CGMS

CGMS-35, NOAA-WP-13
Prepared by D. McGinnis
Agenda Item: I/1
Discussed in WG1

NOAA consideration of the Open Archival Information System Reference Model (OAIS-RM)

In NOAA-WP-13, NOAA informed the CGMS that the Comprehensive Large Array-data Stewardship System (CLASS) has adopted the Open Archival Information System (OAIS) Reference Model (OAIS-RM) to provide a framework and general guidelines in conducting its interactions with its data producers. The benefit CLASS perceives in using OAIS is that it provides a common set of functions, processes, and documents that are required to accomplish data transfers and a common terminology to establish the scope of the effort and the respective responsibilities of the data producers and the archive. To implement this for a pilot project between NOAA's CLASS and NASA, a workgroup was formed for the development of a data submission process using recommendations in the OAIS-RM. This case study will present the benefits and challenges that were encountered when enhancing CLASS to support the storage and distribution of NASA's Earth Observing System (EOS) data.

NOAA consideration of the Open Archival Information System Reference Model (OAIS-RM)

1. Introduction

The National Oceanic Atmospheric Administration (NOAA) and the National Aeronautics and Space Administration (NASA) worked together on a pilot project to initiate the transfer of archive responsibilities for EOS data to NOAA utilizing the Comprehensive Large Array-data Stewardship System (CLASS). To complete this task a workgroup was formed with personnel from both agencies. This workgroup started the development of a data submission process using recommendations for Space Data System Standards from the International Organization for Standards (ISO) Reference Model for an Open Archival Information System (OAIS), reference CCSDS 650.0-B-1, January 2002, (ISO 14721:2003).

CLASS has been identified by NOAA as its primary Information Technology (IT) system to support the secure storage, access, and distribution capabilities for multiple NOAA data collections. Currently, CLASS supports data from NOAA's Polar-orbiting Operational Environmental Satellites (POES), Geostationary Operational Environmental Satellites (GOES), Defense Meteorological Satellite Program (DMSP), and European Organisation for the Exploitation of Meteorological Satellites (Metop) missions and in the future it will hold data from the Jason-2/OSTM, NPP, NPOESS, and NEXRAD programs. CLASS has also been designated as the system that will support the long term archive services for the data collections of NASA's EOS missions that are designated as NOAA responsibilities.

The CLASS development team has adopted the OAIS Reference Model to provide a framework and general guidelines in conducting its interactions with its data producers. The benefit perceived by CLASS in using OAIS is that it provides a common set of functions, processes, and documents that are required to accomplish the data transfers and a common terminology to establish the scope of the effort and the respective responsibilities of the data producers and CLASS. The model identifies the need for a Submission Agreement (SA) between the producer, in this case NASA, and the archive, in this case NOAA.

While the NOAA/NASA workgroup has clearly benefited from its use of the OAIS Reference Model, it has had to adjust and tailor its implementation of the model to meet its specific needs. The model represents a synthesis of best practices across a wide range of archive programs and as a result is fairly high-level in its descriptions. The workgroup had to extend or define additional processes and documents to accommodate the details and complexity of the EOS long-term archive project. However, the Reference Model has definitely assisted the project and accelerated its early progress. The overall effort provides an excellent case study in the use of a standards-based process in the design and implementation of a long-term archive.

2. AGENCIES RESPONSIBILITIES

2.1 NOAA CLASS

The National Environmental Satellite, Data, and Information Service (NESDIS), a line office within the National Oceanic and Atmospheric Administration (NOAA), is responsible for archiving and disseminating environmental data collected by a variety of ground-based and space-based observing systems. The Comprehensive Large Array-data Stewardship System (CLASS) is NOAA's mechanism for securely archiving large-volume data and data products, and for making this data available to researchers, commercial users, and the public. The volumes of data collected by future satellite-based instruments and observation systems (as well as the growing customer demand for this data) will overwhelm the existing archive and distribution systems without a dramatic expansion in capacity: CLASS is the vehicle for the necessary expansion.

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, which will be integrated with a robust, large-volume, rapid-access storage and retrieval system that is capable of storing the incoming large-array environmental data, in situ data, and operational products as well as receiving users' on-line data requests, automatically processing these requests, and providing the requested data on the most appropriate media. This system will provide standardization in media, interfaces, formats, and processes for the very large datasets produced by satellites and radars. 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 an essentially "hands-off" operation.

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, Jason-2, NPP, NPOESS, and METOP; numerous in situ observation programs also contribute to the information processing challenge. CLASS has also been designated as NOAA's system that will support the long-term archive requirements for selected NASA EOS data products.

CLASS is a NESDIS initiative to develop and implement a single Information Technology (IT) system for the archive, access, and distribution capabilities of certain NOAA data over time. CLASS plans to support many disparate data types and data sets. CLASS will be hosted at multiple sites while providing a single Web interface to the user and tools for building custom interfaces. This multi-site capability is intended to improve system availability and scalability while enhancing data integrity through replication at geographically dispersed sites. CLASS is available at

www.class.noaa.gov.

2.2 NASA EOS

NASA's Earth Observing System (EOS) program was initiated in the 1980's and formally established in 1991 as a U.S. Presidential initiative to be NASA's contribution to the U.S. Global Change Research Program. Its purpose was to enable the study of the Earth as a system and it consisted of three major elements. The first was a constellation of satellites that would acquire long-term records of remotely sensed environmental variables on a global scale. The second was a comprehensive science research program that would produce and validate the satellite data products and use them in studies of Earth system science. The final component was the EOS Data and Information System (EOSDIS) that would provide the data and information systems and services to the EOS program during the active life of the individual EOS missions. By interagency understandings and agreements, the long term preservation and stewardship of the EOS data are the responsibility of NOAA and the U.S. Geological Survey (USGS) for atmospheric and oceanographic data products and for land processes data, respectively.

EOSDIS has been under development since the early 1990s. It is a distributed system that is deployed and operated at eight Distributed Active Archive Centers (DAACs) across the U.S. It supports the full range of data and information systems and services for EOS and other NASA Earth science programs including data production, active archive, search and order, data distribution, and direct data access. The data product generation occurs both at the DAAC facilities and at remote facilities developed and managed by the EOS scientists. More recently, these remote facilities have assumed data archive and distribution functions for selected EOS products. EOSDIS also includes a distributed framework that supports the interconnection and interoperability of these components.

3. EOS Long term pilot project

A Pilot Project was conducted with NASA to begin incorporating EOS data into NOAA CLASS using Moderate Resolution Imaging Spectroradiometer (MODIS) Level 0 data from the Terra and Aqua missions. The pilot project utilized the process defined in the Producer-Archive Interface Methodology Abstract Standard (PAIMAS); reference CCSDS 651.0-B-1, May 2004, which identified requirements, design elements, implementation issues, and responsibilities associated with preservation. This pilot project is an important step in archiving valuable NASA satellite information in CLASS.

The goal of the Pilot Project was to use the MODIS Level 0 and associated ancillary data products to guide the specification, implementation and test of an interface that would support the transfer of these products to CLASS. During the test of this interface, the forward streams of MODIS Level 0 data from both the Terra and Aqua missions (~70 GB/day for each mission) plus several smaller mission orbit and attitude products were transferred and ingested into CLASS over a four day period. In addition, historical MODIS data were retrieved from archive and staged to support a stress test of the interface. Over a six hour period, data were transferred and ingested at six times the forward stream of the two platforms.

The workgroup determined that several components were out of scope of the pilot and therefore not addressed. These included the documentation to describe the MODIS data and the software used to generate the data. CLASS did not examine whether the available MODIS documentation would be sufficient for supporting the long-term independent understandability of the data. In addition, because this was a pilot project, the SA did not include a definition of the ownership of the data. Both of them would have been addressed if this were not a pilot project.

3.1 Initial Contact

There is a notion within OAIS-RM that it is wise to tell the Data Producer that the Archive is expecting to follow the OAIS Reference Model in its interactions with the producer and that there is a formal negotiation of content, roles, and responsibilities that will follow the initial contact. CLASS followed this recommendation and suggested the adoption of the OAIS-RM framework to guide the efforts during the pilot project. In fact, this was CLASS's first implementation of the OAIS-RM for the negotiation between CLASS and a data producer.

3.2 Negotiation

NASA and NOAA used the iterative development of a data submission agreement (SA) as the negotiation mechanism for the Pilot Project. Initially, the CLASS team provided NASA with an outline for the SA that represented a comprehensive specification of the data transfer, including the high-level and detailed functional and performance specifications, interface mechanisms, operational processes and responsibilities, and expected levels of service to be provided. During the negotiation process, it was decided to use the SA as the high level agreement between the two agencies on the overall level of service to be provided and to separate the other material into subsidiary documents, tailored to the needs of the Pilot Project, and including a Requirements Specification, an Interface Control Document (ICD) and an Operations Agreement (OA). While the CLASS program has its own set of requirements that address IT support to all of its designated campaigns, this Requirements Specification identified requirements focused on support of NASA's EOS data and defining the Pilot Project. The ICD would govern the implementation of the interface between the NASA archive and CLASS for the Pilot Project and the OA would list the operational systems, procedures and points of contact.

While these decisions were made during the Pilot Project negotiation process, they should be generally applicable to the transfer of data between NASA and the NOAA CLASS system. As the high-level, interagency agreement, the SA would require the appropriate signatures at the Program level and would take precedence over the subsidiary documents. The ICD and OA would be signed and maintained at the Project level, enabling a more effective response during the implementation process. For the true, operational data transfers from NASA to the NOAA long-term archive, requirements would need to be developed at both the Program and Project level.

Metadata in the SA

One example of the tailoring that CLASS implemented was the inclusion of additional

metadata in the Submission Agreement. CLASS is required to provide Federal Geographic Data Committee (FGDC) compliant metadata that describe CLASS holdings. The ISO recommendation points out that much of that information is “more easily available or only available at the time when the original information [science data] is produced” thus stressing the need for data producers to play an active role in creating and maintaining standards compliant metadata. Metadata are necessary for many purposes, such as calibration and data processing. This requires open standards that allow for a great degree of flexibility in data storage and retrieval. The experience of the NOAA National Data Centers (NNDC) is certainly consistent with this need. These data centers now find it difficult or impossible to create metadata for older data holdings that were not well documented when they were originally provided to the data center. In light of these observations, CLASS submission agreements should seek to clarify metadata requirements, both from the producer and the archive. Whenever possible, these requirements should be addressed in standards-compliant ways. For example, when contact information is provided, the FGDC contact template should be used as an indicator of the required information. When parameters in the data products are described, those descriptions should include the FGDC required information at a minimum. This approach will certainly lead to improvements in the quality of metadata available to CLASS users and in the scientific usability of the data that CLASS provides.

For this pilot project, the workgroup modified the SA to include additional metadata information to better meet the FGDC standard. In fact, CLASS has also modified its standard SA template to include these metadata fields.

4. USE OF OAIS REFERENCE MODEL – BENEFITS AND CHALLENGES

The use of the OAIS Reference Model has served the NOAA/NASA MODIS Pilot Project in a number of ways. In what might first appear to be a simplistic example, the Reference Model has given the data producer and the archive a common terminology with which to frame the discussion. It has also identified a common set of functions that are required to archive data and information and a set of processes to establish the specific requirements that are associated with those functions. Further, it identifies a set of documentation to capture and record those requirements and specifications. Although some of these functions and processes have had to be tailored and extended to meet the needs of the pilot project, the Reference Model has provided an excellent foundation on which to build. Specifically, the SA proved to be a very useful tool in defining the submission and dissemination processes.

One source of difficulty in the direct application of the OAIS Reference Model to the long-term archive of EOS data in CLASS is that, in this case, the data producer is not the original data producer but rather is an existing, operational archive itself. The active archives of EOSDIS have been developed over the last decade and have established their own approaches, designs, processes, and documentation to support the archive of data received from the source providers and to disseminate that data to users. This requires the additional step of mapping and adapting the conventions of EOSDIS to the OAIS model being used by CLASS. Additionally, EOSDIS was designed to support the active archive function that includes an ongoing interaction with the EOS science teams as they refine their data processing algorithms, reprocess the raw data, and submit updated versions of the science

products to the archive. During this phase in the data life cycle, the focus is on production of the best science products and not necessarily on their long-term preservation. Some of the steps that ensure the proper stewardship of the data, including the generation of documentation and restructuring data collections must be done after the fact.

Another challenge encountered was that the SA proved clumsy when dealing with the complexity of the science products. Since the decision was made to initially transfer lower level instrument data, the data to generate higher level science products required orbit and attitude data for the Terra and Aqua satellites that carry the MODIS instrument. Furthermore, ancillary data sets that come from different sources and are in different formats for each satellite had to be stored. They exist in EOSDIS in both raw and processed forms that must be linked with instrument data and the processing software used to generate the higher-level science products. In order to document these complexities and inter-relationships, the SA structure required significant enhancements.

5. Summary

The OAIS Reference Model has been incorporated into NOAA's approach in the development of CLASS and was employed in the pilot project as a forerunner to NOAA's assumption of the responsibility as the long-term archive for EOS data. The project team used the OAIS mechanisms to establish and document the interagency transfer of data and found those concepts to be very useful in structuring the discussions and capturing a complete set of goals and requirements. The pilot project was the first opportunity that CLASS had to exercise the OAIS mechanisms and to learn the details of the Reference Model. CLASS also learned that even for the few MODIS data sets that were addressed, the long-term preservation and stewardship of EOS data and all of its associated information was an extremely complex problem. The OAIS Reference Model helped the pilot project and the experience and lessons learned by the project team should assist future users of the model. Critical to enabling interoperability with other digital archives is our compliance with the OAIS reference model.

To summarize, the OAIS reference model is applicable to all digital archives, their Producers and Consumers. It established common terms and concepts for comparing archival concepts and implementations, but it does not specify a particular implementation. It identifies a minimum set of responsibilities that must be discharged for an archive to call itself an OAIS archive. It provides detailed models for archival function and for the information associated with archives. Although not discussed in this paper, it also provides perspectives on migration, emulation and interoperability among OAISs.