CGMS

CGMS-35, NOAA-WP-39
Prepared by B. Barkstrom
Agenda Item: II/2
Discussed in WGII

Use of the Open Archive Information System Reference Model for Long-Term Preservation of Historical Data Sets and for Enhancing Interoperability for Future Data Sets Including GEOSS

NOAA-WP-39 describes how the Open Archive Information System Reference Model (OAIS RM) provides a widely accepted basis for enhancing the prospects for long-term preservation of historical data sets and for enhancing interoperability for future data sets including those created as part of the GEOSS program. There are three areas where the OAIS RM provides a solid basis for proceeding:

The Reference Model provides a clear description of the roles and responsibilities of an archive data producer and the user community (identified in the RM as the "Designated Community).

The Reference Model provides an abstract model for the couplings that must exist between data content and the descriptive information that long-term preservation requires.

The Reference Model gives a reasonably complete description of the functions an archive must enact in order to ensure information understandability over the long periods involved in archival preservation. These functions include preservation planning – which then encompasses the planning needed to deal with IT obsolescence and evolution.

Use of the Open Archive Information System Reference Model for Long-Term Preservation of Historical Data Sets and for Enhancing Interoperability for Future Data Sets Including GEOSS

1 INTRODUCTION

The Open Archive Information System Reference Model (OAIS RM) [CCSDS, 2002] provides an ISO standard description of an archive's environment and of the functions an archive needs to perform. In this paper, we deal with the features of the Reference Model that appear most important to enhancing the long-term preservation of historical data sets and for enhancing interoperability of future data sets. We are particularly concerned with identifying the way in which the Reference Model enhances our understanding of the information technology (IT) that we can apply to the problems of preserving information.

In the past year, NOAA has moved toward using the OAIS RM as a standard to describe the relations between the NOAA National Data Centers and the Comprehensive Large Array-data Stewardship System (CLASS). Based on this ISO standard, the Data Centers are the archives, while CLASS becomes the provider of the enabling Information Technology (IT). This change has clarified the roles and responsibilities of the Data Centers and of CLASS. Use of the ISO standard has also provided substantial clarity in dealing with the issues of data and metadata organization.

2 MAIN TEXT

The OAIS RM identifies the archive as being responsible for obtaining Submission Agreements with the Data Producers who provide the archive with its data, as well as identifying the Designated Community for whom the archive's contents need to be understandable over the long term. This definition of responsibility also requires that the archive quantify, as well as possible, the engineering requirements for ingest and for data access and ordering. The Submission Agreement (as well as the expectations of the Designated Community) also create the background for dealing with the data organization within the archive and for providing access aids that assist users in finding and exploiting data.

Setting Input and Output Engineering Expectations

One of the key benefits of the OAIS RM is improved quantification of input and output expectations for the archive and its IT components. The Submission Agreement of the OAIS RM calls for the specification of the data model that the archive will ingest, as well as for the schedule of Submission Sessions during which the archive receives the data and metadata. In most cases, the archive will be receiving files as input. As a result, the archive can estimate the requirements for ingest throughput on a firm, contractual-type basis, substantially improving the input engineering requirement specification.

Equally important, the archive can quantify its identification of the communities who need data and services and can quantify these expectations, rather along the lines of modern approaches to marketing. Such approaches call for the archive to quantify their expectation of the number of potential users for each kind of data and then to estimate an adaptation model that provides the rate of data search sessions and the rate at which user communities receive data.

Defining Collection Organization and Descriptive Information (Metadata)

A second key area where the OAIS RM assists the archive and its IT components lies in suggesting a structure and nomenclature for describing file collections and their associated metadata. While the OAIS RM is rather abstract, and thus does not provide detailed Earth science metadata descriptions (which are left to such standards as ISO 19115:2003, ECS, or Unidata metadata prescriptions), it does provide a Unified Modeling Language framework for organizing the Information Packages that contain the data and metadata submitted to the archive and stored there. The discussion of collections in the OAIS RM is clearly improved over other discussions of collections in the standards, allowing a clear separation between the rapidly changing, individual file inventory metadata and the much more slowly changing collection-level metadata.

Long-Term Information Preservation Issues

While there is a great deal of information about current issues in Earth science archives, the OAIS RM encourages us to also think about several long term issues that place information at risk. Indeed, the requirements for long-term preservation are so stringent that it is clear that we will need to develop much more systematic approaches to managing the risk of loss. This suggests adopting an approach that uses the three standard components of risk management:

1. Quantify the threats that may cause information loss
2. Estimate the probability of each threat and the probable loss as a result
3. Identify the most cost-effective strategy for managing the risk

With some work, we can use this approach even in the face of such threats as losses due to software obsolescence (particularly with respect to proprietary software), operator error, hardware data corruption, and institutional stability (as the Library of Congress expressed the problem). The latter is perhaps the most difficult to deal with, as it seems likely to require strategies for building trust and replicating data across many institutions, agencies, and even governments.

3 CONCLUSIONS

Overall, the use of the OAIS RM appears to offer substantial advantages in clarifying institutional roles and responsibilities, improving our ability to formalize metadata and apply standards, as well as think about long-term preservation issues that might otherwise escape our vision until it was too late to prevent data loss.