

The Open Archival Information System (OAIS) Reference Model and its Usage

Donald Sawyer¹, Lou Reich², David Giaretta³, Patrick Mazal⁴, Claude Huc⁴, Michel Nonon-Latapie⁴,
Nestor Peccia⁵

¹ National Aeronautics and Space Administration, ² Computer Sciences Corporation, ³ British National Space Centre, ⁴ Centre National d'Etudes Spatiales, ⁵ European Space Operations Centre/European Space Agency

1.0 Introduction

The OAIS Reference Model was developed by the Consultative Committee for Space Data Systems (CCSDS) as a work item under the ISO Technical Committee 20, Sub-committee 13. It is a framework for understanding and applying concepts needed for long-term digital information preservation (where long-term is long enough to be concerned about changing technologies). It is also a starting point for a model addressing non-digital information. It does not specify any implementation.

This model is targeted to several categories of reader:

- Archive designers,
- Archive users,
- Archive managers, to clarify digital preservation issues and assist in securing appropriate resources, and
- Standards developers.

The model has already been widely adopted as a starting point in digital preservation efforts:

- Digital libraries (e.g., Netherlands National Library),
- Traditional archives (e.g., US National Archives),
- Scientific data centers (e.g., National Space Science Data Center), and
- Commercial organizations (e.g., Aerospace Industries Association preservation working team)

It has recently been published (reference [1]) by the CCSDS as a final standard (Blue Book). It is also in the final stage of publication as an ISO standard and will be identified as ISO 14721: 2002.

2.0 Open Archival Information System

What is meant by an “Open Archival Information System?” ‘Open’ is simply referring to the fact that this standard was developed in an open forum and is freely available.

The “Information” part is more difficult and can have subtle ramifications. For now, information is simply any type of knowledge that can be exchanged, and that data refers to the way this knowledge is represented in the exchange. This will be expanded upon later.

The phrase “Archival Information System” is used to refer not only to the hardware and software, but also the people who are involved in acquiring information, preserving it, and making it available to those needing the information.

There are many terms that need to be used in well defined ways in order to construct a reference model. The OAIS has a glossary of such terms, and a few of the more important of these are defined below when they are needed.

2.1 Environment Model

The modeling starts by giving a view, as shown in Figure 1, of the OAIS as a box with three primary interfaces.

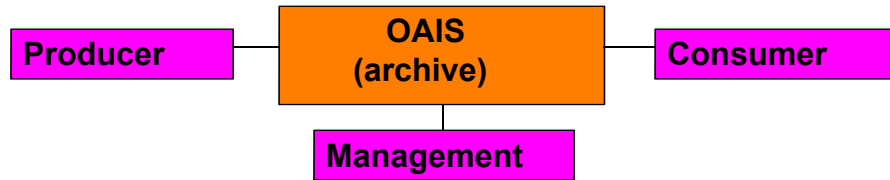


Figure 1: OAIS Environment Model

- Producers play the role of those who provide the information to be preserved
- Management plays the role of those who set overall OAIS policy where the OAIS is only one of their concerns. Day-to-day administration of the OAIS is handled by an Administration function within the OAIS box.
- Consumers play the role of those who interact with the OAIS services to find information of interest and to access this information.

Later, the OAIS box will be expanded into six functional areas. Although not described here, the OAIS Reference Model also identifies a minimum set of responsibilities that must be discharged for an archive to call itself an OAIS archive.

2.2 Information Modeling

As mentioned above, information is expressed by some type of data. It is the interpretation of the data, using additional representation information, that yields the information desired. This is shown in Figure 2 schematically. Consider a simple example to clarify the relationships.



Figure 2: An Information Object

Consider a data object to be a particular string of 128 bits in a file. Given the information that these bits are to be interpreted by applying the ASCII standard, an understanding the data (bit string) as a sequence of ASCII characters is obtained. This process has converted the data object (bit string), using the ASCII standard (Representation Information), into an Information object that is more meaningful than the

original bit string. Note that in order to preserve the information object, it is necessary to preserve not only the bit string, but also the ASCII representation information and the association between the two.

Of course the Representation Information may be much more complex than the ASCII standard, and so the Information Object may be much more complex than a sequence of characters.

A key information-modeling concept in the OAIS is the Information Package. Think of it as a container, as shown in Figure 3, which holds two types of information, called Content Information and Preservation Description Information.

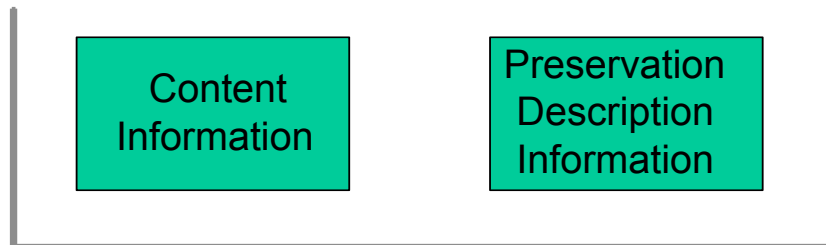


Figure 3: Information Package Definition

Note that each of these is an Information Object and thus will have its own Data Object and Representation Information. The Content Information's Data Object is referred to as the Content Data Object.

The Content Information is defined to be that information that is the original target of preservation. For example, suppose the objective is to preserve the content of a book in electronic form. It could be decided that the Content Information is all the information that allows a re-creation of a view of the book, from its cover through all the pages, including figures, etc. This could be constructed as, or received as, a single data file in Adobe's PDF format. This would be called the Content Data Object. The associated Representation Information, needed to provide the end view of the book, would be contained in the Adobe software as it has the information to map the bits of the file into the view that is to be preserved.

Alternatively, it might be that the book is really just text organized into chapters. It can be adequately represented simply as a text file with no need to use PDF or other complex formatting. Just what constitutes the Content Information to be preserved is not always obvious, and may need to be negotiated with the Producer.

Note that in the general case, the Content Data Object doesn't have to be a digital object. It could be a physical object, such as moon rock or a piece of film. The Representation Information would be used to add meaning about what was being preserved.

In addition to the Content Information, an Information Package may also contain a type of information called Preservation Description Information. The purpose of this information is to assist in preserving the Content Information, and it is broken down into four sub-categories.

- First, the Reference Information is used to provide one or more systems of identifiers by which to identify the Content Information. For example, this might include bibliographic attributes and/or a Digital Object Identifier.

- Second, the Provenance Information describes the history of the Content Information, including the chain of custody, so that Consumers can better judge how much to trust the information.
- Third, the Context Information relates the Content Information to other information outside the Information Package. This provides Consumers with an understanding of how the information being preserved relates to a wider environment.
- Finally, the Fixity Information is used to help ensure that the Content Information is not altered in an undocumented manner. For example, this might include checksums and digital signatures.

The Preservation Description Information is an essential part of the Information Package used by the OAIS for its preservation function.

While an Information Package typically contains two types of information, Content Information and Preservation Description Information, there are also three variants of the Information Package depending on where the package is being used in the OAIS environment.

- The first of these is the Submission Information Package, used to provide information to the OAIS by the Producer. Typically it is subject to negotiation between the two.
- The second of these is the Archival Information Package. It is used by the OAIS to hold the Content Information and Preservation Description Information as it performs its preservation function. Note that it may take several Submission Information Packages to form a single Archival Information Package, or one Submission Information Package may result in several Archival information Packages.
- The third of these is the Dissemination Information Package. It is used to provide requested information to the Consumer. Note that it may contain only a part, or all, of one or more Archival information Packages as determined by the OAIS in response to requests.

The use of the three variants of an Information Package are shown in Figure 4.

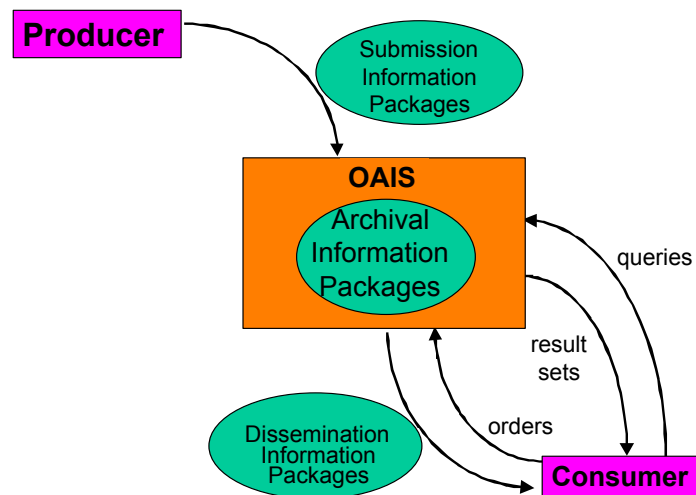


Figure 4: External Data Flow View

The Submission Information Package is submitted to the OAIS by a Producer. The OAIS holds and preserves the information using Archival Information Packages. In response to Consumer queries and resulting orders, Dissemination Information Packages are returned.

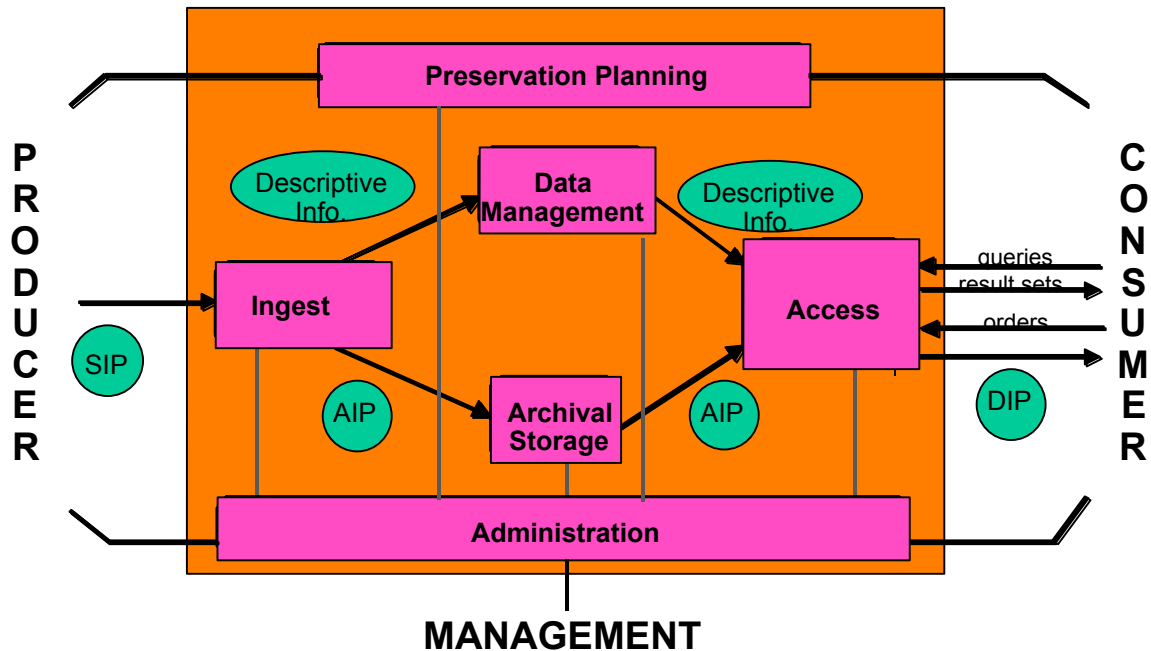
The OAIS reference model goes into additional detail regarding the modeling of an Archival Information Package, however it is not appropriate to present that detail here. Having looked at the information modeling aspects of the OAIS reference model, it is time to take a brief look at the modeling of archive functions.

2.3 Functional Modeling

Six primary functions have been identified, as previously noted.

- Ingest is the first, and this entity provides the major interface between the OAIS and the Producer. It accepts Submission Information Packages from Producers during a Data Submission Session. This session may be comprised of a delivered set of media, or it may be a single telecommunications session. The Submission Information Packages will conform to agreements reached between the Producer and the OAIS as defined in the Submission Agreement. Ingest prepares Archival Information Packages and Package Descriptions for storage and subsequent access.
- Archival Storage is the second, and this entity accepts Archival Information Packages, maintains them, and provides them upon request.
- Data Management is the third, and this entity accepts Package Descriptions from the Ingest function and other types of metadata needed to support overall OAIS operations.
- Administration is the fourth, and this entity is responsible for managing the overall operation of the OAIS on a day-to-day basis.
- Preservation Planning is the fifth, and this entity is responsible for monitoring technology evolution and the needs of the Designated Communities, and for forming preservation strategies and techniques to support the OAIS preservation function.
- Access is the last function, and this entity supports Consumers in identifying, locating, and accessing the information of interest.

The conceptual relationships of the six functional areas, along with the three variations of information packages, are shown in Figure 5.



SIP = Submission Information Package □
AIP = Archival Information Package □
DIP = Dissemination Information Package

Figure 5: OAIS Functional Entities

This Figure may be understood as follows:

Conceptually, a Submission Information Package is provided by a Producer to the Ingest entity. An AIP is created and delivered to Archival Storage. Related Descriptive Information is provided to Data Management. A Consumer searches for, and requests, information using appropriate Descriptive Information and access aids. The appropriate AIP is retrieved from Archival Storage and transformed by the Access entity into the appropriate Dissemination Information Package for delivery to the Consumer. This is all under the guidance of the Administration entity. Preservation strategies and techniques are recommended by Preservation Planning and put in place by the Administration entity.

Within the OAIS the functional entities are broken into sub-functions. The purpose is to more clearly identify the types of functions involved, not to promote a specific implementation. The readers should consult to the OAIS Reference Model for these details.

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.

3.0 Usage of OAIS Reference Model

There are a growing number of organizations that have adopted the "Reference Model for an Open Archival Information System (OAIS)" as a starting point for their digital preservation efforts. These

include the Networked European Deposit Library (NEDLIB) (reference [2]), the National Library of the Netherlands (reference [3]), and the British National Library (reference [4]). The latter two organizations have solicited vendor support in developing an implementation that addresses the OAIS concepts. The Research Library Group and the Online Computer Library Center have a number of projects underway, including work on trusted archives (reference [5]), that make extensive use of the OAIS model. The Library of Congress is hosting an XML based packaging standard for digital publications, known as METS (reference [6]), that recognized the OAIS information modeling concepts in the Archival Information Package. However the remainder of this paper will focus on the usage made by two space science archives, one in France and one in the United States.

3.1 Centre de Donnees de la Physique des Plasmas (CDPP)

The CDPP has developed a computer system called SIPAD (System for Preservation and Access to Data and Information). This has been developed and is being operated to fulfil the CDPP requirements for acquiring, managing, preserving, and making information related to Plasma Physics available to researchers. This includes multi-mission information from both space-based and ground-based instruments, and also related information such as orbit and attitude data, documents, bibliographic references, and catalogues. All the information that might be needed by scientific researchers to support the use of the Plasma Physics data are sought for inclusion within SIPAD.

A variety of access services are provided, including making data selections by time, mission/experiment, and keywords. Browse images and event tables are provided to help select interesting time intervals. A data request can include data from multiple data sets, and transformations of data may be performed before delivery via media or network. The related documentation and abstracts can be navigated and retrieved.

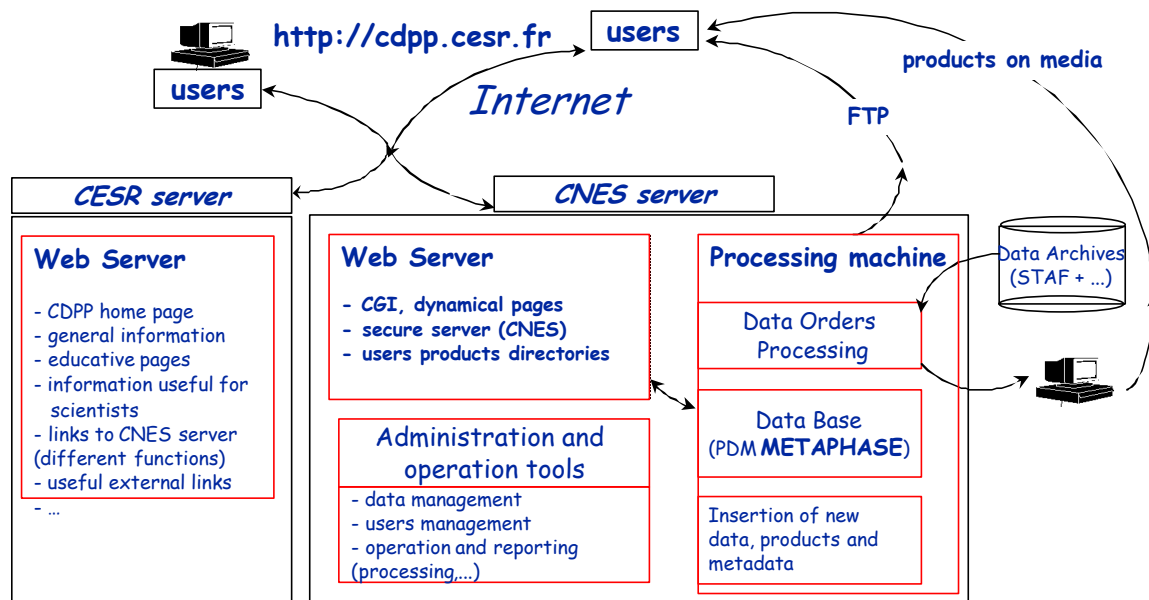


Figure 6: CDPP System Architecture

The system architecture, as shown in Figure 6, has been informed by the OAIS Reference Model as it existed in 1997/1998. This is when SIPAD was developed, and it recognizes the different functions called out in the OAIS model. Referring to Figure 6, the OAIS Ingest function is addressed primarily by

the CNES server Processing Machine function devoted to the insertion of new data, products, and metadata. The Data Management function is addressed by the Processing Machine Data Base. The Administrative function is addressed by the CNES Administration and Operation Tools function. The Access function is addressed by the CESR and CNES Web Servers, and by the Processing Machine's Data Orders Processing function. The Archival Storage function is primarily addressed by the external STAF function. STAF is a separate CNES facility devoted to data preservation at the file level. The location of these files and the media on which they are stored may be changed without impact on client systems such as SIPAD.

There are a number of additional standards used by SIPAD. These include CCSDS EAST descriptions of data and the CCSDS date/time format. These support data documentation, searching, and automated data processing with field extractions to satisfy requests. The CCSDS Data Entity Dictionary Specification Language (DEDSL) is also used to document the data entities and attributes supplied to the archive. These various types of metadata are delivered to the SIPAD using the CCSDS Parameter Value Language (PVL).

The strong standards approach allows the software to be generic with respect to multiple missions and information categories. This facilitates cost-effective evolution and secure ingest and processing of the data from the Producers.

3.2 National Space Science Data Center (NSSDC)

The National Space Science Data Center has been in existence since 1972 and has collected over 20 TB spread across over 4300 data sets. This is a very heterogeneous collection and includes Space Physics, Astrophysics, and Planetary data from mostly space-based instruments. Like CDPP, it collects a variety of supporting information including orbit and attitude data, documents, bibliographic references, and catalogues. It provides long-term and multi-mission data management, preservation, and dissemination to a variety of customers (Designated Communities) including Space Science researchers, the general public, and other archives. As examples for the later two categories, it provides the public with CD-ROMs from the Planetary Data System archive, and it provides a long-term archive and backup function to the High Energy Astrophysics Archive.

A variety of access services are provided including search functions based on time, mission/experiment, and relevant keywords. It provides anonymous FTP access to copies of a significant fraction of its Space Physics data, and it provides the ability to selectively retrieve and display plots of parameters from many of these data sets. The associated documentation may also be readily accessed.

The NSSDC has recently begun a re-engineering of various aspects of its archival system. This is described in an NSSDC Newsletter (reference [7]) and has taken advantage of the OAIS Reference Model concepts. This architecture includes a separation of functions along OAIS lines and the implementation of an Archival Information Package (AIP). Currently data are being migrated into AIP forms and given to the Data Ingest and On-line Access Subsystem (DIONAS) system for long-term storage and preservation on Digital Linear Tape. New data received may arrive already in AIP form, or they will be put into AIPs during the Ingest process. The IMAGE project is an example of the former as they use NSSDC supplied AIP packaging software to create packages with the needed metadata and then they transfer them via FTP to NSSDC. This greatly facilitates a cost-effective Ingest process.

Figure 7 shows a logical view of the architecture in terms that readily map to the OAIS Reference Model. The Archival Storage function is provided by the DIONAS software. It also keeps track of the files it has placed in the anonymous FTP data dissemination environment.

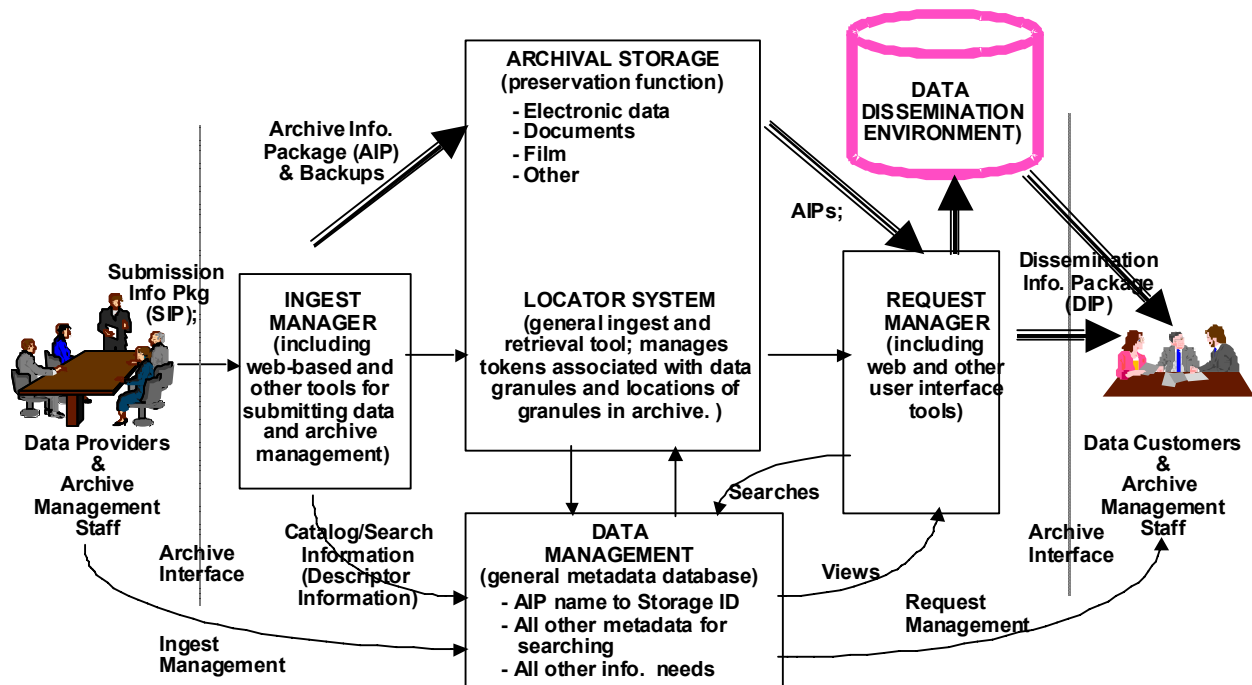


Figure 7: NSSDC Archive - Logical Architecture

This standards approach allows the data to be managed, and migrated to new media, independently of the nature of the data. The management software and data can be updated independently. In addition, the AIP implementation has allowed the specification of important metadata to accompany the data, and it provides the ability to give projects packaging software that significantly improves the NSSDC ingest efficiency.

4.0 Conclusion

The OAIS Reference Model has been rapidly adopted by a wide variety of organizations, particularly those that have not had digital archives in the past but now recognize they need them. However the usage among science archives, which have long had digital information, has also continued to grow. They are finding it useful as a checklist, and in some cases are re-engineering some of their processes to take advantage of OAIS concepts. Some vendors are also beginning to support the OAIS concepts and thus it can be said that this standard has already met the objectives laid out at the beginning of its development.

References

- [1] *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.-B-1. Blue Book. Issue 1. Washington D.C. January 2002. <<http://www.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf>>
- [2] *NEDLIB home page*. <<http://www.kb.nl/>>
- [3] *National Library of the Netherlands home page*. <<http://www.kb.nl/index-en.html>>
- [4] *British National Library home page*. <<http://www.bl.uk/>>

- [5] *RLG and OCLC Issue Final Report on Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report. Press Release.* May 2002. <<http://www.rlg.org/pr/pr2002-repositories.html>>
- [6] *Metadata Encoding & Transmission Standard home page.* <<http://www.loc.gov/standards/mets/>>
- [7] *NSSDC News, December 2002 Issue.* <http://nssdc.gsfc.nasa.gov/nssdc_news/dec00/dec00_toc.html>