CLASS PLANS PROVIDING ARCHIVE, ACCESS, AND DISTRIBUTION SERVICES FOR GOES-R DATA

Robert Rank[1], Scott McCormick[2], Constantino Cremidis[3]

[1]NOAA/NESDIS/OSD/GSD, USA, [2] Global Science & Technology, Inc., USA, [3]CSC, USA

ABSTRACT

The Comprehensive Large Array-data Stewardship System (CLASS), a National Oceanic and Atmospheric Administration (NOAA) Information Technology (IT) enterprise solution, supports NOAA's data archive and science data stewardship missions by providing the IT portion of an archive for environmental data. Currently, CLASS is part of a major 10-year growth program to add new data sets and functionality to support a broader user base. The CLASS requirements to support this growth are defined by NOAA's data centers in conjunction with their users and with data producers to ensure that science and other user requirements are clearly defined. These requirements are captured in documents that include requirements specifications, Interface Control Documents, and Submission Agreements that drive the changes needed to handle the future increases in user and data volumes. The paper will present CLASS' approach for other major data campaigns such as NPP, NPOESS, and NEXRAD, historical data, and its plans for going forward with the GOES-R data campaign.

Index Terms— CLASS, requirements, campaigns, stewardship, archive

1. METHODOLOGY

Placing data on-line for access via the Web is a high priority in accordance with the Federal Government's eGov initiative. Data storage and retrieval systems will continue to be upgraded to support effective and efficient access with special focus on Internet interfaces, and on-line data that support the objectives of the CLASS concept of operations and ensure that the Nation has access (including Section 508 compliance) to their data and information.

New satellite observation campaigns are being prepared for launch and operations. The volumes of data to be collected by these campaigns dwarf the data streams managed by existing archive and distribution systems within NESDIS. The size, number, and frequency of data sets to be stored and distributed will require significant expansion of capacity for moving, storing, processing, and distributing data. New and continuing remote-sensing campaigns include GOES, POES, DMSP, NEXRAD, NPP, JPSS, Jason-2/3, and METOP. Numerous in situ observation programs also contribute to the information processing challenge.

CLASS has also been designated as NOAA's IT enterprise solution that will support the long-term archive requirements for NOAA data and products. The CLASS concept was developed as a framework to provide integrated data support while accomplishing the needed capacity expansion.

1.1. CLASS Approach for New Data

Considering the importance and complexity of early identification of metadata requirements in particular and the problems associated with data management in general, the CLASS project has a team dedicated to work the details of adding new data into CLASS: the CLASS Requirements Definition Team. The CLASS Requirements Definition Team serves as the technical liaison between the CLASS team and the NOAA data centers and external stakeholder organizations during the process of negotiating and defining data management requirements for both potential and new data campaigns. The CLASS Requirements Definition Team also performs gap analysis and requirements management for these campaigns and works with system engineers to conduct impact assessments.

1.2. CLASS Experience with New Campaigns

The CLASS Requirements Definition Team is currently engaged in two large new campaigns and has recently completed two others. The two current campaigns are the National Polar-orbiting Operational Satellite System (NPOESS), now the Joint Polar Satellite System (JPSS), and the Geostationary Operational Environmental Satellite (R-series). The initial satellite in the NPOESS program is the NPOESS Preparatory Project (NPP) satellite scheduled to launch in late 2011. The CLASS RDT has worked with the NPP Integrated Program Office (IPO) and NOAA's National Climatic Data Center (NCDC) to prepare the CLASS system for the NPP satellite launch. This began with the development of system and technical requirements for the ingest, archival storage, data management, access, and

dissemination of the NPP instrument data and products, auxiliary and ancillary data, software, documentation, and calibration/validation data.

Given the large volumes of data expected, system engineers on the CLASS team have performed thorough analyses of network bandwidth, hardware performance, disk and tape capacities, and server throughput to help ensure that overall system performance will support the ingest, access, and dissemination of 4TB of NPP data per day.

The campaigns recently completed are Jason-2 and the Meteorological Operational satellite programme (MetOp), a joint venture between the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) and NOAA. In all cases, CLASS has followed or is following the Open Archival Information System – Reference Model (OAIS-RM) [3].

1.3. Opportunities for Improvement

While working with these new campaigns, CLASS has learned a variety of lessons and has identified various opportunities for improvement. The two most important opportunities for improvement identified by the CLASS Requirements Definition Team are:

First, although the OAIS-RM indicates that a Submission Agreement is necessary between the archive and the producer, CLASS' experience with these three campaigns show that two additional documents are, in some occasions, necessary to complement the Submission Agreement. These two documents are: 1) an Interface Control Document (ICD) documenting the technical aspects of the electronic data transfer including defining network channels, file and data formats, processing workflow, and acknowledgment protocol, and 2) an Operations Agreement (OA) to document the operational aspects of data transfers such as IP addresses and routine and emergency point of contacts. This is due to the fact that for most very large data campaigns, the producer of the data is a science organization that it is concerned with the science aspects of producing the data, while a different group is concerned with the technical and operational aspects of timely creation and distribution of the data. Although both science and operational organizations could be under the same parent organization, CLASS' experience is that there are enough differences to warrant the development of three different, albeit related, documents.

Second, during the early development of new campaigns, the User Community and the Archive, represented by the NOAA National Data Centers (NNDCs) and CLASS, need to actively engage the data producer to define standards for data formats and metadata, and to produce the necessary documentation that must be made

available once the data is produced. The development of large data campaigns (like GOES-R) presents a unique opportunity to the user community to steer the data producer towards producing data in formats that are easy to work with, that are compatible with the tools used by the user community, and that have enough documentation to enable continued usage of the data even after the 'experts' have moved to new projects. This supports the notion of independent understandability so that the information in the archive is sufficiently complete to allow it to be interpreted, understood and used by consumers for many years to come.

The Archive's interests are similar to those of the user community, but for different reasons. In regard to data formats, the Archive is interested in data formats that (1) are directly usable by the user community, therefore not requiring the Archive to make data format conversions before distributing the data; (2) facilitate data preservation by not relying on specific software for processing it; and (3) facilitate archival storage by being compact and standard. A compact file is easier to receive, copy, archive, and distribute; a standard data format is easier to process using open source and other generally available tools capable of processing it and it is also likely that the Archive already has experience with it. In regard to metadata and documentation, the Archive is interested in the availability of collection and granule level metadata that facilitate data discovery, and enough documentation to ensure long-term information preservation.

In support of the GOES-R campaign, the CLASS RDT with support from many others within the CLASS project is actively engaged in a number of integrated product teams and working groups to develop interface requirements leading up to an ICD. Recent work has focused on use cases to ensure that nominal and non-nominal processes within the interface between the GOES-R product distribution system and the CLASS ingest function are designed properly. The requirements for this interface must also include a definition of the metadata that is needed by the archive for independent understandability.

The RDT is also developing functional requirements for the search, access, and dissemination of GOES-R data. It is working closely with the three NOAA National Data Centers to develop and refine these requirements. While a Submission Agreement is also being developed, a separate requirements document will provide the data centers and CLASS teams with a more structured set of requirements that can be baselined much earlier in the satellite campaign than the submission agreement.

2. CENTRAL CONCLUSIONS

The significant increase in high-quality digital data to be delivered by GOES-R, along with other large-volume data

contributors such as NPP/NPOESS and NEXRAD, have led NOAA to develop an agency-wide solution for large-array data archive and access. This new system, the Comprehensive Large-Array data Stewardship System (CLASS), is currently being implemented by the Office of Systems Development (OSD) and the three NOAA National Data Centers. The existing deployment of CLASS has operational nodes at the National Climatic Data Center (NCDC) in Asheville, NC and the National Geophysical Data Center (NGDC) in Boulder, CO. The centralized development/integration/test environments are maintained at OSD in the NOAA Satellite Operations Facility (NSOF) in Suitland, MD. The CLASS operational node at NGDC has the advantage of being accessible by the extensive research community in the Boulder region via broadband connection. NGDC recently implemented the National Lambda Rail to allow broad band data access via the commercial internet. The NSOF development/integration/test environment enhances the sharing of resources among development and integration environments.

Effective systems must be in place to support the scientific and research community, as well as address data management issues that are associated with massive volumes of data. Therefore, CLASS will implement an architecture for an integrated, national environmental data access and archive system to support ingest, archive, access, dissemination, and long-term stewardship requirements of the large array, remotely sensed data sets.

3. REFERENCES

[1] CLASS Project Team, *CLASS Level 1 Requirements Document and Mission Success Criteria*, Version 1.16, November 2008

[2] CLASS Project Team, *CLASS Target System Architecture Overview*, March 2009

[3] Consultative Committee for Space Data Systems (CCSDS), *Open Archival Information System – Reference Model (OAIS-RM)*, Issue 1.1, August 2009

[4] National Oceanic and Atmospheric Administration, *NOAA Enterprise Architecture*, Version 4, March 2009