JP1.8

CLASS FUTURE PLANS

John J. Bates *
NOAA National Climatic Data Center, Asheville, North Carolina

Richard G. Reynolds
NOAA NESDIS Office of Systems Development, Suitland, Maryland

Constantino Cremidis
Computer Sciences Corporation, Camp Springs, Maryland

Carlos Martinez
TMC Technologies, Fairmont, West Virginia

ABSTRACT

The National Oceanic and Atmospheric Administration (NOAA) has developed the Comprehensive Large Array-data Stewardship System (CLASS) to archive and provide access to the data from current satellite-based observing systems. CLASS is also being designed to handle the significant increases in data volume that will come from planned satellite launches. Finally, CLASS will ultimately be capable of supporting current in situ data sources.

NOAA has defined a path forward for the CLASS system in order to better serve our customers. A cornerstone of this plan is to increase communications with and input from our customers. An important next step in this process is a users' forum at this AMS conference to be held on Wednesday, January 12, 2005, at 12:15 pm. This forum will provide NOAA customers with an overview of current and future plans for CLASS. Customers will also be able to provide feedback on their experiences with CLASS and their requirements for the future evolution of CLASS.

1. INTRODUCTION

There are important improvements in functionality, capacity, interoperability, and of course, more data that are planned for CLASS over the next several years.

The following capabilities and additional functionality are planned for CLASS over the next two years:

- Coordination with users to define additional functionality
- Improved metadata and the ability to search it
- Enhanced geospatial search and discovery
- Advanced data discovery
- Dataset manifest and more generalized ingest processes
- Interface with the National Environmental Satellite, Data and Information Service (NESDIS) e-commerce System (for delivery of CLASS data on physical media)
- Compatibility with the ISO standard for Open Archive Information Systems (OAIS) reference model
- OAIS-based Submission Agreements between CLASS (Archive) and data providers (Producers)
- Hierarchical Data Format-5 (HDF5) format

Additional details on these plans are given below.

It is estimated that more than 70 petabytes of data, not including the mirror sites, will be added to CLASS by the year 2017. These estimates (Figure 1) include the following data:

- Meteorological Ops / EUMETSAT Meteorological Observation Satellite (MetOp)
- National Polar-orbiting Operational Environmental Satellite System (NPOESS)
- NPOESS Preparatory Project (NPP)
- NASA's Earth Observing System (EOS) Moderate Resolution Imaging Spectroradiometer (MODIS)
- Geostationary Operational Environmental Satellites (GOES) statistical information
- NASA's Jason-1 satellite
- GOES-R satellites
- NEXRAD
- Climate Products
- EOS satellites (beyond MODIS)
- In situ data sources: Automated Surface Observing System (ASOS)
- Reclamation of historic data
- Reprocessing of historic data

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

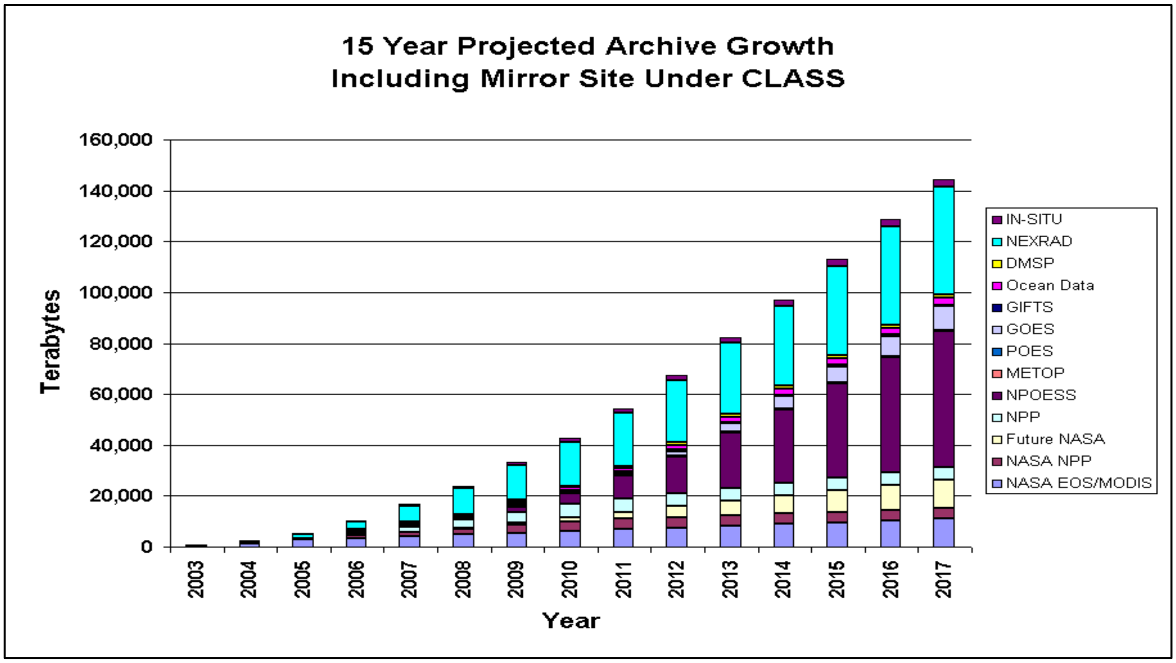

Figure 1. Projected future data volumes in CLASS by observing system.

In order to be able to ingest, archive, access, search, and distribute these massive amounts of data, CLASS has planned a number of capacity improvements, including the following:

- Enhanced Storage Area Network (SAN) disks
- Potential third operational site for expanded ingest and distribution capabilities
- Additional data archive capacity
- Increased telecommunications throughput
- Automated load balancing
- Enhanced interoperability with data centers

In planning future enhancements, CLASS must take customer needs and expectations into account. Customers expect computer-based systems to continuously improve so CLASS must improve as well. CLASS must keep up with and take advantage of improvements in information technology, including faster hardware, expanded Internet bandwidth, improved software, and emerging standards. It must evolve to provide faster access to data and easier ways of finding, browsing, ordering, receiving, and using data and derived products.

2. DESIGNATED USER COMMUNITY

The community of potential users of CLASS runs the gamut, from someone with limited computer knowledge to elementary school teachers (and students) to internationally recognized experts in remote sensing. Attempting to serve such a heterogeneous community with a single system would present an enormous, maybe even impossible, challenge. Therefore, CLASS will first focus on serving a subset of all possible users. It will aim initially to serve the scientific and technical community. CLASS customers will be expected to be scientifically and technically literate and familiar with tools for analyzing scientific and gridded data. However, they should not need to have extensive pre-existing experience with satellite instruments or CLASS datasets. The CLASS user base has grown exponentially over the last decade (Figure 2) and is expected to continue to rapidly grow over the coming decade.

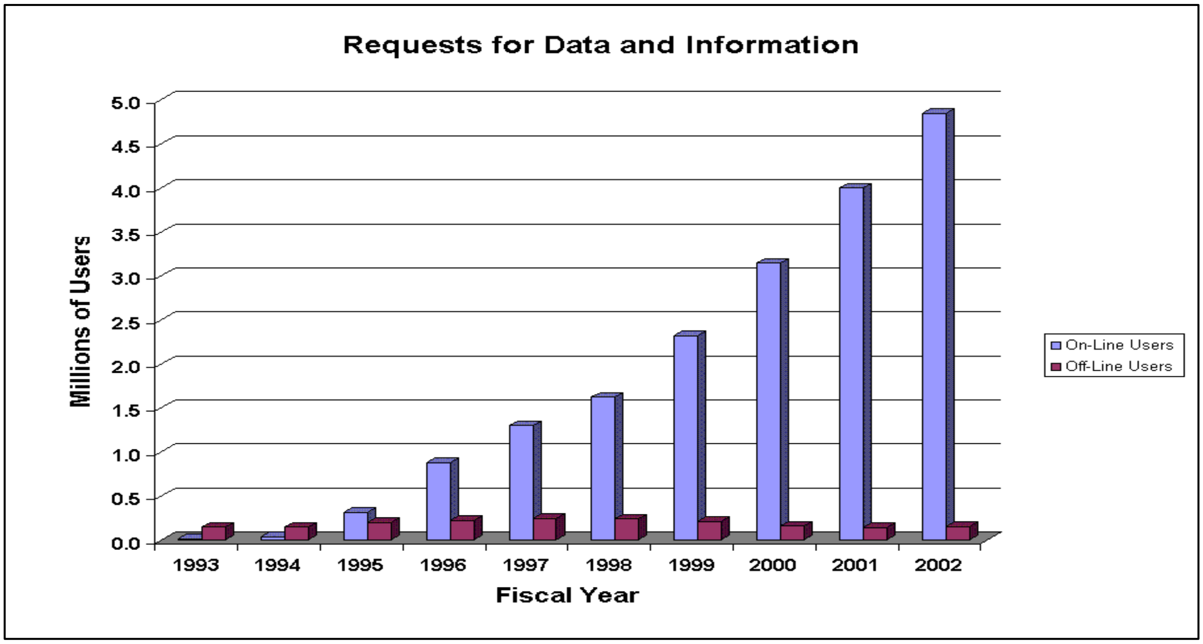

Figure 2. Increase in user access to CLASS and an earlier pilot system.

3. DATA DISCOVERY

3.*1 Metadata Search*

CLASS will include a mechanism to allow its customers to search for and locate data and products that best meet their requirements. CLASS will support searches through two interfaces: interactive and automated. An interactive search would allow users to search for data by specifying criteria through a Web browser and then viewing metadata and/or browse images to determine if the data meet their needs. An automated search would allow agents (software routines) to search for CLASS datasets via standard Application Programming Interfaces (API).

Both search mechanisms depend upon sufficient and accurate OAIS Descriptive Information for all CLASS datasets. To be most effective, this information should be in the form of metadata available in accordance with Federal and industry standards. This metadata must be complete enough for customers to determine if a dataset includes the parameters, time period, and geographic location they are looking for, at the resolution, frequency, quality, and accuracy they require.

CLASS currently provides access to customers through a Web browser interface. However, the interface should be more metadata rather than dataset-centric. That is, it should present a general interface without users needing to specify a particular instrument or sensor. The search interface should allow users to search via specific field values or through free text (albeit with controlled vocabularies). Users should be able to specify field values through pull-down menus or, where appropriate, interactive maps. Free text searches should support Boolean operators.

The search system should be dynamic, that is, it should allow the user to set values of specific fields that only appear for certain types of data. For example, if a user selects “radar data” then he/she would be given the additional option of selecting an individual radar site.

Development and maintenance of multiple interfaces should be considered to best meet the needs of the diverse customer base. While novice users would be best served through extensive menus and forms, expert users might prefer a terse command line interface or computer to computer interfaces.

The search system should be extensible to support searches on any metadata maintained by CLASS. It should offer customers the capability to identify special attributes of the images they need. For example, customers could identify cloud-free data sets by examining browse images, but this can be very time consuming. They should be able to search for only images that are cloud free.

To ensure its datasets can be located by a wide user community, CLASS should provide descriptions of its datasets to popular external environmental catalogs such as NOAA Server and the Global Change Master Directory (GCMD), in a Federal Geographic Data Committee (FGDC) compliant format.

As CLASS is extended to provide access to other NOAA and non-NOAA datasets, it should be able to automatically collect and ingest metadata describing those datasets, assuming the metadata adhere to FGDC or ISO standards.

CLASS should include a mechanism to respond to automated requests for information concerning datasets from a system outside CLASS. CLASS should provide access to these requests through a Web-services interface to its metadata, which would return metadata in a standard machine readable form (e.g., FGDC-compliant XML).

Many Earth observations catalogues that require search-level interoperability have adopted Z39.50, the international standard used for catalogue search. Therefore, CLASS should provide access to its metadata via the Z39.50 protocol. CLASS should provide additional functionality by complying with the OGC Catalog Services Specification.

Industry standards for richer and more capable search, discovery, and retrieval are now emerging. These require that systems describe not only their datasets, but also the services that they can provide through a well-defined Web-services interface. CLASS should provide a Web-services interface as soon as standards in this area are widely adopted within the CLASS user community.

3.2 *On-line Browse*

Browse or quick-look data or images are often needed for users to determine if a particular set of data meets their requirement. For example, cloud-free images are needed for many remote sensing applications. To meet this requirement CLASS must provide a basic browsing and visualization capability that extends across the full breadth of CLASS data. The browsing capability should provide geolocated and time-referenced graphics and images suitable for the evaluation of CLASS data through standard Web browsers. Browse images may be reduced resolution samples of actual images or may be derived from compositing, averaging, or other processes, whichever provides the most meaningful result for intended uses of the data.

Browse images should be available for delivery through both the interactive and automated search interfaces.

4. DATA DELIVERY

Obviously, at the most basic level, CLASS must be able to deliver data and products to its customers. These data and products should be accompanied by complete OAIS Packaging Information that describes the file names, directory structure, etc. For CLASS to fully satisfy its users, it must also provide powerful and flexible tools that allow users to tailor output to their own needs and circumstances.

4.1 *Subscriptions*

At times customers might prefer to receive a given product on a regular schedule, i.e., set up a subscription. In addition to the normal criteria used to select a dataset, the customer should also be able to define the time range when the subscription is valid (e.g., from May 1 to September 30) and the frequency at which data will be delivered (e.g., daily). Once defined, the subscription could either push the data to the user automatically, or notify the user that the data is available. CLASS should support both approaches.

CLASS should manage subscriptions specific to an individual customer rather than as a generic subscription for all customers who are interested in a particular dataset. In that way, customers will only receive products that meet their requirements and will not need to evaluate the product to determine its suitability whenever a notification is received. Furthermore, individually-managed subscriptions would allow customers to receive products tailored to their needs (e.g., subsets or data delivered in a specific format).

4.2 *Delivery via the Internet*

CLASS is able to deliver data and products to customers over the Internet using standard protocols. Traditional standards such as FTP and HTTP must be supported.

CLASS will also investigate modern alternatives that have been developed to serve the peer-to-peer file sharing community. One example is Gnutella, which offers many advantages over FTP such as support for automatic restarts of interrupted downloads and parallel downloading in slices from multiple servers.

Interoperability between Earth observation data systems depends crucially on technical program interfaces, typically described through standard service definitions. Such service definitions precisely specify the syntax and semantics of all data elements exchanged and fully describe how systems interact at the interface. At present, several standard APIs are available and no single standard has emerged as the clear leader. However, CLASS will use one or more of the most common APIs that are currently in use within its community. These include OPeNDAP, OGC Web Map/Feature/Coverage Server, and SOAP/WSDL. CLASS will keep abreast of developments in this area and adapt standards as they become widely accepted.

4.3 *Delivery via Removable Media*

Although the growth of Internet coverage and bandwidth allows CLASS to distribute an increasing volume of data on-line, for the foreseeable future some customers will require off-line delivery of media. As an example, recently a customer discussed setting up a subscription that would require delivery of about 140GB/day. Given the large volume of data they preferred to receive it via removable media rather than on-line.

CLASS will continue to include the capability for customers to order data be delivered via removable media. A variety of up-to-date media should be available such as cartridge tapes, 8mm tapes, CD-ROM, and DVD-ROM. While the capability of charging for delivery of this portable media is a requirement, CLASS will interface with the NESDIS e-commerce System for these financial transactions.

5. DATA UTILITY

5.1 *Standard Data Formats*

To be effectively used by customers, CLASS data and products should be delivered in a format that the customer is familiar with and for which he/she already has software tools. Unfortunately, even within a single community and application, several choices are often available. For example, for multi-dimensional gridded data (e.g., from models) netCDF and GRIB are widely used. HDF and HDF5 are also sometimes used for this purpose but are more commonly used for raster data.

Since no single standard is accepted or can meet the needs of all of CLASS's customers, CLASS will provide the capability for customers to receive data in any one of a few standard formats. Different formats are appropriate for different types of data so the available choices must depend upon the type of data or product needed.

5.2 *Documentation and Metadata*

Data and products received from CLASS must be independently understandable**.** That is, they should have sufficient documentation to allow the information to be understood and used without needing the assistance of the experts who produced them. The amount of metadata required to adequately characterize the data depends upon the user and purpose for which the data will be used. As an illustration, consider the following two examples.

1) Robert would like to include a visible satellite image of a hurricane in his term paper. Assuming that the image is available in a standard format that can be displayed by a Web browser, the only metadata that he needs are the name of the hurricane and the date of the image.

2) Susan is studying the impact of tropical storms on sea surface temperature. She would like computed sea surface temperatures of the Gulf of Mexico immediately before and after the passage of a hurricane. Along with the images, she needs detailed information on the date and time of each, the source satellite, the spectral bands and algorithm used to compute the temperatures, the calibration coefficients that have been applied, and the expected error range.

For CLASS to best satisfy the needs of its users, it must be able to provide a different level of documentation and metadata to different users. Therefore, CLASS must collect and maintain metadata detailed enough to meet the needs of the most demanding users that can be imagined, but should allow users to retrieve only the level of metadata that they require.

6. CONCLUSIONS

CLASS is a combined process to both reengineer legacy data storage and access systems and blend new and efficient technologies to ensure the stewardship of existing (e.g., POES, GOES, NEXRAD, in-situ) and rapidly approaching large-array data sets (e.g., NPP/NPOESS, EOS, METOP, NEXRAD). It is an aggressive plan to safeguard, enhance, expand, and automate NOAA's capability to ingest, store, quality control, preserve, and access its vast environmental data holdings. It is based upon a focused effort to ensure the Information Technology (IT) infrastructure is in place and working before the arrival of significantly larger and more complex environmental data (e.g., NPP/NPOESS and GOES-R). As such, it acts as an assurance of environmental data availability to the Nation through the development of a virtual mirror-site.

7. REFERENCES

Consultative Committee for Space Data Systems (CCSDS) *Reference Model for an Open Archive Information System* (CCSDS 650.0-B-1), May 2004

NOAA, CLASS Concept of Operations (1004_V_1.0_CLASS Concept of Operations), 5 April 2002.

NOAA, CLASS Dual Site Architecture & Operational Concept (1038_V_1.0_Dual Site Operations), 30 June 2003.