SCIENCE DATA STEWARDSHIP THROUGH A GLOBAL SCIENCE DATA NETWORK

David CLARK

ICSU/World Data Center Panel
NOAA/NESDIS/NGDC
325 Broadway
Boulder, CO 80302 USA

David.M.Clark@noaa.gov

The International Council for Science (ICSU) operates 49 World Data Centers (WDC) globally. When the WDCs were established in the early 1950's, the concept of global science was just beginning to be developed to address fundamental questions about the Earth system, e.g., plate tectonics, reversals of the earth's magnetic field and oceanic circulation. Global science today still addresses these questions, but many more have been added as our understanding of the Earth system increases. ICSU's WDCs have taken on the challenge to evolve into a new system of data centers to help address these new and daunting global issues.

The WDCs will become a Global Science Data Network (GSDN) which will incorporate new technologies, new types of data, new participants, and new organizational principles. A key strength of this new concept will be the emphasizing of human resources focused on science-driven data management. It will improve links between the scientific community and the WDC system, and will emphasize the preservation of scientific data and making the data more useful. To be successful, the GSDN must ensure that the data are exercised, reprocessed, and reused for both original and new applications; improved through periodic analysis and evaluation, and made accessible through new technologies and innovative techniques.

The GSDN will build on the solid expertise and experience of the existing World Data Centers. As in the current WDC System, the network will draw on the scientific expertise of its centers and foster scientist-to-scientist links. It will ensure the preservation not only of the basic scientific data, but also the information and knowledge developed as a result of working with the data. It will continue to support common data policies, assuring full and open access to data and data products to scientists in all countries of the world.

To begin to implement this new network, the WDCs will seek to expand the WDCs around the world; it will initiate the development of partnerships between established WDCs and appropriate organizations in developing countries; it will establish a WDC Technical Task Team to take the lead on the development of new and innovative data management techniques; it will begin to actively address end-to-end data management, long-term archival, and data rescue functions across the WDC system; and it will forge formal liaisons between the WDC System and key international programs and organizations.

The long-term preservation and effective utilization of scientific data are some of the daunting challenges that arise in the understanding of the Earth system. Maintaining a useful archive of data describing the Earth involves more than routine preservation of the media. Establishing a GSDN that actively updates the scientific content of the data as well as maintains the integrity of the archive is an important aspect of today's global science.

1. Introduction

The World Data Center System was established nearly 50 years ago to meet the needs of scientists in the International Geophysical Year. Since then, new centers and disciplines have been added, and new technologies have been adopted. Overall, however, the System has not yet reached its full potential to serve international science. It is envisioned that a stronger, more robust, and more comprehensive system, capable of supporting both basic and applied scientific research around the world, truly global in reach, and committed to its founding principles of free and open access to data, long-term preservation of data, and improvement in data quality will be initiated in the future.

A key element of this vision is the evolution of a “Global Science Data Network” from the existing collection of centers (WDC Modernization Task Team, unpublished). Such a network will certainly include new technologies, new data, new participants, and new organizational principles. A key element of this GSDN will be the commitment to scientific data stewardship. To make this work, a strengthening of the human resources focused on science data management will be needed. It will be essential to improve links between international science programs and the WDC system, and to provide the technology and training to facilitate state-of-the-art data management and research by scientists around the world, especially those in developing countries.

It is anticipated that the network will continue under the non-governmental auspices of ICSU, building on the solid expertise and experience of the World Data Centers. As in the current WDC System, the network will draw on the scientific expertise of its centers and foster scientist-to-scientist links. It will support the exchange of results and expertise across scientific disciplines and world regions. It will ensure the preservation not only of the basic scientific data, but also the information and knowledge developed as a result of working with the data. It will continue to support common data policies, assuring full and open access to data and data products to scientists in all countries of the world. It will enhance the efforts of the scientific community around the world to support sustainable development, focusing in particular on building scientific capacity in developing countries.

There are several aspects that are logically part of the Global Science Data Network. It should have a network of human resources with the skills and expertise in the relevant scientific disciplines. An infrastructure that supports the long-term operation of the network is necessary. Information technology should be incorporated that is appropriate for using the data in scientific applications. And lastly, a continuing commitment to scientific data stewardship is needed. The concept of scientific data stewardship, or science stewardship of data, is a key element of the GSDN and one that needs the most development by the existing system of WDCs. Scientific stewardship of data can be described as 1) building consistent and high-quality records of environmental observations with associated metadata; 2) partnering with the scientific community (and others) through provision of high quality data and services for use in scientific studies; and 3) use of new and innovative technology to provide access to data and products.

2. Global Science Data Network

As we begin the new century, many problems faced by our society remain and new questions are identified every day. The challenge for scientific community and for the World Data Centers is to begin to develop ways to begin to answer these pressing questions (Clark, 2001). The concept of a Global Science Data Network is a response to this challenge and is currently being formulated (WDC Modernization Task Team, unpublished). While concept is still in development, there are several key concepts that should be considered part of the GSDN.

2.1 Human Resources

The current WDC system is well represented in the developed world and continues to expand in many parts of the world such as China. The strong interest on the part of a number of institutions in becoming new WDCs is a sign of the continuing value of the system to the international scientific community. However, it is clear that expansion in some areas of the developing world such as Africa and Latin America is unlikely without active efforts on the part of the WDC system. In some cases, there may already be strong research institutions that have the capacity to become WDCs with only a small investment of time and resources. However, in many countries, the underlying scientific capacity remains weak and the resources available for data management and archiving are very limited.

The WDCs, individually and collectively, will work to build the capacity to manage and archive scientific data throughout the world. Local scientists have local knowledge and understanding that are vital to developing accurate understanding of local and regional environmental processes. Scientists in developing countries need to have access to, and control over, data and information resources relevant to their regions of interest not only to conduct their own research but also to work effectively with those concerned with global scientific applications. By working directly with such scientists to build their capacity to manage and archive data, the WDC system can contribute both to overall global access to important new sources of environmental and associated data—in areas where such data have been most lacking—and to the long-term development of a truly distributed system of scientific data archiving and dissemination built on solid foundations in every country of the world.

One method to develop this capacity to effectively manage scientific data will be initiated through partnerships with others in the GSDN. The WDC system is proposing a formal partnership program, which will encourage existing organizations, especially in developing countries, to work cooperatively with existing WDCs to develop databases, expertise (both scientific and technological), and an infrastructure to effectively manage scientific data. This partnering program can be focussed for a

specific region or to address a particular discipline. It is hoped that the first partners in this program will be announced in 2003.

2.2 Infrastructure

One aspect of a GSDN, which is needed in some form, is a commitment to a long-term infrastructure for operation of the network. The infrastructure needed to adequately maintain and distribute scientific data is not trivial. A commitment to long-term support by the organization is needed if the data are to be maintained well beyond its initial collection. As will be noted, use of the data years or decades after its collection is a key element of scientific data stewardship. This aspect of the GSDN can be somewhat daunting because long-term commitments are sometimes difficult to make and to implement. Some mechanism for a long-term commitment is needed whether it is through a governmental entity, university or Non-Governmental Organization.

2.3 Technology

There are currently 49 WDCs around the world. Interactions among existing data centers are high due to the long history of WDC cooperation and the growing international use of the Internet. In today's world, advancing technology has on the one hand enabled the WDCs to cope with large and varied amounts of data, but on the other hand has opened a technological gulf between many of the WDCs. The previously mentioned emphasis on capacity building with in the GSDN helps address this problem.

Modern technology can be used to bridge this gulf on several fronts. Rapid decreases in the costs of information technology and continuing expansion of networks and their bandwidth are making significant new capabilities available to WDCs of all sizes and in diverse locations. Often, the main barrier to adoption of new technology is not the up front cost of hardware, but access to the software, expertise, and training needed to fully utilize the new technology to meet local needs. World Wide Web applications are at the forefront of some of these applications. WWW mapping and exchanging data are commonplace. Innovations such as Napster and the more recent GRID technology promise that the pace of new technological applications will not slow in the foreseeable future.

A GSDN will use this technology in many ways. The establishment of "mirror sites" between WDCs is now almost commonplace. This technology not only allows better access to data in a particular discipline but also encourages exchange and interaction between geographically separated scientific communities. The development of common catalogs, standard metadata formats and other technology dependent infrastructure will be the foundation of the GSDN. The WDCs have already formed a Technology Task Team that is looking at these aspects of the GSDN.

2.4 Scientific Data Stewardship

The key element in implementing a GSDN is the commitment to scientific data stewardship. As noted earlier, scientific data stewardship is the building of consistent and high-quality records of environmental observations with associated metadata; the partnering with the scientific community (and others) through provision of high quality data and services for use in scientific studies; and the use of new and innovative technology to provide access to data and products (Diamond, et al., in press).

Most of the scientific stewardship of data will take place when the data are made available for users other than those who originally collected that data. Scientific data are collected by observing and monitoring systems for operational use, or by researchers to help answer specific scientific questions. Ultimately, the data set is "archived" for use by other scientists. A data set can be combined or compiled with other data sets to form a comprehensive database. The first premise of scientific data stewardship is that the data set must be the best quality as possible and the data sets well documented with complete metadata. Recalibration of data sets should be done when appropriate. Adding new data to archived databases often reveal inconsistencies in the data. Documentation of the data should be complete enough to allow the data to be used many years from its initial collection date.

Improving the quality of data sets, in addition to routine internal consistency checks, often entail that the data are used in new scientific applications, sometimes different than those for which the data were originally intended. These applications can be either at the data archive as data studies initiated by the scientific staff at the data center or co-operatively with partners in the scientific community. The use of archived data for new projects and programs is a key element of scientific data stewardship. As new data

are collected in the course of a program, these are added to the database, causing some data to be revised, updated or deleted, thereby improving the quality of the database.

The innovative use of technology in providing access to data goes beyond the use of information technology to formulate the basic structure of the GSDN. Today requirements for state-of-the-art tools for accessing and using scientific data are ever increasing. Moore's law says that processing speeds of computers are doubling every 18 months. Disk storage costs decrease 50% each year and storage capacity doubles every 24 months. Fiber optic network speeds double each year. Users of scientific data will be using methods like data mining, data discovery and distributed computing. Innovative technology like Napster (exchanging data, i.e. music, directly over the WWW without central control) and GRID computing (client computing with server-like control using the WWW) will be at the next wave of scientific applications. The incorporation of this new technology into the scientific stewardship of a database will be required to keep pace with the requirements of the scientific data users.

3. Conclusion

A key element of the Global Science Data Network will be making scientific data stewardship part of the concept plan. Science data stewardship has the potential to provide the framework that will allow use of scientific data for long-term studies of the Earth's systems. When incorporated into a GSDN, science data stewardship makes the GSDN more responsive to scientific users needs as well as provides that the data are of highest quality, appropriate and adequate for the applications and that the newest technology tools are used for access and use of the data. The GSDN will be unique in this aspect. It will have a network that is dedicated to the human resources as well as to the technological advances. But as a critical complement, it will focus on the scientific integrity of the data. This is the important aspect that scientific data stewardship brings to the Global Science Data Network.

4. References

D. Clark. *Global Science and ICSU's WDCs: Challenges for the New Century*. In: Abstracts of the IGBP Open Science Conference, Amsterdam, The Netherlands, July 10-13, 2001, IGBP Secretariat. (2001)

H. Diamond, J. Bates, D. Clark, R. Mairs, and G. Sharman. *Archive Management: The Missing Component.* Presented at Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data symposium held at Institut Aeronautique et Spatial Complexe Scientifque de Rangueil, Toulouse, France, 5-7 November 2002. (In press)

WDC Modernization Task Team. *Towards a New World Data Center System: Meeting Global Needs*. ICSU WDC Panel. (unpublished).