Archive Management: The Missing Component

Howard J. Diamond howard.diamond@noaa.gov
John J. Bates john.j.bates@noaa.gov
David M. Clark david.m.clark@noaa.gov
Robert L. Mairs robert.l.mairs@noaa.gov

National Oceanic and Atmospheric Administration (NOAA)/NOAA Satellite and Information Services

Abstract

The National Oceanic and Atmospheric Administration's (NOAA) National Environmental Satellite, Data, and Information Service (NESDIS) views the area of Archive Management, as comprising three components: Information Technology Infrastructure, Customer Service, and Scientific Stewardship. This last component, Scientific Stewardship, can be characterized as the long-term preservation of the scientific integrity, monitoring and improving the quality, and the extraction of further knowledge from the data. As our data volumes increase, this component, while being recognized as being important, has suffered. Without this component, Archive Management is static, sterile, and lacks the true ability to provide meaningful information and knowledge derived from the archived data. Proper Scientific Stewardship is performed by scientists and data managers knowledgeable in the scientific assessment of a particular data type, and the practice will ensure more effective data management.

1. INTRODUCTION

The National Oceanic and Atmospheric Administration's (NOAA) National Environmental Satellite, Data, and Information Service (NESDIS) views the area of Archive Management, as comprising three components: Information Technology Infrastructure, Customer Service, and Scientific Stewardship. This last component, Scientific Stewardship, can be characterized as the long-term preservation of the scientific integrity, monitoring and improving the quality, and the extraction of further knowledge from the data. As our data volumes and complexities have increased, this component, while being recognized as being important, has suffered.

2. ARCHIVE MANAGEMENT

Archive management as viewed from a data center has three basic components: (1) the information infrastructure necessary to the store, process, and access the data; (2) the customer service features required to get users efficient and effective access to the data; and (3) finally, what we view as the missing yet most critical area of archive management, Scientific Stewardship. Essentially, Scientific Stewardship is a data management discipline that seeks to ensure the quality, calibration, and use in and for scientific applications beyond the initial use of the data of scientific information. It seeks to provide meaningful and derived information, knowledge, and ultimately wisdom from the archived data that can be applied to solving real-world scientific problems are a result of increasing amounts and complexities of environmental data (e.g., satellite and *in-situ* observations), improvements in observational instrumentation, the development of new environmental products and datasets, and the growth of a more sophisticated and knowledgeable user community. These data are used to track hurricanes, tornadoes, and other severe storms; predict future weather and El Niño-Southern Oscillation (ENSO) climate events; monitor numerous ocean phenomena ranging from global coral reef bleaching to the level of the global seas to the quality of the coastal water environment; measure world-wide climate change and the state of the Earth's ozone layer; and detect and monitor a wide range of environmental hazards, including fires, ash from volcanic eruptions, droughts, sea ice, and conditions conducive to flash floods and aircraft icing. Although much progress has been made in recent years, the fact remains that environmental data are underutilized. This situation is being exacerbated as a new generation of more advanced instruments with finer spatial and spectral resolutions producing orders of magnitude more data come on-line.

3. CHALLENGES

NOAA spends over a billion dollars each year on the observing systems that collect environmental data from all over the world. These data are used for a wide range of environment prediction programs – from severe weather forecasting in which data must be used within minutes of collection, to climate prediction programs that use data from the past in order to project the climate over the next 100 years. NOAA continues to move forward

with the development of new observing systems and initiatives in response to these needs. The enhanced systems and initiatives are producing new data, and are driving the need to develop and maintain new and more accessible data sets to be used in answering important questions and solving real-world problems.

In the past decade, it has become apparent that the Earth's environment is a fully coupled, complementary system. Therefore, to adequately understand what is occurring in the U.S. and our adjacent waters, we must be aware of, and understand, phenomena that are occur globally. Regional or national environmental data are no longer sufficient to fulfill NOAA's mission. In order to meet the needs of modern environmental forecasters, NOAA must acquire near-real-time global observations for local, regional, national, and international analyses and predictions. As more and more nations understand that environmental phenomena are not purely local in cause and effect, there are greater demands for cooperation in the collection and sharing of environmental data, on-line data exchange via the Internet, and the creation of global databases that can be searched via the Internet.

Environmental observation platforms including, but not limited to satellites, have become key to the success of many components of NOAA's missions. However, no one nation – let alone one operation – can provide complete global observational coverage. Therefore, it has become critical for NOAA to negotiate and acquire space based data from a growing number of national and international missions. NOAA must partner with other U.S. Federal government agencies and international consumers/operators of environmental data and information. However, the increasing numbers of these missions are placing ever-increasing demands on NOAA's data processing, distribution, archiving and access systems.

No previous decade has seen the magnitude of changes in the volume of data coming into NOAA for processing and archiving as those experienced in the 1990s. However, that explosive growth is nothing compared to what is going to happen between now and the year 2015. We expect a growth in our data holdings to increase exponentially to over 20 petabytes of archived data. Even as current observing systems continue to provide data, new satellite systems such as NASA's series of Earth Observing Satellites (EOS), the National Polar-orbiting Operational Environmental Satellite System (NPOESS), and the NPOESS Preparatory Project (NPP) will be, or are, going into operations. These systems will provide massive amounts of new data, which will present formidable challenges for NOAA.

New users groups are evolving. One such group, weather derivatives, is potentially a $75 billion international industry. This vital new industry includes financial management companies, insurance and reinsurance companies, energy companies, and other industries whose costs are affected by weather and climate extremes in the environment. This new risk management tool uses financial instruments rather than traditional insurance policies to manage the risk of losses due to extremes in weather and climate.

In general, user requests increased throughout the 1990s. Although off-line data user requests doubled, the truly exponential growth has been in the number of on-line users. This currently averages nearly 900,000 per month and is increasing. While on-line requests have increased, it is important to realize that only a small portion of NOAA's data archive is available to the user on-line. As on-line access to NOAA's data expands, the user's average level of technical sophistication and scientific expertise is changing. On-line users are searching for information and answers to specific questions rather than for access to data.

4. ADDRESSING THE CHALLENGES

There is reason to expect that the information technology advances we have seen in the last ten years will continue for the foreseeable future. With these advances, NOAA has made significant progress in its ability to archive and provide access, and will continue to leverage on these advancing technologies. Management of these data can be accomplished only through a rapid expansion in storage capacity, increased communications bandwidth, and automation of the means of data ingest, quality control, and access. The Comprehensive Large Array-data Stewardship System (CLASS) program will act as the connection in NOAA's effort to meet these challenges and pave the way to accommodate the additional massive data volumes expected over the next several years.

The ability to ensure on-going scientific stewardship for NOAA's environmental data and information will only be possible through extensive enhancement of NOAA's current data ingest, quality assurance, storage, retrieval, access, and migration capabilities. This goal will be met through the development and implementation of a standardized archive management system. Such an archive management system will have to be integrated with a robust, large-volume, rapid-access storage, and retrieval system that is capable of a number of functions. These functions include the ability to store large volumes of incoming large-array environmental data and operational products, automatically process on-line data requests from users, and provide the requested data on the most appropriate media.

The target architecture goal will be one which will, through life cycle replacements and upgrades, bring the current NOAA National Data Centers under a single

archive and access architecture that will be under formal configuration management control. This will allow elimination of duplication of effort; minimize stand-alone systems, establishment of an infrastructure to accommodate the large-array data sets, and a reduction in the overall operational and system maintenance costs. The foundation system that is being used is NOAA's highly successful and stable Satellite Active Archive (SAA) http://www.saa.noaa.gov/. Recognized as a stable, modular, well-built system, the SAA approach provides the maximum flexibility while minimizing development work and costs. The heart of the development centers on the upgrade of communications capabilities, an increase in computer storage and power, the use of commercially available modular hardware and software, and the expansion of the World Wide Web access to the data and information through new or enhanced database management, search, order, browse, and sub-setting techniques.

No previous decade has seen the magnitude of changes in the volume of data coming into NOAA for processing and archive as those experienced in the 1990s. Already, there are significant new volumes of data from NEXRAD and DMSP being preserved as part of the NOAA archives. However, that explosive growth is nothing compared to what is planned between now and 2015 (Figure 1). Even as current observing systems continue to provide data, new satellite systems such as MetOp, EOS, and NPP will be going into operations within the next few years. In addition, NPOESS and the next generation GOES (GOES-R) will follow towards the end of this decade. These systems will provide orders of magnitude more data which will present formidable challenges for NOAA. At the same time, new *in situ* observations from widely dispersed automatic reporting platforms are generating significant increases in conventional observation data which NOAA will be required to manage for the long term.

In general, user requests for NOAA environmental data and information increased throughout the 1990s. However, with the growth of the World Wide Web as a ubiquitous technology, a global market was created nearly overnight for NOAA's data and information services. And, although off-line data requests doubled, the truly exponential growth has been in the number of on-line users who extend far beyond NOAA's traditional user community. While these on-line requests have dramatically increased, it is important to realize that only a portion of NOAA's data archive currently is available to the user on-line. As on-line access to NOAA's data expands, the user's average level of technical sophistication and scientific expertise is changing.

On-line users now want to search for information and answers to specific questions rather than simply for access to data. Users, no longer content to wait days for their data or information, are demanding on-line ordering, search, and browse capabilities with electronic file transfer for data delivery. New user groups require near-real-time access to data to support decision-making and rapid response needs. Increasingly, users want information rather than data, as information and products derived from observations are frequently more useful to business and industry than the original data. Scientists and advisors have a critical need for long time-series of historical and recent environmental data to assess long-term trends, evaluate current status, and predict future conditions and events. Therefore, the timeliness and completeness of NOAA's environmental records are crucial.

5. SCIENTIFIC STEWARDSHIP

Much progress has been made in the utilization of environmental data since the first meteorological satellite was launched in 1960. During the 1960s and 1970s, the advances in satellite instruments raced ahead of computing capacity and analysis techniques required to use these data effectively in weather forecasts. Satellite imagery analysis by local weather forecasters was an immediate success, but the use of quantitative satellite data and products in computerized weather forecasting lagged behind. By the late 1990s, however, very fast computers and sophisticated methods of merging vast amounts of satellite and *in-situ* data with numerical forecast models were becoming available.

The evolution of satellite capabilities is imposing a requirement for significant scientific effort to accommodate the new data. Between now and the year 2010, the volume of potentially useful and routinely available environmental observation data will grow by a factor of approximately 100,000. This includes data from new operational NOAA and NASA, as well as missions from other countries and agencies. To accelerate the use of this future environmental data in operational weather forecasts, NASA and NOAA have formed a collaborative Joint Center for Satellite Data Assimilation (JCSDA) in order to develop an end-to-end process for the operational utilization of satellite observations. The JCSDA will be a center distributed among several centers of expertise. These NOAA centers will include NESDIS, the National Centers for Environmental Prediction (NCEP), NOAA's Oceanic and Atmospheric Research (OAR) office, and the NASA Data Assimilation Office (DAO). Each will bring its own area of expertise to the joint effort and by collaborative efforts will make efficient and rapid advances in the use of satellite data in weather forecast models.

The JCSDA will promote the development of common weather forecast models for research and operations. At the present time each U.S. forecasting center runs its own

models, and data assimilation advances made at one center are not easily transferred to other centers. Common models will make this process efficient, and components required by data assimilation will be developed for wider community use. This will include community radiative transfer models, surface emissivity models, and surface physics models.

5.1 Principles of Scientific Stewardship

The concept of scientific stewardship within NOAA means providing the data and information services necessary to answer the global change scientific questions of highest priority, both now and in the future. The NOAA scientific stewardship program has five principles as follows:

Ensure Observing System Quality. To provide for the real-time monitoring of climate-scale biases in the global suite of satellite and in-situ observing systems by monitoring observing system performance. Since subtle spatial and temporal biases can create serious problems in future use of the data, we must develop the tracking tools necessary for detection of biases in the climate record. These biases can then be minimized or eliminated through efficient communication and coordination of information related to network performance using both in-situ and satellite observations.

Provide Basic Information Technology (IT) Support. To document Earth system variability and change on global, regional, and local scales. This will be accomplished by building and maintaining a high quality base of data and information and establishing the best possible historical perspective critical to effective analysis and prediction.

Develop a Climate Processing System. To provide the necessary algorithms to ensure that understanding of key climate processes can be derived from space-based systems and the combination of space-based and *in-situ* systems. The best possible scientific understanding of critical climate and global change issues can only be reached when all opinions and ideas can be explored. Thus, an active program of engaging the research community, establishing partnerships with industry, and increasing interactions with local and regional governments to develop a processing system for satellite and *in-situ* observations are envisioned.

Document Earth System Variability. In order to better document the overall Earth system we need to build and maintain the highest quality climate database and establish the best historical perspective. This will optimize data and information services in order to make research easier and more effective by ensuring that those services are simple, straightforward, direct, and responsive. This 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 amount of users

Enable and Facilitate Future Research. Because action is required now, and climate and global change societal imperative questions may not come into focus for many years, we must invoke the concept of stewardship to justify this effort. This aspect of stewardship involves providing the basic information technology, hardware, telecommunications, and software support to guarantee that the data can be safeguarded and communicated both within NOAA and to outside users for generations in the future. As new global change imperative questions arise, and in order to safeguard the interests of future generations, we must make data sets easily available on the Internet and emerging Grid technology outlets. These data sets will be used to update scenarios and assessments, and to identify and respond to emerging questions that the scientific community will be looked to for providing answers.

5.2 Building the Climate Record

Over 20 years of satellite and *in-situ* data are now available for climate analysis and detection of climate trends. However, a completely new polar-orbiting environmental satellite system, NPOESS, will begin operation in 2008. The NPOESS system will have new weather and climate monitoring instruments as well as new instruments for monitoring ozone. A prototype of the NPOESS spacecraft, the NPP, will be launched in 2006. A substantial research effort will be required in order to ensure continuity in the earth's climate record between the current operational system that has been in operation over the past 20 years and NPOESS. The proper time and data sets for this research effort will be provided by NPP, which will provide the bridge to NPOESS. The NPP will allow coincident climate observations between the old satellite instruments and the new ones that will begin operation with NPOESS and continue for many years. Construction of a seamless climate record between the current satellites and NPOESS, as well as other instrumented climate data, is a very important yet difficult challenge.

5.3 Examples of Scientific Stewardship

In parallel with an increasing number of satellite and *in-situ* products and applications, there are a growing

number of user communities that NESDIS serves and interacts with. These include weather forecast offices; NOAA's environmental prediction centers specializing in hurricane, severe storm, aviation, space environment, hydrological, ocean, and climate forecasts; the national and international climate community concerned with climate change; federal, state, and local resource managers responsible for coastal environmental monitoring; officials responsible for reacting to environmental hazards (e.g., fires); the aviation and shipping communities; those involved in drought mitigation activities; and local and international groups assessing the health of coral reefs.

5.3.1 Data Quality. NOAA's current and near-future environmental instruments have been designed to measure properties of the Earth and its atmosphere for application primarily to weather forecasting. The basic measurement of the satellite instruments is of the intensity of the radiation upwelling from the Earth-atmosphere system, from which geophysical properties are derived mathematically. It goes without saying that the basic radiation measurements need to be highly accurate. This means that the instruments need to be calibrated accurately on orbit, and NOAA has a vigorous program to achieve this.

There is a growing demand to detect climate change from satellite observations. The instruments and NOAA's observing strategies were not designed to provide the long-term continuity, stability, and consistency in the observations that this application requires. Furthermore, NOAA occasionally makes changes in sensor characteristics or measurement techniques, and these present serious problems, as do time gaps in the observing period. For constructing climate-quality data sets, it is imperative that we include supporting ground-based *in-situ* observations, which validate and complement the satellite observations.

In the first category, the recommendations included the following elements: (a) maintaining constant local observing times for polar satellites; (b) obtaining continuity or, even better, one-year overlap between observations of successive satellites; (c) improving instrument calibration systems to account for on-orbit drifts in radiometer gain; and (d) making information available on instrument performance status and changes that might affect the observations. Although current satellite systems do not comply with many of these recommendations, NOAA intends to work towards compliance in its future systems, e.g., NPOESS.

In the second category, the recommendations included: (a) maintaining the current observational and primarily radiosonde-based upper air observing system; (b) improving the accuracy and reliability of observations of the stratosphere; (c) assuring temporal continuity and consistency of the data record; (d) generating and making available a "climate data record" of the observations and associated metadata; (e) upgrading the radiosonde temperature and humidity sensors while maintaining continuity of calibration; and (f) exploring options for a significantly improved next-generation atmospheric sounding system. NOAA plans to comply with all of these.

The recommendations from the third category dealing with the climate data record included: (a) reinvigorating the full range of activities within NOAA to ensure a long-term climate record; (b) monitoring and making available the performance and error characteristics of the space-based and *in-situ* networks; and (c) establishing a dialogue and information exchange on the climate data records and the sensors that are their source. Here too, NOAA expects to comply.

5.3.2 Long-term Stewardship. The use of a variety of environmental data for climate studies has progressed from being experimental to routine. These data sets have proven to be of high value for climate studies. They have been used in regional and global temperature and upper tropospheric humidity trend studies, in studies of the ozone hole, and in studies of clouds and rainfall. The Intergovernmental Panel on Climate Change (IPCC) and various World Climate Research Programs (WCRP) have in turn, used these products in assessments. Further applications of satellite data to climate studies, particularly for retrieval of column CO_2, are currently under development and appear promising.

As we eagerly move into the next generation of instrumentation, it is now clear that the constellation of operational satellites will be the backbone of the long-term global climate observing system. There are major challenges we face, however, in moving to the next generation systems and in preserving and using the past data. How do we maintain a seamless time series of fundamental observations during this transition? How do we deal with the quantum jump in data volume as well as for decades of data? How do we ensure that user access to these data archives is simple, straightforward, direct, and responsive?

Answering these challenges will take a concerted effort by instrument scientists, climate scientists and computer scientists. There must be extensive collaboration between the research and operational climate communities. Computer scientists and climatologists will need to provide a sound and effective means to ensure that all necessary data is preserved and remains accessible in easy-to-use formats. It will also take a long-term commitment to provide resources to enable preservation of the climate archive from the first generation satellite systems of the 1980s and 1990s, through the transition satellite systems of the EOS era, to the second generation

of operational systems in the future NPOESS.

5.3.3 Target Architecture. The target architecture goal will be one which will, through life cycle replacements and upgrades, bring the current NOAA National Data Centers under a single archive and access architecture that will be under formal configuration management control. This will eliminate the duplication of effort, minimize stand-alone systems, build the infrastructure to accommodate the large-array data sets, and reduce the overall operational and system maintenance costs. The foundation system that is being used is the highly successful and stable Satellite Active Archive (SAA). Recognized as a stable, modular, well-built system, the SAA approach provides the maximum flexibility while minimizing development work and costs. The heart of the development centers on the upgrading of communications capabilities, increasing computer storage and power, use of commercially available modular hardware and software, and expansion of the World Wide Web access to the data and information through new or enhanced database management, search, order, browse, and sub-setting techniques.

5.34 Future Developments. As new technologies are developed they will also have to be factored in. Recent developments in Grid technologies, viewed as the Internet infrastructure for the 21st century, is designed to facilitate collaboration in a more seamless distributed processing environment with the intent of having a more uniform data access. Grids, as they are known, are persistent environments that enable software applications to integrate instruments, displays, computational and information resources that are managed by diverse organizations in widespread locations. While this in many cases gets associated with high-end super computing, it is probably the next step in the evolution of the Internet for the dissemination of information in a more seamless manner. A popular example of Grid technology is the SETI@home screensaver that allows the unused computer resource time of PCs to participate in the ongoing Search for Extraterrestrial Intelligence by analyzing data specially captured by the world's various radio telescopes. NOAA is working with the Committee on Earth Observing Satellites to begin investigating the use of Grid technology in order to facilitate a more seamless exchange of data among satellite data agencies.

5.3.5 Implementing Scientific Data Stewardship. The actual implementation of scientific data stewardship covers not only the archiving plans for all the various satellite and *in-situ* data sources, but it also involves applications with a number of groups and activities as follows: (1) data character; (2) mission groups; (3) interdisciplinary groups; and (4) external grants. The 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 mission groups are specific to each observing platform (e.g., NPOESS), ramp up during the implementation of the platform and then transition to the data character group during stable operations; these groups have the competency in the specifics of each mission along with complete documented metadata. The interdisciplinary groups address major theme areas (e.g., water and energy cycles) and use all instruments and blend with all data sources to solve climate and global change science questions in order to help provide data and information assessments and options. Finally, the external grants program uses expertise from existing NOAA grants and contracts to assure the involvement of academia and industry; and works with other Scientific Data Stewardship groups to take advantage of directed research with cooperative institutes.

6. Comprehensive Large Array-data Stewardship System (CLASS)

Now through the judicious application of recent technological advances, it is possible to provide incremental improvements in service capabilities at reasonable costs. In an effort to take full advantage of these opportunities, NESDIS has initiated an environmental data and information archiving and access activity which will focus on improving the ability of NESDIS to be the steward for NOAA's environmental data and information and to maintain a permanent archive that is easily accessible to the world science community and to other users. The keystone of this activity is the Comprehensive Large Array-data Stewardship System (CLASS) Project.

The CLASS Project is designed to enhance the NOAA capability to provide environmental data and information archive and access services to the Nation through the effective application of modern, proven techniques and technology. The project places special emphasis on the ability to efficiently archive the vast quantities of NOAA satellite and *in situ* observational data currently being collected and to be collected; to safely and permanently preserve those valuable data for future generations to use; and to provide rapid data access in a cost-effective and secure manner.

There is reason to expect that the information technology advances we have seen in the last ten years will continue for the foreseeable future. With these advances, NOAA has made significant progress in its

ability to archive and provide access to its increasing data volumes, and will continue to leverage on these advancing technologies. Management of these data can be accomplished only through a rapid expansion in storage capacity, increased communications bandwidth, and automation of the means of data ingest, quality control, and access. The CLASS project will act as the connection in NOAA's effort to meet these challenges and pave the way to accommodate the additional massive data volumes expected over the next several years.

There are a number of aspects to the successful implementation of CLASS. First, there are those aspects which are purely mechanical or technical in nature. They include, for example, communicating the data from the source to the primary and backup storage locations; quality control and pre-processing of the data; storage of the data on media such as tapes and disk; and, post-processing the data to extract information. In addition, there are those issues concerning the virtual on-line search, discovery, retrieval, display, and customer order processing capabilities for the user community. All these various tasks must be accomplished securely, quickly, and efficiently to meet the needs of NOAA's user community.

The ability to ensure on-going scientific stewardship for NOAA's environmental data and information will only be possible through extensive enhancement of NOAA's current data ingest, quality assurance, storage, retrieval, access, and migration capabilities. This goal will be met through the development and implementation of a uniform archive management system, which will be integrated with a large-volume, rapid-access storage and retrieval system capable of storing the incoming large array environmental data, *in situ* data, and operational products as well as receiving a user's on-line data request, automatically processing the request, and providing the requested data on an appropriate media. This system will provide standardization in media, interfaces, formats, and processes. Additionally, the system will facilitate ongoing migration, preservation, and validation to new technology and media. This system is modular in design, built to integrate with automated real-time or near-real-time systems that deliver data. Transaction processing will be implemented to enable essentially autonomous operations and, where appropriate, the system will allow users to pay for data or services through credit card or automated billing.

The target architecture goal will, through life cycle replacements and upgrades, bring the current NOAA National Data Centers under a single archive and access architecture that will be under formal configuration management control. This will allow elimination of duplication of effort, minimize stand-alone systems, build the infrastructure to accommodate the large array data sets, and reduce the overall operational and system maintenance costs. The foundation system that is being used for CLASS is the highly successful and stable Satellite Active Archive (SAA) which provides the maximum flexibility while minimizing development work and costs. The immediate goals for CLASS are to support upcoming campaigns and the new data they will create. The CLASS architecture includes plans for a central portal to all NOAA environmental data. In addition to the data stored within the CLASS Archive and Distribution System, CLASS will provide a data discovery function that will assist users in locating environmental data stored in other NOAA and some non-NOAA archives. In addition to the file-level search currently provided by SAA, CLASS will provide directory level searches, and connect users to other systems. This CLASS functionality is also based on an existing NOAA system, the NOAAServer.

7. GEOSPATIAL ENHANCEMENTS

CLASS is being designed to allow improved access to the wealth of satellite data via the latest data discovery and geospatial techniques and tools. This effort is being led by the CLASS team primarily located at NGDC with assistance from NODC and NESDIS headquarters.

A critical aspect to making these data readily available to our customers is to ensure that the satellite metadata are current, accessible, and that they follow the appropriate metadata standards [1]. Currently, the NESDIS Satellite Metadata is physically located at several different locations and only minimally takes advantage of national (Federal Geographic Data Committee, FGDC) and International Standards Organization (ISO) draft ISO 19115 metadata standards. As a note, the current FGDC standard will be ISO 19115 compliant. To correct this situation, the CLASS project has initiated a number of tasks to improve our current data discovery and geospatial techniques and tools, and to develop new ones as appropriate.

One task is designed to implement a robust NOAA Metadata Repository (NMR) management system within CLASS. Metadata, which describe the content, quality, condition, and other characteristics of data are critical to data discovery, usability, quality, interoperability and automatic processing. The planned NMR architecture will primarily handle large array data sets but also be capable of handling smaller data sets as well. The plan for developing the NMR represents a comprehensive approach to creating high-quality metadata and revitalizing NESDIS systems for managing it. The proposed effort relies on state-of-the-art data management tools and techniques and extensive experience and expertise in the NESDIS Data Centers and satellite data processing groups. The plan emphasizes building on this expertise to create a metadata management and access system that will serve as the preeminent example for other

NOAA Line Offices and agencies at all levels of government.

The NMR will be based on relational databases (Oracle), internet mapping software (ArcIMS), XML and XML stylesheets, and Java access tools (Blue Angel Technologies). The NESDIS metadata team has extensive experience with all of these tools and they are presently running in several NESDIS locations as part of the National Virtual Data System (NVDS) and Coral Reef Information System (CoRIS) Projects. This plan for the development of the NMR builds upon the work done by NVDS and CoRIS over the last several years. That work has included populating repositories at NGDC (NVDS) and NODC (CoRIS) with nearly 10,000 metadata records from the Data Centers, other NOAA Line Offices, and CoRIS investigators, and developing and field testing several metadata manager interfaces to the repository. In addition, this project will incorporate work done developing programmatic interfaces to the repositories as part of the overall CLASS Data Discovery Support System.

The foundation of the proposed NMR system will be made up of Oracle databases with access provided using the Enterprise Blue Angel Metadata Repository Java Classes. This foundation also includes a set of XML files that support Z39.50 searches, metadata transfer to other systems [e.g., Global Change Master Directory (GCMD), the FGDC Clearinghouse, and the Geography Network)], and display using XML stylesheets. It will also rely on a spatial database (either Oracle or Informix) and ESRI's geographic information systems (ArcSDE, ArcMAP, and ArcIMS).

Initially, the focus will be on two metadata collections, the SAA Data Family Metadata and the Satellite Data Products Database. These were chosen because these two data collections play a critical role in the creation of FGDC compliant metadata for the existing NESDIS satellite products. The primary goal of this study is to compare the individual fields in the existing metadata collections and to evaluate those fields that have similar meanings (semantics). The FGDC standards with the remote sensing extensions are being used for this crosswalk process. This combination provides a broad collection of well defined metadata fields and a framework for organizing the crosswalks.

The geospatial foundation to be used in CLASS will be based upon standard relational database management systems with spatial extensions [1]. This is being done since these systems are capable of supporting the wide variety of environmental data types to be available via CLASS. In addition, these systems offer support for a number of different data access approaches being investigated.

As such, a second task initiated will work towards the development of tools for processing level 1 and level 2 data to extract the spatial information presently stored in the file record headers and to ingest that data into geospatial databases as orbits and scan lines. Completion of this task will immediately benefit CLASS customers by making all of the data in the database available for both spatial and general SQL queries.

8. CONCLUSION

As we have shown, the missing element of archiving is Scientific Stewardship. In order to build consistent and high-quality records of environmental observations and produce comprehensive analyses of environmental change, it is imperative for data providers such as NOAA to partner with the scientific community by ensuring the provision of high quality data and services as well as the generation of useful and understandable products that can be easily accessed. This will include the provision of value-added products through the development of new algorithms in order to derive environmental parameters from the instrumental record, the generation and validation of environmental data records from satellite and *in-situ* products that are calibrated by i*n-situ* measurements, and the generation and analysis of products that more accurately identify environmental change. This will require increased access and utility to vast archives of data that will be built over the next 15 years and will require the refinement and development of tools and techniques (e.g., Geographic Information System, Grid Technology, data mining) in order to effectively take advantage of the avalanche of more complex environmental data that is to come.

9. REFERENCES

[1] Clark, D.M. *Science Data Stewardship Through a Global Science Data Network.* 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.

[2] Habermann, T., 2002: Geospatial Data Systems and NESDIS Satellite Data (unpublished)

[3] _____, K. Nuttycombe, J. Barkley, T. Gaines, and T. Stevens, 2002: NESDIS Satellite Metadata Model, V1.0 (unpublished).

[4] NESDIS, 2002: http://www.ngdc.noaa.gov/ngdcinfo/aboutngdc.html

[5] NOAA. *The Nation's Environmental Data: Treasuresat Risk, Report to Congress on the Status and Challenges for NOAA's Environmental Data Systems*, Washington, DC, 2001.

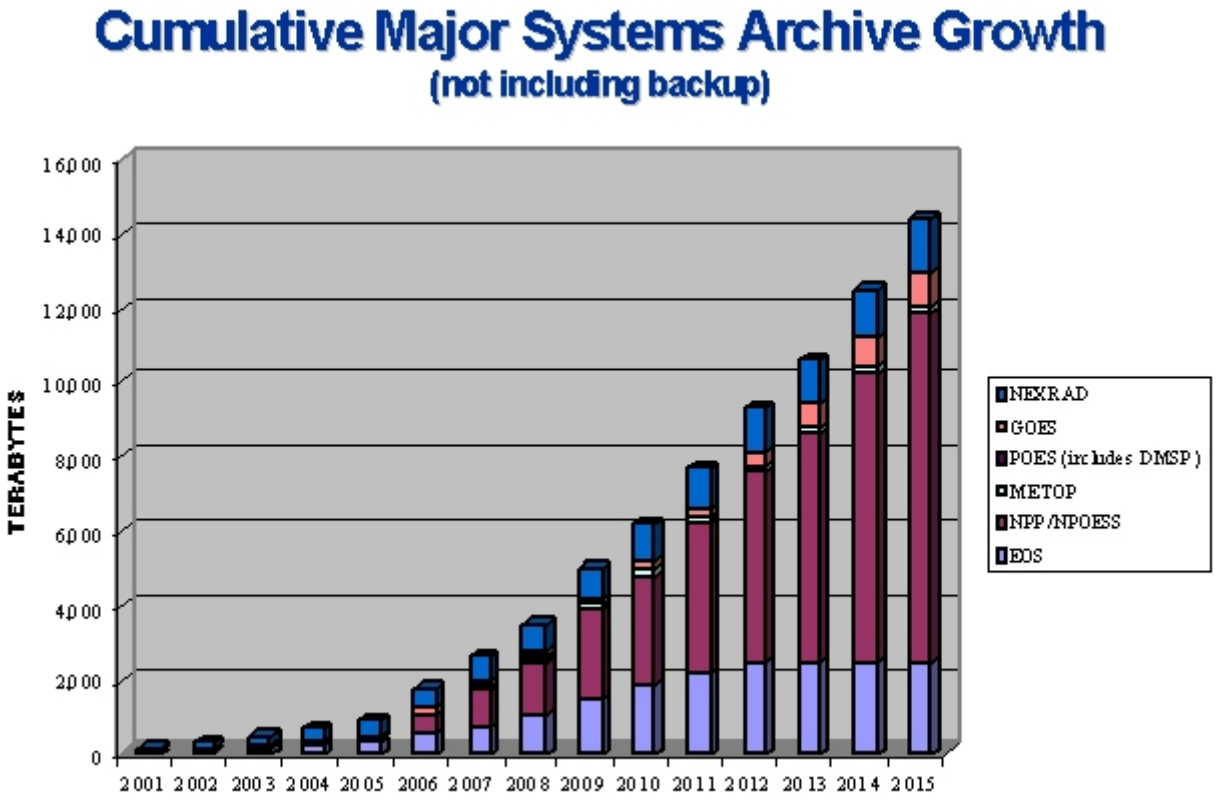
Cumulative Major Systems Archive Growth
(not including backup)
TERABYTES
160 00
140 00
120 00
100 00
80 00
60 00
40 00
20 00
0
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
NEXRAD
GOES
POES (includes DMSP)
METOP
NPP/NPOESS
EOS

Figure 1