J2.3A NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION'S SCIENTIFIC DATA STEWARDSHIP PROGRAM

John J. Bates *
NOAA National Climatic Data Center, Asheville, North Carolina
Richard G. Reynolds
Office of Systems Development, Silver Spring, Maryland
Constantino Cremidis
Computer Sciences Corporation, El Segundo, California
Carlos Martinez
TMC Technologies, Fairmont, West Virginia

1. INTRODUCTION

NOAA's data stewardship program is a new paradigm in data management consisting of an integrated suite of functions to preserve and exploit the full scientific value of NOAA's environmental data. These functions are the proper archival of and timely access to data and metadata, the careful monitoring of observing system performance for long-term applications, the generation of authoritative long-term records from multiple observing platforms, and the assessment of the state of the atmospheric, oceanic, land, cryospheric and space environments. Successful implementation of NOAA's data stewardship program will ensure NOAA's environmental data are of maximum use to the Nation now and in the future. Figure 1 shows schematically the concepts of data stewardship applied to the suite of NOAA observational systems.

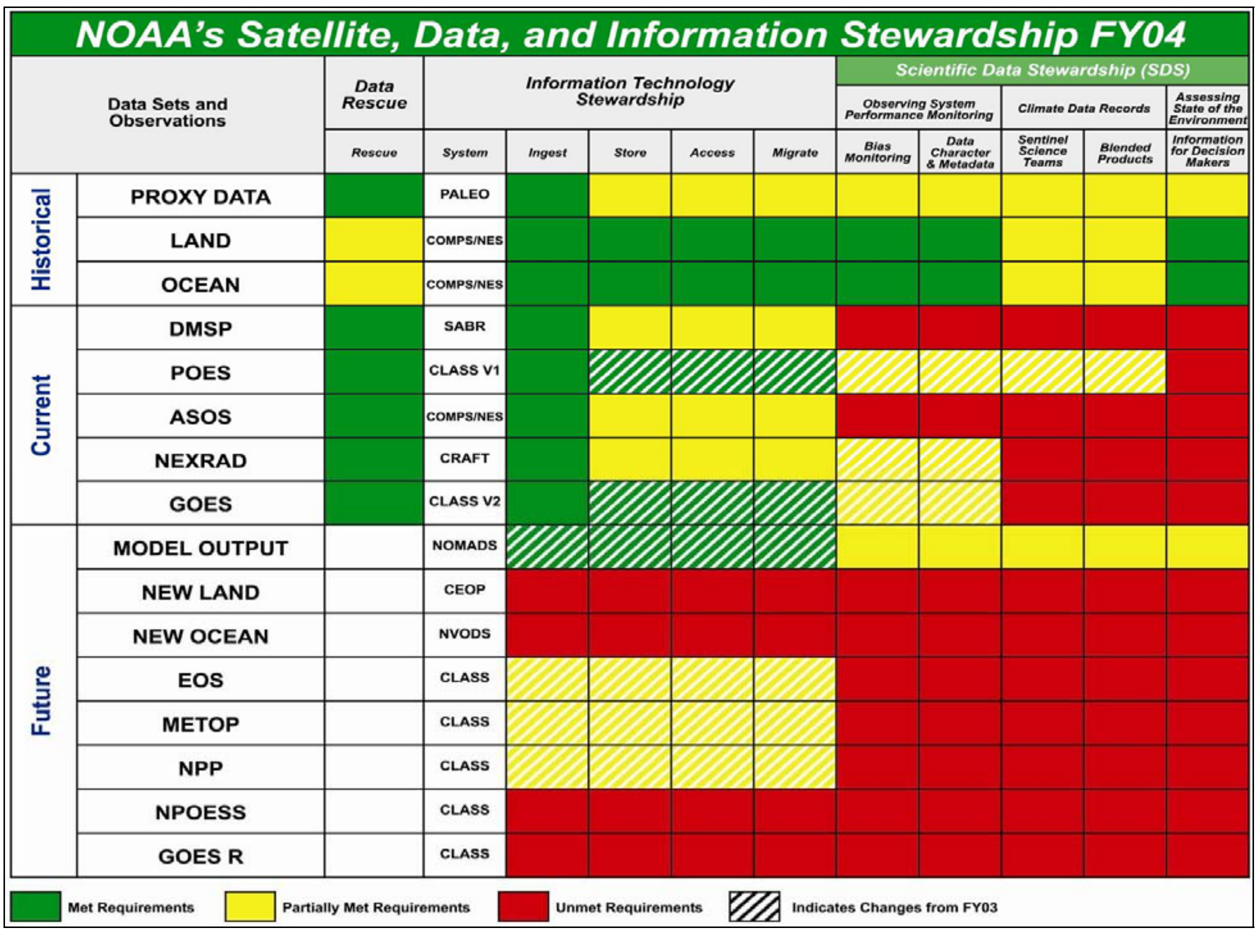

Figure 1. 'Stoplight' diagram of data stewardship activities

* *Corresponding author address:* Dr. John J. Bates, National Climatic Data Center, 151 Patton Avenue, Asheville, NC 28801; e-mail: john.j.bates@noaa.gov.

To promote full exploitation of the scientific value of NOAA’s data by current and future users, four functions, each with several constituent components, must be achieved. The first function is information technology (IT) oriented and is called the Comprehensive Large Array-data Stewardship System (CLASS). The CLASS function, insuring complete archival and access capabilities, requires that metadata, direct observations, and fundamental records from satellite and in situ platforms be comprehensive, complete, and preserved, in perpetuity. Open, efficient access to the metadata, products, and data streams must be insured, and data made available in useful formats. These functions and their components must at all times maintain strong emphasis on the physics of the instrument designed to measure an environmental parameter or parameters. Included in this IT component is a data rescue function to ensure preservation of past data sets, both hardcopy and electronic.

The next three functions are science-oriented and are therefore referred to as Scientific Data Stewardship (SDS). The first SDS function is to provide real-time monitoring of the observing system performance for long-term applications. Such monitoring requires establishing tracking tools necessary for the detection of changes in the observing system, as well as in the observation record. One example is the detection of small biases in the instrumental record. These biases can then be minimized or eliminated through efficient coordination with network operators.

The second SDS function is generating authoritative, long-term records. This function will preserve and enhance the value of the irreplaceable historical data by conducting rigorous data analyses and research to validate and improve these authoritative records, and by reprocessing and enabling others to reprocess the entire data stream from the fundamental measurements when scientific advances warrant it. SDS will use physics-based techniques to fuse together data from disparate observing systems like direct measurements from ground-based networks, as well as indirect measurements from remote sensing instruments on satellites. Note that these elements differ from those that describe operational data, namely timeliness and a minimal reliance on data from outside sources. The authoritative nature and vitality of these products will be maintained through periodic reviews.

The third SDS function uses the authoritative records to assess the current state of the environment and to put it in historical perspective. Long-term trends on local, regional, or global scales can be determined and estimated for the future. In addition, the authoritative records can be used to detect changes in environmental conditions between different time periods and different environmental regimes.

2. OVERVIEW OF THE COMPREHENSIVE LARGE ARRAY-DATA STEWARDSHIP SYSTEM

CLASS supports the NOAA mission to acquire, archive, and disseminate environmental data. NOAA has been acquiring these data for more than 30 years, from a variety of in situ and remote sensing observing systems operated by NOAA and from a number of its partners. NOAA foresees significant growth in both the data volume and the user population for these data, and has therefore initiated this effort to evolve current technologies to meet future needs.

NOAA's National Data Centers and their world-wide clientele of customers look to CLASS as the primary NOAA information technology infrastructure project in which all current and future large array environmental data sets will reside. CLASS provides permanent, secure storage and safe, efficient access between the Data Centers and the customers. The initial objective for CLASS is to support specifically the following campaigns:

- NOAA and Department of Defense Polar-orbiting Operational Environmental Satellites (POES) and Defense Meteorological Satellite Program (DMSP)
- NOAA Geostationary Operational Environmental Satellites (GOES)
- National Aeronautics and Space Administration (NASA) Earth Observing System (EOS) Moderate-resolution Imaging Spectrometer (MODIS)
- National Polar-orbiting Operational Environmental Satellite System (NPOESS)
- NPOESS Preparatory Program (NPP)
- EUMETSAT Meteorological Operational Satellite (Metop) Program
- NOAA NEXt generation weather RADar (NEXRAD) Program
- National Center for Environmental Prediction (NCEP) Numerical Weather Prediction (NWP) Model Datasets

CLASS must be able to ingest, archive, and provide access to long-term satellite climate data records produced from these large-array data sources, both existing and those to be defined in the future. CLASS will build upon systems already in place to implement architecture for an integrated, national environmental data access and archive system to support a comprehensive data management strategy. The goals of CLASS are as follows:

1. Give any potential customer access to all large-array NOAA (and some selected non-NOAA) data through a single portal.
2. Eliminate the need to continue creating “stovepipe” systems for each new type of data, while, as much as possible, using already refined portions/modules of existing legacy systems.
3. Describe a cost-effective architecture that can primarily handle large array-data sets, but is

also capable of handling smaller data sets as well.
4. Support the reprocessing of any or all datasets managed by CLASS.

The development of CLASS is expected to be a long-term, evolutionary process, as current and new campaigns are incorporated into the CLASS architecture.

The primary function of CLASS is to provide a major portal for access to NOAA environmental data, some of which is stored in CLASS itself, and some available from other archives. The most significant processes required to meet this function that are within the scope of CLASS are:

- Ingest of environmental data from CLASS data providers
- Extraction and recording of metadata describing the data stored in CLASS
- Archiving data
- Browse and search capability to assist users in finding data
- Distribution of CLASS data in response to user requests
- Identification and location of environmental data that is not stored within CLASS, and links to the responsible system
- Charging for data, as appropriate. In general, access to these data sets via FTP will be free. Orders for magnetic media or specialized products will be charged a nominal fee for cost of fulfilling user request.
- Operational support processes: disaster recovery, help desk/CLASS support
- Maintaining a user statistics data base and providing standard and ad hoc statistical reports of CLASS users

A schematic representation of the different processes is shown in Figure 2.

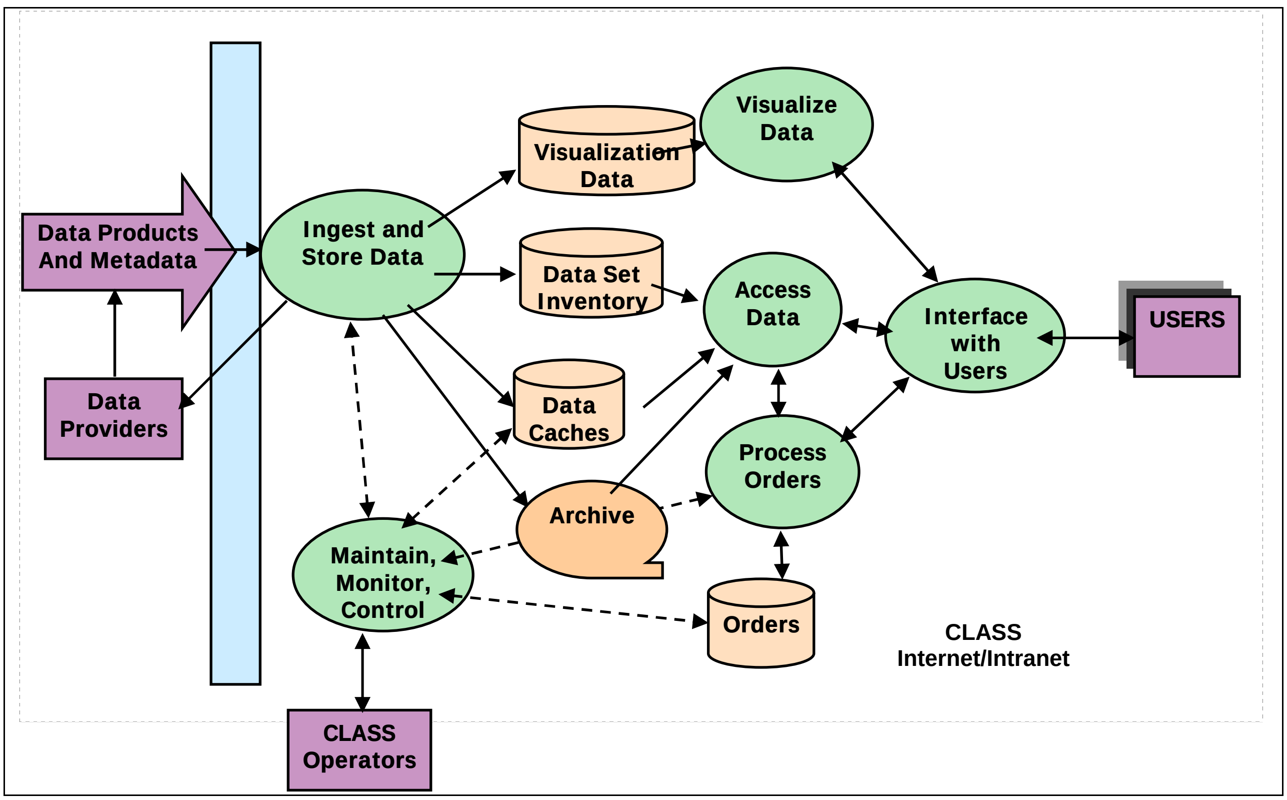

Figure 2. Schematic of CLASS significant processes.

3. OVERVIEW OF SCIENTIFIC DATA STEWARDSHIP

The concept of scientific data stewardship within NOAA means providing the data and information services necessary to answer global change science questions of the highest priority, both now and in the future. High quality, long-term records are needed to address important monitoring and prediction issues. The NOAA in-situ observations, such as surface temperature and precipitation, have long been subjected to extensive scrutiny, quality control, adjustment, and analysis. These steps established confidence in the quality of these data used by decision makers. Similar stewardship functions are required for all the NOAA observations. With successful stewardship, long-term records from all of the NOAA observing systems will reveal their respective maximum potential usefulness regarding a range of critical environmental monitoring and prediction issues. Issues include: atmospheric and oceanic climate change, terrestrial change detection in response to climate changes, space and solar variability, and ecosystem and coastal management.

There are five fundamental principles associated with the NOAA Scientific Data Stewardship Program:

3.1 *Ensure Observing System Quality*

Provide real-time monitoring of climate-scale biases for the global suite of satellite, airborne, and in-situ (atmospheric and ocean) observing systems by monitoring the observing system's performance. Subtle spatial and temporal biases can contribute to serious data quality problems. Automated Network Monitoring that discovers these biases as soon as possible is the first line of defense, before the data becomes part of the long-term historical record. Programs to seek out existing biases in the historical records provide the second level of monitoring necessary for improving the climate data records. Time dependent and other biases can be minimized or eliminated through efficient and effective early processing and periodic reprocessing leading to improving and maintaining the quality of the observing systems.

3.2 *Develop a Climate Processing System*

Provide the necessary algorithms to ensure that understanding of key climate processes can be derived from space-based systems and the fusion of space-based and in-situ systems. The best possible scientific understanding of critical climate and global change issues can only be reached when many opinions and perspectives are explored. Essential to this principle is an active program that engages the research community, establishes partnerships with industry, and facilitates interactions with local, regional, and national governments, agencies, and institutions.

3.3 *Provide Basic IT Support*

Technology now provides the capabilities to implement, maintain, and access the most comprehensive and highest quality historical data and information records possible, critical to the support of effective analysis and prediction activities. A dynamic and flexible strategic plan for the efficient use of IT resources will support rapid adaptation to evolutionary and revolutionary IT changes (e.g., new sensors, telecommunications, storage, commercial-off-the-shelf software, and interoperable hardware). The creation of long-term, contiguous, and quality data records requires the commitment of resources to accomplish these tasks.

3.4 *Document Earth System Variability*

Documenting the Earth's integrated climate system, variability, and change on global, regional, and local scales depends on creating, using, and maintaining the highest quality climate databases and deriving the best historical perspective from these climate data records (CDRs). This will optimize data and information services in order to be responsive. Success will be achieved by establishing end-to-end accountability for establishing long-term, scientifically valid, and consistent records for global change studies. This will ensure that our data and information are available to the maximum number of users.

3.5 *Enable and Facilitate Future Research*

Climate and global change is a long-term societal imperative. Questions and answers often take many years to come into focus due to the nature of the "feedback" and "forcing" systems influencing the global and regional climate systems. This aspect of stewardship encompasses providing the quality climate observations over years and decades, recurring basic information technology hardware and software improvements, and improved models, which can guarantee the preservation of and access to the data, information, and gained knowledge for future generations of scientists and policy makers. Today the Internet, Next Generation Internet, and the emerging grid technology will provide the access for today and the near future. Newly developed data sets will be used to update scenarios and assessments, and to identify and respond to emerging questions that the scientific community is expected to answer.

The implementation of scientific data stewardship has begun and covers not only the archiving plans for all the various satellite and in-situ data sources through CLASS, but it also involves applications with a number of groups and activities as follows: (1) long-term data character groups; (2) current mission instrument groups; and (3) interdisciplinary science groups. The long-term data character group has the mandate for long-term calibration, inter-calibration, and validation of all sensors; collaborates with existing national and international observing system groups; and assures that customers get the highest quality basic data while also responding to data quality questions. The current mission instrument groups are specific to each observing platform (e.g., NPOESS) ramp up during the implementation of the observing platform, and then transitions to the data character group when new mission operations are achieved. These first two groups possess the competencies in the specifics of each mission and produce the documented metadata for fundamental CDRs (FCDRs).

The interdisciplinary groups address major theme areas (e.g., water and energy cycles) use all the instruments, and blend (fuse) data from multiple sources to address climate and global change science questions in order to help provide data and information assessments and options (thematic CDRs - TCDRs).

An external grants and contracts program will use expertise from existing NOAA grants and contracts to assure the involvement of academia and industry in all three of these groups. Grant and contract program members will work with other scientific data stewardship groups to take advantage of directed research with cooperative institutes. Careful thought and support is required to achieve the desired end-to-end observation data management process (collecting, monitoring, processing, product development, access/distribution, archiving, and assessing). An integrated suite of functions to preserve and exploit the full scientific value of the environmental data and information under the stewardship of the NOAA will evolve – provided the basic resources are committed to the goals and objectives. Maximizing the public benefit of the Nation's operational satellites, radars, and past proxy data sets requires implementing the scientific data stewardship program. The involvement and consensus of the research community are essential to accomplishing these objectives. Implementation of scientific stewardship of environmental data from the NOAA observing systems will ensure the quality, usefulness, and accessibility of this national information resource treasure for current and future generations. Figure 3 shows the proposed data flow diagram and how the interactions among NOAA, other agencies, and the climate community are proposed under the SDS program.

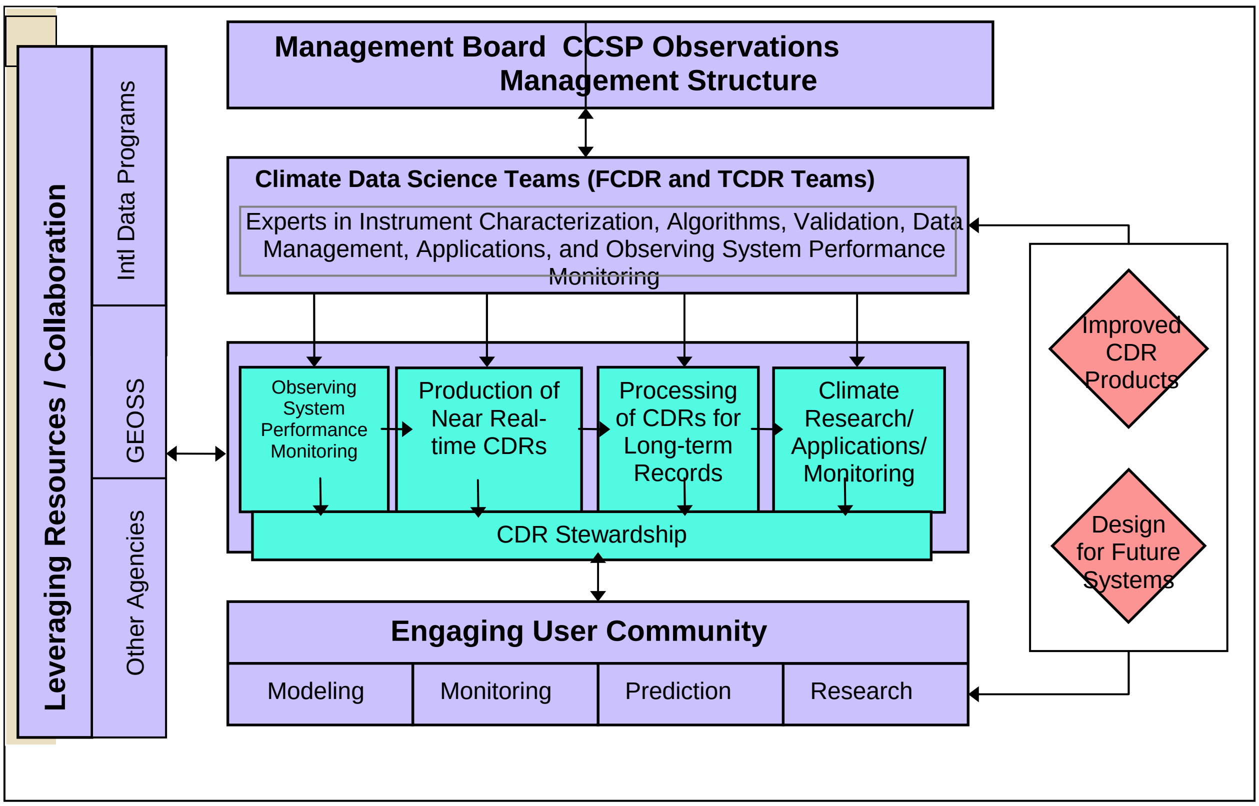

Figure 3. Schematic diagram of data flow and partnerships in Scientific Data Stewardship.

4. CONCLUSION

Scientific data stewardship, with an emphasis upon satellite data and information, is a new era in data management consisting of an integrated suite of functions to preserve and exploit the full scientific value of environmental data and information entrusted to NOAA. These functions provide effective observing system design, careful monitoring of observing systems performance for use in long-term applications, improved quality control, generation of authoritative long-term quality data and records, assessments on the state of the environment, and archive and access to data, metadata, products, and services. Successful implementation of data stewardship will ensure maximum use and public benefit of environmental data and information, now and in the future.