. However, achieving effective scientific involvement will not be easy. Unfortunately, much of the scientific community is not aware of the need to be involved in developing any unified global change data and information management approach. This attitude is part of a pattern: data management has long been considered a secondary aspect of research. Since data and information together will be such a critical element in global change research, a change of attitude is essential. Fortunately, there are many signs that this change is taking place, both by research scientists and within sponsoring agencies.

Active researchers must be participants in the process. They should define needs and create the framework for a data and information system to meet those needs. They should help establish procedures and data centers. It will not be enough for them simply to assent to what a group of data technologists are creating. They must be involved.

There must be incentives for researchers to be involved. For example, the system must respond to scientists' needs. It must be perceived as the optimal way to do research with data. A simple feedback process will have a beneficial effect. For example, data centers

should involve researchers working with the datasets in the development of the data system. When advice is sought and listened to, there is an incentive for involvement.

Not only should incentives be created, but existing disincentives should be removed. User fees above a minimal cost for reproduction for scientific use of data constitute an existing disincentive. The nature of global problems requires access to large datasets. If their cost makes this prohibitive, then exploratory research will be obstructed. The global change data system should have the lowest possible user fee structure. Data should be free wherever possible.

As one example, Landsat data are currently so expensive that the data are generally beyond the reach of the research community. Furthermore, only data from selected areas are acquired and archived regularly. Data from most areas in the world must be requested by a user or a Landsat scene will not be acquired. This is a major problem because we cannot always identify what data will be needed in the future for studies of change detection.

There may be innovative ways to recognize data contributions. The Committee on Geophysical Data has discussed the possibility of creating a CD-ROM "journal," or of refereeing and referencing dataset contributions, as is done with scientific contributions. However, there is a strong community reluctance to recognize the preparation of a scientific dataset as being equivalent to the preparation of a scientific paper. There are encouraging signs that this attitude is changing: the *Journal of Geophysical Research (Space Physics)* has begun soliciting "brief data reports" of datasets that have been submitted to national data centers. Referees are asked to use the protocol to access the data and to comment on the data's relevance and quality.

Creating a New System

A new system for data and information management begins with the establishment of objectives. What datasets are needed to describe global change? What are the highest priorities? Setting objectives and priorities goes beyond data management. The fundamental scientific design of the USGCRP should include setting objectives for the data and information which will be needed. These

objectives must be set with data management input into the scientific process.

Any new system should build on existing elements. Thus, although new elements will be needed, we must make sure that the existing components work appropriately. Then we can move on to create the more complex system. This represents a challenge: can we make the existing elements work? As has already been noted, making them work is not a technical challenge but one of will and resources.

Design for Evolution

New systems must be designed for evolution. History shows that data storage systems, concepts, formats, and computing capability have changed on time scales of 2 to 5 years. It is unlikely that a data system defined today will remain constant for the next decade. The design problem is to develop systems which are flexible enough that data and information about the data can be readily saved for the next 100 years or longer, despite changes. Methods have already been implemented which permit relatively easy migration of data from one medium to another. Systems must be defined to accommodate diversity (e.g., a format designed 10 years before). Metadata must be structured so that they can be readily used by future as well as present systems.

Agencies must accept the responsibility of providing for the stewardship of the data they generate. Data management should be considered at the outset of every project, explicitly defined, and adequately budgeted for the life of the project. Arrangements should be made for the long-term archiving of the data.

Demonstrate Success

A new system should demonstrate success through practical prototypes. Confidence will be built by proving accuracy, by showing competence, and by producing some early products of value. This can be done by beginning feasible pilot projects of high importance. The Master Directory project sponsored by the Interagency Working Group on Data Management for Global Change (IWG) is an excellent

example. (This project is creating a high-level directory of datasets held by many agencies and institutions related to global change. Its success depends on information standards between centers and on operational computer network links.)

Data centers should be created for those disciplines important to global change research that are not included in the network of national data centers. This can be achieved through the establishment of new centers or by expanding the purview and resources of existing centers. A network of discipline-oriented data centers is a necessary component of the system to support global change research. However, because of the extraordinary information requirements that the global change program will make on data management elements, a new mechanism to handle the data will be necessary to augment the existing system.

Establish Information Analysis Services

Information analysis services for global change research should be established, either at existing data centers or as new information analysis centers (IACs). IACs should be issue oriented, complementing the existing data centers, which are principally discipline oriented. They should accept a much broader concept of information management than traditional data centers. They should support a broad user community, from individual researcher to agency policymaker. The information analysis services should extend, not replace, the existing system of data centers. In fact, it has been argued that only by maintaining a strong disciplinary component in most data centers will the requisite expertise be available for quality control. The provision of information analysis services to complement the discipline-based services should be a key component to the strategy of information management for the USGCRP.

Apply Appropriate Technology

New data and information systems should apply appropriate technology. This demands a choice between over-engineering and

under-utilization of technical advances. We face a future with an ever-increasing number of scientists and engineers sensing the environment and using the data, an ever-increasing spatial and temporal coverage of the sensors, and an ever-increasing number of channels on each platform. Coping with the data will require flexible computer-based solutions to store, find, and retrieve usable portions of the data.

Technologically appropriate solutions will likely not be permanent. New datasets that do not fit the current computer solution will appear, prompting a need for new algorithms.

It is impossible to give here specific guidance for solving these problems. However, we must ensure that they are effectively dealt with in a future data management plan.

Centralize the Data Directories

Locating data, both nationally and internationally, will be helped by the establishment of a centralized data directory. This directory will have information about information. It will be created by a joint effort, initially by the national data centers, and eventually with the help of international data centers. The data directory should be accessible electronically and user friendly. It should provide as much information about datasets as possible, including location; access policies and procedures; and information about the data's completeness, accuracy, general usefulness, documentation, and limitations. There should be no charge to use the directory except the telephone or electronic mail cost of connecting.

Upgrade Standards for Quality Assurance and Documentation

The quality assurance and documentation standards of datasets important to global change research must be upgraded. Quality assurance and documentation should be at the heart of a data management system supporting the global change program. Only after extensive testing by independent reviewers should important global research datasets be considered accurate. Depending on the data in-

volved, the effort can be extensive and often the single most expensive step in processing a dataset for distribution. This process, analogous to independent peer review of a journal article, maximizes the integrity of the information. It is necessary for the USGCRP. Future research and policy decisions will rest in part on these important datasets.

Data documentation must pass the well known "20-year test." That is, will someone 20 years from now, not familiar with the data or how they were obtained, be able to find datasets of interest and then fully understand and use the data solely with the aid of the documentation archived with the dataset? This is a tough test, yet one that must be passed for many of the data collections if long-term global environmental programs are to be successful.

Documentation must do more than describe the values represented in each field and the format information that is needed to read the data tape. It must fully document the dataset from all possible points of view. At a minimum, dataset documentation should contain:

- Identification of contributors
- Background information
- Scope and purpose of the program
- Data collection procedures
- Station history
- Description of instrumentation
- Definition of calibrations applied
- Full variable definitions
- Definition of calculated variables
- Description of adjustments
- Quality assurance at the center
- Modifications made at the center
- Limitations of the data
- Systematic and random errors
- Data transport verification statistics
- Full or sample listings
- Input/output routines on the transport medium
- References

Complete data documentation is a crucial portion of data processing and, along with quality assurance, will be the heart of successful data management activities for global change.

Encourage Data Management Research

Data management has developed in large part through the efforts of its practitioners. To meet the needs of global change research, a major research effort directed to environmental data management that involves computer scientists, data managers, librarians, archivists, and user scientists is needed over the next decade. Issues to address include the following:

- The design of an indexing system tied to the nature of the database.
- Provisions for browsing through complex datasets from a distant work station.
- Increased capability for visualizing databases.
- Principles for using compressed data in databases.
- Guidelines for the length of time data should be kept in "live" databases.
- Software design for the management of large databases.
- Application of expert systems to the management of large databases.

Important to the discussion here is NASA's plan for an elaborate system to handle data generated by the Earth Observing System (EOS). The EOS Data and Information System (EOSDIS) is being planned in parallel with the EOS flight mission to ensure that scientists will be involved not only in the analysis of the EOS measurements, but also with the storage and archiving of EOS data. EOSDIS is the largest single component of the federal data and information management system being created for the USGCRP and, as such, is important to this discussion. EOSDIS represents a major step in data and information management design in that it will handle extremely large databases. It has received advice from many groups and individuals, such as the NRC Panel to Review NASA's Earth Observing

System in the Context of the USGCRP (NRC, Committee on Global Change, 1990b) and NASA's Science Advisory Panel for EOS Data and Information (NASA, 1989). An excellent report by Dutton (1989) proposes EOSDIS design concepts. Although an analysis of such a complex system as EOSDIS is not appropriate within the scope of this report, the groundwork laid by EOSDIS will be significant for the development of a broader national data and information management system for global change data.

A Federal Data Policy

The USGCRP will depend on scientists sharing their data with each other. The timely submission of data to national centers requires a policy to ensure it. The policy must recognize the needs of principal investigators to protect their intellectual investment and to encourage their continued efforts to collect useful global change data.

The Ocean Sciences Division of the National Science Foundation (NSF) has long been a leader in maintaining a data policy. Researchers supported by the division must agree at the outset to share their data within a reasonable time. When the Committee on Geophysical Data (CGD) began its review of global change data and information management, it regarded the text of the then-current NSF Ocean Data Policy as a model of what should be done for all U.S. researchers involved in global change research.

The IWG has developed a *Policy Statement on Data Management for Global Change*. The participating agencies have endorsed the policy as part of the work of the Committee on Earth and Environmental Sciences.

A National Information System

A system to meet the data and information needs of the USGCRP should evolve step by step, based on scientific requirements. Until scientific needs are clear, it is unwise to develop elaborate technical plans. The CGD has included a possible implementation

approach in Chapter 5, along the lines of the strategy suggested in Chapter 4.

5. A Vision of a National Information System for Global Change

In the previous chapters of this report we listed some weaknesses of the current data management system and gave a partial list of the elements that a new system must include. In this chapter we sketch one possible realization of a data system that appears to satisfy all the concerns listed in the previous sections. The advantage in presenting such a vision is that in so doing contradictions, redundancies, or missing elements are more readily identified. However, we stress that other realizations are possible.

We imagine a national information system (NIS) for global change built on existing systems, so that it can take full advantage of the strengths of existing centers, datasets, software, and expertise. Our approach does not require extensive reformatting or refiling of data nor the replacement of existing systems; rather we describe enhancements, methods for increasing the scientific content of the datasets and the overview of the system's operation, development of coordination and data exchange procedures, and improved user services. The approach given here should be consistent with a manageable and affordable development effort.

Hierarchical Structure

We envision a "virtual system," which, as described below, implies that the user sees a uniform and coherent structure. In actuality the system will be built by many participating organizations out of many different parts on different computers using different software. An additional layer of software and/or hardware will be used to tie together the pieces, so that the user can navigate through the entire system virtually as if it were a single entity.

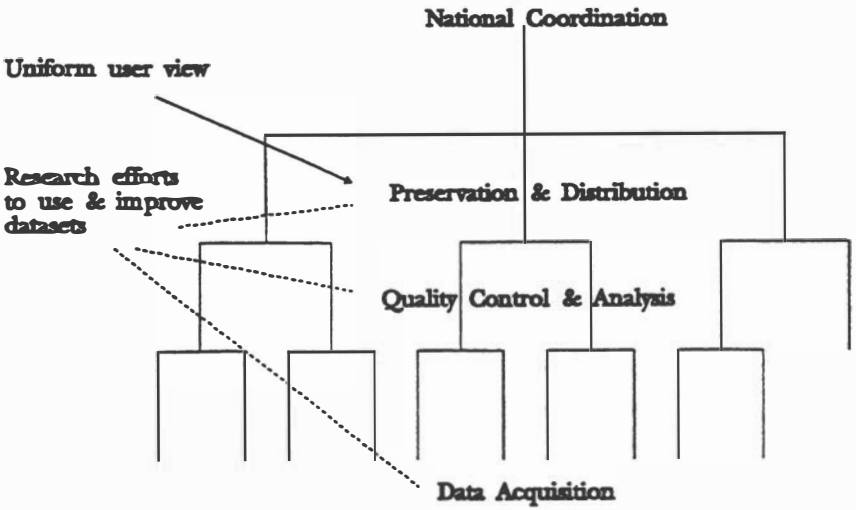

Figure 1

The National Information System (NIS) for global change as it might be organized in a four-level hierarchy:

1. Data acquisition level.
2. Quality control and analysis level.
3. Preservation and distribution level.
4. National coordination level.

Conceptually, then, the user sees the NIS as organized in a four-level hierarchy, sketched in Figure 1:

1. Data acquisition level.
2. Quality control and analysis level.
3. Preservation and distribution level.
4. National coordination level.

The previous sections of this report indicate that many elements in the diagram are currently missing or do not function reliably. Many critical datasets are not gathered in a way that they can be synthesized into a global dataset. Many such datasets may not be available to the scientific community at large. The external quality control level may be missing or unsatisfactory, or it may not be possible to control the quality of the data adequately. The preservation and distribution levels may not receive all the required data and may fail to preserve it. The system as a whole may not adequately serve the scientific community, especially users who are not experts in the particular dataset. Often missing are the connections between scientists active in research, modelling, or data assimilation and the centers. This loop is essential to improve the datasets. Finally, the national coordination level is missing.

Data and information about data flow upward to the preservation and distribution level as shown in Figure 1. The feedback loops involve active research efforts and serve either to ensure quality or to identify errors or needed improvements for later uses of the data.

The levels in Figure 1 are not directly mapped to physical locations; some centers may carry out the functions of several levels, or one function may be parcelled out to several groups. An example of a multifunction center is the Carbon Dioxide Information Analysis Center (CDIAC), which includes data gathering, quality control, documentation, preservation, and distribution of data.

Some data, such as marine expendable bathythermograph data from the Pacific Ocean, are collected through an operational program or by individual scientists. These data are telemetered to the National Oceanographic Data Center (NODC), where the quality of the data is controlled for valid ranges. Then the data are sent to scientists at the Scripps Institution of Oceanography, where the data quality is again

considered and the data are documented. The cleaned-up products and documents are returned to NODC for preservation and distribution. The National Geophysical Data Center (NGDC) has similar arrangements with external scientists for some of its holdings. The planned Planetary Data System (PDS) has a different separation of functions: the National Space Science Data Center (NSSDC) preserves the data, while a central NASA/PDS node handles the distribution service, passing requests to lower nodes.

Functions of Each Level

Each level within the hierarchy has distinct functions as outlined below:

1. Data Acquisition Level:
   - Gather data with suitable instrumentation.
   - Process and calibrate data to an agreed-upon level.
   - Supply data and metadata to the quality control level.
   - Respond to queries and requests for more information or more detailed data either from the higher nodes or from users.

2. Quality Control and Analysis Level:
   - Gather data from one or more suppliers.
   - Check quality and consistency.
   - Prepare products such as data summaries, figures, and browse files.
   - Pass products to the distribution/preservation level.
   - Gather metadata.
   - Prepare catalog and inventory information and pass to higher nodes.
   - Work with scientists using the datasets for modelling, synthesis, and assimilation.

3. Preservation and Distribution Level:
   - Acquire catalog information, browse information, and validated datasets.

- Prepare directory information for local holdings and subnode holdings.
- Make information available to higher nodes.
- Work with users to supply metadata and browse information.
- Arrange delivery of datasets or facilitate delivery from any higher or lower node.
- Serve as the primary focus for user requests.
- Arrange for accounting of data costs and security.
- Aid in the development of user interfaces and browse capabilities.
- Establish feedback loops with scientists to ensure quality of the datasets.
- Arrange for long-term data preservation.

4. National Coordination Level:
   - Arrange for transfer of information and user requests between various preservation/distribution centers.
   - Define standards to allow heterogeneous elements to act as cooperating parts of a system.
   - Ensure that lower nodes are functioning properly and effectively.
   - Maintain an overall directory and lexicon (possibly distributed among the discipline nodes).

User View

An NIS user should be able to enter the system at any one of the distribution nodes and thereafter have transparent access to any information in the system. Requests would be transferred down to subnodes or up to the coordination node to be referred to another node. The user should see one user interface and deal with all the information as if it were stored in one coherent database.

By referring to "information" rather than "data," we emphasize the necessity that metadata accompany the data. Metadata allow a contemporary user in another discipline to find and intelligently use

the data, help future users to understand limitations of the data, and simplify the search and retrieval of data.

Another essential element of the NIS will be ongoing scientific research efforts connected with the global change datasets, both individually and in concert. A great increase in this kind of activity is needed, and the connections between researchers and data management activities must be greatly improved, recognized, funded, and (when appropriate) formalized. Advice and products from these groups must be incorporated into the information system.

A Virtual National Information System

Rationale

By a "virtual national system" we mean that the aggregate behaves as though it were a single system. The scientific and operational groups that collect the data, carry out quality control, preserve and distribute the data, and use the data for other scientific studies must have well defined interfaces with responsibilities toward the overall system. This must hold even though the elements involved may belong to different agencies and disciplines, may use different hardware and software, and may have existed for years or have been created only for global change research. Since each group can exchange information according to the specified protocols, the whole acts as a system that streamlines the flow of data from the collector to an interdisciplinary scientist or a global modeller.

A virtual system has many advantages. By incorporating existing components directly, the system gains immediate access to a tremendous amount of data and expertise. This in turn builds stronger ties between researchers and existing data centers, develops a national center for coordination, and strengthens mechanisms for acquiring and assembling the required datasets.

Software and hardware will be essential to automate searches and transfers of data and requests. This is the binder that holds the disparate elements together and, as such, must function smoothly and accurately. Agreements among agencies, organizations, and research groups on what information to exchange and how to do it will be

essential to avoid interruptions in the data flow. Only motivated people will make it all happen.

High-level software is needed in this virtual system. It should permit a query of the existing and newly developed database management systems to retrieve requested information and pass requests and information from one part of the system to another. An NIS user should not have to know about the different individual databases, data centers, agencies, and computers. Software and hardware developments can concentrate on defining protocols and standards (being careful not to make them limiting) and implementing the communications process.

Individuals familiar with the databases can be responsible for integrating them into the system. Since the databases used in the individual components are not altered, they retain optimization for local use. They simultaneously serve a larger community. Changes in local systems and the addition of new centers will not affect the NIS except perhaps for making some information temporarily unavailable. Likewise, new technology, both hardware and software, can be integrated into the system, in effect by replacing one of the modular parts. Personnel able to handle both the database and the scientific aspects of the work at each center can be identified and brought into the system.

Current understanding is not enough to design and build such a system in one centrally directed giant leap, even though the technology for the separate components appears to be available. The approach must be stepwise, using prototypes that are tested both for utility and for the potential to become operational in a cost-effective manner and that help define the next step.

Necessary Steps

To implement the system imagined here starting from our current situation, we would take the following steps, which are grouped into three areas:

1. Flow of Information:
   - Establish a plan for developing priorities for the datasets needed for the development of an earth system science.
   - Ensure reliable sources of documented data (this is what the U.S. Global Change Research Program initiative has emphasized).
   - Ensure quality control of the data and generation of products usable by researchers who may be unfamiliar with the details of the data collection equipment.
   - Ensure that data and data products reach a data center where they can be distributed and preserved in a timely manner.
   - Establish new data centers and groups as needed.
   - Ensure that data centers, old and new, preserve data, distribute them, and provide feedback mechanisms to assess the scientific quality of the data they hold.
   - Establish a structure to coordinate centers and overall activities and define responsibilities.

2. Scientific Involvement:
   - Establish procedures, funding, and incentives for researchers to work with data centers in exercising datasets, generating new products, updating quality assessments, and influencing the data center.
   - Set up scientific advisory and control mechanisms at each level.

3. Unimpeded Access:
   - Develop protocols and methods for obtaining a uniform user view of all of the information in the NIS.
   - Minimize paperwork and user costs.
   - Set up effective user support personnel and procedures.
   - Support research in scientific data management.

Flow of Information

One objective of the NIS development is to disrupt operations at existing data centers to the least extent possible. However, many current datasets do not supply full metadata; efforts in cooperation with interested and concerned scientists will be required to enhance current databases with the necessary metadata, calibration and validation information, references, and documentation. In the section Scientific Involvement, suggestions are given for specific recommendations that would help.

Each center must ensure that the operations at each level are carried out efficiently and accurately. The suggested structure requires a national center for coordinating the centers, which would allow data to flow from one center to another and then to the user. Such a coordinating center must be able to work effectively across national agencies, a tricky issue that we bring up but leave open.

Scientific Involvement

The NIS must establish links to and cooperative efforts with research groups. Data management must be perceived as an activity critical to a successful research program; incentives for participating in NIS prototyping and construction must be created. Experts on methods and calibration of datasets will be needed to annotate extant datasets. Expert data analysts are needed who can take advantage of the growing database and produce worthwhile products which can themselves become part of the system. Assimilation modellers are needed who can produce dynamically consistent and complete datasets.

Examples of specific steps that can be taken are as follows:

1. Establish postdoctoral programs and visiting scientist fellowships. These programs would bring a researcher to a data center for a specific data-intensive project, from which the data center would receive a specific product.

2. Fund academic groups through the data centers to perform quality control and documentation of selected data at their own institutions or to generate products based on data from the data centers.

3. Establish staff scientist positions at the data centers. This must be handled carefully, since a scientist remains a research scientist by publishing useful results in the open literature, not by performing administrative tasks and technical consulting.

4. Collocate data centers with research groups. The National Center for Atmospheric Research (NCAR) data support section is a good example: while it is well regarded for various reasons, the fact that data managers share building and computing facilities with a representative sample of their users, whose voice can be heard by the hierarchy at NCAR, provides a strong and immediate feedback loop to improve performance.

Scientific papers and publications from these groups will help add to a wider perception of the value of the NIS. While there are many journals where science can be published, a high-quality journal where the data products are identified and thereby published should be considered.

Advisory groups and users who find the system valuable in their own research will provide essential feedback and impetus for changes as well as inducement for more scientists to use the system. A system can be most elegant and beautiful yet fail if the users are not there. If NIS is used successfully, scientists involved in global change research will also be convinced of the value of expert management of and ready access to interdisciplinary information.

Unimpeded Access

Each subsystem (e.g., a solar-terrestrial data center) must be melded into the virtual system in the best possible way. This will require work at each site. To do this, the terminal interfaces currently supported on each system will have to be supplemented with an

interface to the NIS that can receive and transmit queries and information according to the protocols established through the coordinating center. Internal query and response standards or protocols to be used in the NIS must be developed.

Communications and control programs for various systems will need to be developed and installed on the hardware used by the different centers. Networks that can accommodate the work must be selected, and the systems must be hooked into them. Preliminary user interface(s) must be developed, tested, and used in demonstrations for the system.

The NIS must develop a strong and viable support system. This will require ample professional documentation of the standards and first-generation software. New technical experts will be needed to aid in defining specifications and to develop the prototype. These will include programmers and system designers, as well as data managers and scientists.

Knowledgeable user support people who are retained for long periods will be vital. Too many users have been discouraged in the past by difficulties in obtaining rapid and knowledgeable answers to queries, a job best done when the support personnel also appreciate the scientific content and meaning of a dataset. Here again, joint projects involving data center personnel and scientific users (rather than a simple service from the data center to the user) can ensure interested, motivated, and knowledgeable support personnel.

6. The Next Steps

Looking beyond this strategy, plans must be made for creating a global change data and information system. Defining the program needs for research data and information is the highest priority. This process has been started (NASA, Earth System Sciences Committee, 1988). However, a thorough examination of U.S. Global Change Research Program (USGCRP) objectives matched to the corresponding needs for data and information has yet to be done. This will be a difficult task. Researchers must work with data managers to do the job. The Committee on Geophysical Data is prepared to participate in the process.

Definition of needs should come before system design. System design is technically feasible and relatively straightforward. Thus, some such studies have already begun. However, it is probably unwise to do the relatively simple technical design studies before the relatively complex and difficult definition of needs has been tackled.

We must not lose sight of the objective: a global change data and information management system must provide the means for gaining an understanding of global change processes. No matter how technically advanced or powerful the data and information system might be, if it does not support the research, it will have failed.

The USGCRP is a scientific challenge. Designing a data and information management system to serve it is also a challenge which the scientific community can respond to and support.

References

Committee on Earth Sciences (1989a). *Our Changing Planet: The FY 1990 Research Plan*, Federal Coordinating Council for Science, Engineering, and Technology, Office of Science and Technology Policy, Washington, D.C., 118+65 pp.

Committee on Earth Sciences (1989b). *Our Changing Planet: A U.S. Strategy for Global Change Research*, Federal Coordinating Council for Science, Engineering, and Technology, Office of Science and Technology Policy, Washington, D.C., 38 pp.

Committee on Earth Sciences (1990). *Our Changing Planet: The FY 1991 U.S. Global Change Research Program*, Federal Coordinating Council for Science, Engineering, and Technology, Office of Science and Technology Policy, Washington, D.C., 60 pp.

Committee on Earth and Environmental Sciences (1991). *Our Changing Planet: The FY 1992 U.S. Global Change Research Program*, Federal Coordinating Council for Science, Engineering, and Technology, Office of Science and Technology Policy, Washington, D.C., 90 pp.

Dutton, J.A. (1989). The EOS data and information system: concepts for design, *IEEE Trans. Geosciences and Remote Sensing*, 27, 109-116.

Interagency Working Group on Data Management for Global Change (1990). *Recommendations from an Interdisciplinary Forum on Data Management for Global Change Report OIES-5*, Office for Interdisciplinary Earth Studies, University Corporation for Atmospheric Research, Boulder, Colo., 75pp.

National Aeronautics and Space Administration (1989). *Initial Scientific Assessment of the EOS Data and Information System, Report EOS-89-1*, Goddard Space Flight Center, Greenbelt, Md., 37 pp.

National Aeronautics and Space Administration Advisory Council, Earth System Sciences Committee (1988). *Earth System Science, A Closer View*, University Corporation for Atmospheric Research, Boulder, Colo., 208 pp.

National Research Council, Committee on Data Management and Computation, Space Studies Board (1982). *Data Management and Computation, Volume 1: Issues and Recommendations*, National Academy Press, Washington, D.C., 167 pp.

National Research Council, Committee on Geophysical Data (1988). *Geophysical Data: Policy Issues*, National Academy Press, Washington, D.C., 40 pp.

National Research Council, Committee on Global Change (1989). *Toward an Understanding of Global Change*, National Academy Press, Washington, D.C., 213 pp.

National Research Council, Committee on Global Change (1990a). *Research Strategies for the U.S. Global Change Research Program*, National Academy Press, Washington, D.C., 291 pp.

National Research Council, Committee on Global Change (1990b). *The U.S. Global Change Research Program: An Assessment of the FY 1991 Plans*, National Academy Press, Washington, D.C., 107 pp.

Abbreviations and Acronyms

CDIAC . . . . . Carbon Dioxide Information Analysis Center
CEES . . . . . Committee on Earth and Environmental Sciences
CGD . . . . . Committee on Geophysical Data
CODMAC . . . . . Committee on Data Management and Computation, Space Studies Board
EOS . . . . . Earth Observing System
EOSDIS . . . Earth Observing System Data and Information System
EROS . . . . . Earth Resources Observation System
IAC . . . . . Information Analysis Center
IWG . . . . . Interagency Working Group on Data Management for Global Change
JEDA . . . . . Joint Environmental Data Analysis Center
NASA . . . . . National Aeronautics and Space Administration
NCAR . . . . . National Center for Atmospheric Research
NCDC . . . . . National Climatic Data Center
NESDIS . . . . . National Environmental Satellite, Data, and Information Service
NGDC . . . . . National Geophysical Data Center
NIS . . . . . National Information System
NOAA . . . . . National Oceanic and Atmospheric Administration
NODC . . . . . National Oceanographic Data Center
NODS . . . . . NASA Ocean Data System
NRC . . . . . National Research Council
NSF . . . . . National Science Foundation
NSIDC . . . . . National Snow and Ice Data Center
NSSDC . . . . . National Space Science Data Center
PDS . . . . . Planetary Data System
USGCRP . . . . . U.S. Global Change Research Program
WDC . . . . . World Data Center