Consultative Committee for Space Data Systems

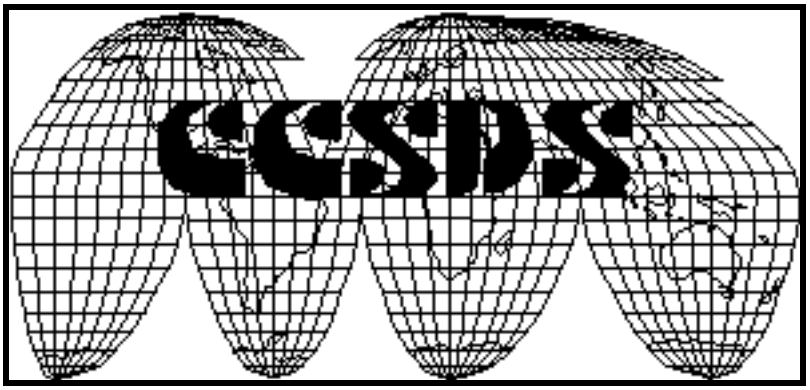
RECOMMENDATION CONCERNING SPACE DATA SYSTEMS STANDARDS

Reference Model for an Open Archival Information System (OAIS)

CCSDS 650.0-W-5.0

WHITE BOOK

April 21, 1999



Dear Reader:

The following version of the OAIS Reference Model is designated CCSDS 650.0-W-5.0. This version, called White Book 5.0, reflects the updates, both specific and general, agreed to at the ISO Toulouse International Workshop and in subsequent teleconferences. It also addresses some of the comments received at the Digital Archive Directions (DADs) workshop. It is prepared for review at the ISO Newport Beach International Workshop and, after what are expected to be limited changes, for release and formal review as a draft ISO standard and CCSDS Red Book.

Comments may be sent to:

Lou Reich
louis.i.reich@gsfc.nasa.gov

or

Don Sawyer
donald.sawyer@gsfc.nasa.gov

AUTHORITY

Issue:	White Book, Issue 5.0
Date:	October, 1998
Location:	Toulouse, France

This document, when it has been approved for publication by the Management Council of the Consultative Committee for Space Data Systems (CCSDS), will reflect the consensus of technical panel experts from CCSDS Member Agencies.

This document is published and maintained by:

CCSDS Secretariat
Program Integration Division (Code MG)
National Aeronautics and Space Administration
Washington, DC 20546, USA

FOREWORD

This document is a technical Recommendation for use in developing a broader consensus on what is required for an archive to provide permanent, or indefinite long-term, preservation of digital information. It may be useful as a starting point for a similar document addressing the indefinite long-term preservation of non-digital information.

This Recommendation establishes a common framework of terms and concepts which comprise an Open Archival Information System (OAIS). It allows existing and future archives to be more meaningfully compared and contrasted. It provides a basis for further standardization within an archival context and it should promote greater vendor awareness of, and support of, archival requirements.

Through the process of normal evolution, it is expected that expansion, deletion, or modification to this document may occur. This Recommendation is therefore subject to CCSDS document management and change control procedures, which are defined in Reference [1].

DOCUMENT CONTROL

Document	Title	Date	Status and Substantive Changes
CCSDS 650.0-W-1	Report Concerning Space Data Systems Standards: Reference Model for an Open Archival Information System (OAIS)	April 1997	Original Issue
CCSDS 650.0-W-1.1		July 1997	Revised information model and reduced text and complexity in Section 2. Partial response to Silver Spring Workshop directions.
CCSDS 650.0-W-1.2		Sept. 1997	Further revision of information model to incorporate concept of packaging information,. More complete response to Silver Spring Workshop directions.
CCSDS 650.0-W-2.0		Oct. 1997	Complete revisions re: Silver Spring Workshop instructions.
CCSDS 650.0-W-3.0		April 1998	Complete revisions re: Frascati workshop instructions
CCSDS 650.0-W-4.0		May, 1998	Complete revision from Frascati and Houston workshops
CCSDS 650.0-W-5.0		April, 1999	Complete revisions from Toulouse workshop and subsequent teleconferences

CONTENTS

1 INTRODUCTION

1.1 PURPOSE AND SCOPE

The purpose of this document is to define the ISO Reference Model for an **Open Archival Information System** (OAIS). An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for one or more **Designated Communities**. It meets a set of such responsibilities as defined in this document and this allows an OAIS archive to be distinguished from other uses of the term 'archive.' The term 'Open' in OAIS is used to imply that this standard and future related standards are developed in open forums, and it does not imply that access to the archive is unrestricted.

The information being maintained has been deemed to need **Long Term Preservation**, even if the OAIS itself is not permanent. **Long Term** is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely. In this reference model there is a particular focus on digital information, both as the primary forms of information held and as supporting information for both digitally and physically archived materials. Therefore the model accommodates information that is inherently non-digital (e.g. a physical sample) but the modeling and preservation of such information is not addressed in detail. This reference model:

- provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access.

- provides the concepts needed by non-archival organizations to be effective participants in the preservation process.

- provides a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future archives.

- provides a basis for comparing the data models of digital information preserved by archives and for discussing how data models and the underlying information may change over time.

- provides a foundation that may be expanded by other efforts to cover long-term preservation of information that is NOT in digital form (e.g., physical media, physical samples).

- expands consensus on the elements and processes for long-term digital information preservation and access, and it promotes a larger market which vendors can support.

- guides the identification and production of OAIS related standards.

The Reference Model addresses a full range of archival information preservation functions including ingest, archival storage, data management, access, and dissemination. It also addresses the migration of digital information to new media and forms, the data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives. It identifies both internal and external interfaces to the archive functions, and it identifies a number of high level services at these interfaces. It provides various illustrative examples and some 'best practice' recommendations. Finally, it attempts to define a maximal archive to provide a broad set of useful terms and concepts, but it defines a minimal set of responsibilities for an archive to be called an OAIS.

1.2 APPLICABILITY

The OAIS model in this document may be applicable to any archive. It is specifically applicable to organizations with the responsibility of making information available for the long term. This includes organizations with other responsibilities such as processing and distribution in response to programmatic needs.

This model is also of interest to those organizations and individuals who create information that may need long-term preservation and those that may need to acquire information from such archives.

The model, including the functional and information modeling concepts, are relevant to the comparison and design of facilities which hold information, on a temporary basis, for two reasons:

- When taking into consideration the rapid pace of technology changes or possible changes in a Designated Community, there is the likelihood that facilities, thought to be holding information on a temporary basis, will in fact find that some or much of their holdings will need the same type of attention as that given by permanent archives.

- Although some facilities holding information may themselves be temporary, some or all of their information may need to be preserved indefinitely. Such facilities need to be active participants in the long-term preservation effort.

Standards developers are expected to use this model as a basis for further standardization in this area. A large number of related standards are possible. A road-map for such development is briefly addressed in Section 1.4.

This Reference Model does not specify an implementation. Actual implementations may group or break out functionality differently.

1.3 RATIONALE

A tremendous growth in computational power, and in networking bandwidth and connectivity, has resulted in an explosion of organizations who are making information available in electronic forms. Transactions among all types of organizations are being conducted using electronic forms that are taking the place of more traditional forms such as paper.

Preserving information in electronic forms is much more difficult than for forms such as paper and film. This is not only a problem for traditional archives, but for many organizations that have never thought of themselves as performing an archival function. It is expected that this reference model, by establishing minimum requirements for an OAIS archive along with a set of archival concepts, will provide a common framework from which to view archival challenges, particularly as they relate to digital information. This should enable more organizations to understand the issues and to take proper steps to ensure long term information preservation. It should also provide a basis for more standardization and therefore a larger market that vendors can support in meeting archival requirements.

1.4 CONFORMANCE

This document specifies the minimum features of an OAIS for defining the serviced community, for fulfilling the responsibilities or services provided, and for maintaining the OAIS content.

A conforming OAIS archive implementation shall support the model of information described in Section 2.2. The OAIS Reference Model does not define or require any particular method of implementation of these concepts.

A conforming OAIS archive shall fulfill the responsibilities listed in Section 3.1. Section 3.2 expands on the range of possible activities and provides additional details on the requirements identified in Section 3.1.

It is assumed that implementers will use this reference model as a guide while developing a specific implementation to provide identified services and content. This document does not assume or endorse any specific computing platform, system environment, system design paradigm, system development methodology, database management system, database design paradigm, data definition language, command language, system interface, user interface, technology, or media required for implementation.

A conformant OAIS archive may stand-alone or may be implemented as part of a larger system which provides additional services to users that are beyond those required of an OAIS.

The OAIS Reference Model is designed as a conceptual framework in which to discuss and compare archives. As such, it attempts to address all the major activities of an information preserving archive in order to define a consistent and useful set of terms and concepts. A

standard and other document that claims to be conformant to the OAIS Reference Model shall use the terms and concepts defined in the OAIS Reference Model in same manner.

1.5 ROAD-MAP FOR DEVELOPMENT OF RELATED STANDARDS

This Reference Model, developed by CCSDS Panel 2 in response to ISO TC20/SC 13, serves to identify areas suitable for the development of OAIS related standards. Some of these standards may be developed by Panel 2 ; others may be developed by other standardization bodies. However, any such work undertaken by other bodies should be coordinated in order to minimize incompatibilities and efforts. Areas for potential OAIS related standards include:

- standard(s) for the interfaces between OAIS type archives.
- standard(s) for the submission (ingest) of digital data sources to the archive.
- standard(s) for the delivery of digital sources from the archive.
- standard(s) for the submission of digital metadata, about digital or physical data sources, to the archive.
- standard(s) for the identification of digital data within the archive
- protocol standard(s) to search and retrieve metadata information about digital and physical data sources.
- standard(s) for media access allowing replacement of media management systems without having to re-write the media
- standard(s) for specific physical media
- standard(s) for the migration of information across media and formats
- standard(s) for recommended archival practices
- standard(s) for accreditation of archives

1.6 DOCUMENT STRUCTURE

1.6.1 HOW TO READ THIS DOCUMENT

All readers should read the Purpose and Scope (1.1), Applicability (1.2), and Conformance (1.4) sections to obtain a view on the objectives and applicability of the document.

Those who want just an overview of the major concepts should also read OAIS Concepts (2.0) and OAIS Responsibilities (3.0).

Those who will implement OAIS archives or administer them on a daily basis should read the entire document.

1.6.2 ORGANIZATION BY SECTION

Section 1 provides purpose, scope, applicability, and definitions sub-sections typical of many standards. It also provides rational for the effort, conformance requirements, and a road-map for development of related standards.

Section 2 provides a high level overview of the major concepts involved in an OAIS archive. It provides a view of the environment of an OAIS archive and the roles played by those who interact with it. It discusses what is meant by "information" and what is necessary to preserve it for the long term. It contains key information concepts relevant to OAIS conforming implementations.

Section 3.1 defines mandatory responsibilities an OAIS archive must discharge in preserving its information, and section 3.2 provides clarifying material of the types of activities that may be needed in many archives to discharge these responsibilities.

Section 4 provides model views needed for a detailed understanding of an OAIS archive. It breaks down the OAIS into a number of functional areas and it identifies some high level services at the interfaces It also provides detailed data model views of information using UML diagrams.

Section 5 provides some perspectives on the issues of information preservation using digital migration across media and across new formats or representations. It also provides some perspectives on the issues of preserving access services to digital information using software porting, wrapping, and emulation of hardware.

Section 6 is an introduction to the various alternatives for archive to archive associations to provide increased or more cost-effective services.

The annexes are not part of the Recommendation and are provided for the convenience of the reader.

Annex A provides scenarios of existing archive operations.

Annex B relates parts of this reference model to other standards work.

Annex C. provides a brief tutorial on the Unified Modeling Language (UML).

Annex D provides a list of informative references.

Annex E provides a layered model of information.

1.7 DEFINITIONS

1.7.1 ACRONYMS AND ABBREVIATIONS

AIC - Archival Information Collection
AIP - Archival Information Package
AIU - Archival Information Unit
ASCII - American Standard Code for Information Interchange
CAD - Computer-Automated Design
CCSDS - Consultative Committee for Space Data Systems
CD-ROM - Compact Disk - Read Only Memory
CEOS - Committee on Earth Observing Satellites
CIP - Catalog Inter-operability Protocol
CRC - Cyclical Redundancy Check
DED - Data Entity Dictionary
DBMS - Data Base Management System
DDL - Data Description Language
DIP - Dissemination Information Package
DVD - Digital Video Disk
EBCDIC - Extended Binary Coded Decimal Interchange Code
ECS - EOSDIS Core System
EOSDIS - Earth Observing System Data and Information System
FITS - Flexible Image Transfer System
GIF - Graphics Interchange Format
HFMS - Hierarchical File Management System
ICS - Interoperable Catalogue System
IEEE - Institute of Electrical and Electronic Engineers
IMS - Information Management System
ISBN - International Standard Book Number
ISO - Organization for International Standardization
LSDA - Life Sciences Data Archive
NARA - National Archives and Records Administration
NASA - National Aeronautics and Space Administration
NSSDC - National Space Science Data Center
OAIS - Open Archival Information System
ODL - Object Description Language
PDI - Preservation Description Information
PDMP - Project Data Management Plan
PDS - Planetary Data System
PSDD - Planetary Science Data Dictionary
SIP - Submission Information Package
UML - Unified Modeling Language
UNICODE - Universal Code
WWW - World-Wide Web

1.7.2 TERMS

There are many terms used in this reference model which need to have well defined meanings and these are defined in this section. When first used in the text, they are shown in bold and are capitalized. Subsequent use employs capitalization only.

As this reference model is applicable to all disciplines and organizations that do, or expect to, preserve and provide information in digital form, these terms can not match all of those familiar to any particular discipline (e.g., traditional archives, digital libraries, science data centers). Rather, the approach taken is to use terms which are not already overloaded with meaning so as to reduce conveying unintended meanings. Therefore we expect all disciplines and organizations will find that they need to map some of their more familiar terms to those of the OAIS Reference Model. This should not be difficult and is viewed as a contributor, not a deterrence, to the success of the Reference Model. For example, archival science focuses on preservation of the 'record'. This is an entity that documents some type of transaction. This term is not used in the OAIS Reference Model, but one mapping might equate it with the OAIS terms of 'content information' in an 'archival information package' (see definitions below and sections 2.2 and 4.2 for context)

Access: This OAIS entity contains the services and functions which make the archival information and externally-available services visible to Consumers.

Access Aids: Software or documents that allow Consumers to locate, analyze, and order Archival Information Packages of interest.

Access Collection: A collection of AIPs that is defined by a Collection Description but for which there is no Packaging Information for the collection in Archival Storage.

Access Methods: A method for retrieving an Archival Information Package based on its name or identifier which is available to authorized users.

Adhoc Order: A request that is generated by a Consumer for information the OAIS has indicated is currently available.

Administration: This OAIS entity contains the services and functions needed to control the operation of the other OAIS functional entities on a day to day basis

Archive: An organization that intends to preserve information for access and use by one or more Designated Communities.

Archival Storage: This OAIS entity contains the services and functions used for the storage and retrieval of Archival Information Packages.

Archival Information Collection (AIC): An Archival Information Package whose Content Information is an aggregation of other Archival Information Packages.

Archival Information Package (AIP): An information packaging concept that requires the

presence of Content Information and all the associated Preservation Description Information that is needed to preserve the Content Information over the long term. It has associated Packaging Information.

Archival Information Unit (AIU): An Archival Information Package whose Content Information is not further broken down into other Content Information components, each of which has its own complete Preservation Description Information. It can be viewed as an "atomic" AIP. An example of an AIU would be a table of numbers representing temperatures in a certain region with all the associated documentation describing how and where the temperatures were measured, what instruments were used to make the measurements, who made the measurements, why they were made, what processing has been performed on the measurements and who has had custody of these measurements since they were first created, how the measurements relates to other information, how the measurement can be uniquely referenced by others, etc.

Associated Descriptions: Information describing the content of an Information Package from the point of view of a particular Access Aid.

Client: An application which exchanges information with another application (see also Consumer).

Collection Description: A type of Package Description that is specialized to provide searchable information about a collection.

Common Services: Supporting services such as inter-process communication, name services, temporary storage allocation, exception handling, security, and directory services necessary to support the OAIS.

Component: An independent, separately identifiable, part of a Content Information or PDI Information Object

Consumer: The role played by those persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail.

Content Information: That set of information that is the primary target for preservation. It is distinguished from Preservation Description Information which is used to assist in the preservation of the Content Information. An example of Content Information could be a single table of numbers representing, and understandable as, temperatures but it excludes the documentation which would explain its history and origin, how it relates to other observations, etc.

Context Information: Information that documents the relationships of the Content Information to its environment. This includes why the Content Information was created and how it relates to other Content Information objects existing elsewhere.

Data: The representation forms of information. Examples of data include a sequence of

bits, a table of numbers, the characters on a page, the sounds made by a person speaking, a moon rock.

Data Submission Session: A delivered set of media or a single telecommunications session that provides Data to an OAIS. The Data Delivery Session format/contents is based on a data model negotiated between the OAIS and the Producer in the Submission Agreement. This data model identifies the logical constructs used by the Producer and how they are represented on each media delivery or in the telecommunication session.

Data Dissemination Session: A delivered set of media or a single telecommunications session that provides data to a Consumer. The Data Dissemination Session format/contents is based on a data model negotiated between the OAIS and the Consumer in the Request Agreement. This data model identifies the logical constructs used by the OAIS and how they are represented on each media delivery or in the telecommunication session.

Data Dictionary: A formal repository of terms used to describe data.

Data Management: This OAIS entity contains the services and functions for populating, maintaining, and querying a wide variety of information such as catalogs and inventories on what may be retrieved from Archival Storage, processing algorithms that may be run on retrieved data (if any), consumer access statistics, security controls, OAIS schedules and procedures.

Data Management Data: Data created and stored in Data Management persistent storage that refer to operation of an archive. Some examples of this data are accounting data for consumer billing and authorization, policy data, subscription data for repeating requests, and statistical data for generating reports to archive management.

Data Object: Either a Physical Object or a Digital Object.

Designated Community: An identification of a set of potential Consumers who should be able to understand a particular set of information.

Descriptive Information: That set of information, consisting primarily of Package Descriptions, which is provided to Data Management to support the finding of preserved information by Consumers.

Digital Object: An object composed of a set of bit sequences.

Digital Migration: The transfer of digital information, while intending to preserve it, within the OAIS. It is distinguished from transfers in general by three attributes:

- a focus on the preservation of the full information content,
- a perspective that the new archival implementation of the information is a replacement for the old, and
- full control and responsibility over all aspects of the transfer resides with the OAIS.

Dissemination Information Package (DIP): An Information Package that contains parts or all of one or more AIPs and that is distributed to the Consumer as requested.

Event Based Order: A request that is generated by a Consumer for information that is to be delivered periodically on the basis of some event or events.

Finding Aid: A type of Access Aid that allows a user to search for and identify Archival Information Packages of interest.

Fixity Information: This information documents the authentication mechanisms and provides authentication keys to ensure that the Content Information object has not been altered in an undocumented manner. An example is a CRC code for a file.

Format: The sequential organization of data in terms of its components.

Information: Any type of knowledge that can be exchanged. In an exchange, it is represented by data. Often the representation used is not fully known to the recipient of the data and the data must be accompanied by explicit Representation Information, understandable to the recipient, that is used to interpret the data. An example is a string of bits (the data) accompanied by a description of how to interpret a string of bits as numbers representing temperature observations measured in degrees Celsius (the representation information).

Information Object: A Data Object together with optional Representation Information.

Information Package: An information packaging concept that distinguishes Content Information from associated Preservation Description Information where the Preservation Description Information applies to the Content Information and is needed to aid in the preservation of the Content Information. It has associated Packaging Information used to delimit and identify the Content Information and Preservation Description Information.

Ingest: This entity contains the services and functions that accept Submission Information Packages from Producers, prepares Archival Information Packages for storage, and ensure that Archival Information Packages and their supporting Descriptive Information become established within the OAIS.

Long Term: A period of time which is long enough to be concerned about the impacts of changing technologies, including support for new media and data formats, and with a changing user community, on the information being held in a repository. This period extends into the indefinite future.

Long-term Preservation: The act of preserving information, in a form which can be made understandable to a Designated Community, over the Long Term.

Management: Management is the role played by those who set overall OAIS policy as one

component in a broader policy domain.

Member Description: An Associated Description that describes a member of a collection.

Metadata: Data about other data.

Open Archival Information System (OAIS): An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for one or more Designated Communities. It meets a set of such responsibilities as defined in Section 3.1 of this document and this allows an OAIS archive to be distinguished from other uses of the term 'archive.' The term 'Open' in OAIS is used to imply that this standard and future related standards are developed in open forums, and it does not imply that access to the archive is unrestricted.

Order Agreement: An agreement between the archive and the Consumer in which the physical details of the delivery such as media type and format of Data are specified.

Overview Description: A specialization of the Collection Description that describes the collection as a whole.

Packaging Information: That information that is used to bind and identify the components of an Information Package. For example, it may be the ISO-9660 volume and directory information used on a CD-ROM to provide the content of several files containing Content Information and Preservation Description Information.

Physical Object: An object (such as a moon rock, bio-specimen, microscope slide) with physically observable properties that represent information that is considered suitable for being adequately documented for preservation, distribution and independent usage.

Preserve: Maintain information, in a correct and independently usable form, over the Long Term. Independently usable information has sufficient documentation to allow the information to be understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.

Preservation Description Information (PDI): Information necessary to adequately preserve the Content Information and which can be categorized as Provenance, Reference, Fixity, and Context information.

Producer. The role played by those persons, or client systems, who provide the information to be preserved.

Provenance Information: Information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. Examples of Provenance Information are the principal investigator who recorded the data and the information concerning its storage, handling and migration.

Reference Information: Information that identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Content Information. It also provides these identifiers that allow outside systems to refer, unambiguously, to this particular Content Information. An example of reference information is an ISBN.

Reference Model: A framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that environment. A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist.

Refreshment: A Digital Migration where the effect is to replace a media instance with a copy that is sufficiently exact that all Archival Storage hardware and software continues to run as before.

Repackaging: A Digital Migration in which there is an alteration in the Packaging Information of the AIP.

Replication: A Digital Migration where there is no change to the Packaging Information, the Content Information and the PDI. The bits used to represent these Information Objects are preserved in the transfer to the same or new media instance.

Representation Information: Information that maps Data into more meaningful concepts. An example is the ASCII definition which describes how bits (i.e., Data) are mapped into numbers. Another example is a description of the numbers (i.e., Data) of a table as being the coordinates of a location on the Earth measured in East longitude and latitude.

Representation Network: The set of Representation Information which fully describes the meaning of a Data Object. Representation Information in digital forms needs additional Representation Information so its digital forms can be understood over the Long Term.

Result Set: The set of descriptive records for those AIPs in an OAIS which match the criteria stated in a Consumer query, or other results from a search on Data Management.

Search Session: A session initiated by the Consumer with the archive during which the Consumer will use the archive Finding Aids to identify and investigate potential holdings of interest.

Semantic Information: Information which imparts meaning apart from how other information is organized.

Structure Information: Information that imparts meaning about how other information is organized. For example, it maps bit streams to common computer types such as characters, numbers, and pixels and aggregations of those types such as character strings and arrays.

Submission Agreement: An agreement reached between an OAIS and the Producer that specifies a data model for the Data Submission Session. This data model identifies format/contents and the logical constructs used by the Producer and how they are represented on each media delivery or in the telecommunication session.

Submission Information Package (SIP): The Information Package identified by the Producer in the Submission Agreement with the OAIS

Transformation: A Digital Migration in which there is an alteration to the Content Information or PDI of an Archival Information Package. For example, changing ASCII codes to UNICODE in a text document being preserved is a Transformation.

Unit Description: A type of Package Description that is specialized to provide searchable information about an Archival Information Unit.

2 OAIS CONCEPTS

The purpose of this section is to motivate and describe several key high level OAIS concepts. A more detailed view, and a formal modeling of these concepts, is given in Section 4.

The term “archive” has come to be used to refer to a wide variety of storage and preservation functions and systems. Traditionally, an archive is understood as a facility or organization which preserves records, originally generated by or for a government organization, institution, or corporation, for access by public or private communities. It accomplishes this task by taking ownership of the records, ensuring that they are understandable to the accessing community, and managing them so as to preserve their information content and authenticity. Historically, these records have been in such forms as books, papers, maps, photographs, and film, which can be read directly by humans, or read with the aid of simple optical magnification and scanning aids. The major focus for preserving this information has been to ensure that they are on media with long term stability and that access to this media is carefully controlled.

The explosive growth of information in digital forms has posed a severe challenge not only for traditional archives and their information providers, but for many other organizations in the government, commercial and non-profit sectors. These organizations are finding, or will find, that they need to take on the information preservation functions typically associated with traditional archives because digital information is easily lost or corrupted. The pace of technology evolution is causing some hardware and software systems to become obsolete in a matter of a few years, and these changes can put severe pressure on the ability of the related data structures or formats to continue effective representation of the full information desired. Although some archives may be temporary, some or all of their information may need to be preserved indefinitely. Much of the supporting information necessary to preserve this information is more easily available or only available at the time when the original information is produced. Such temporary archives need to be active participants in the long-term preservation effort and follow the principles espoused in this OAIS reference model to ensure the information can be preserved for the long term. Participation in these efforts will minimize the lifecycle costs and enable effective long-term preservation of the information.

The explosion of computer processing power and digital media has resulted in many systems where the producer role and the archive role are the responsibility of the same entity. These systems, which are sometimes known as Active Archives should ascribe to the goals of long-term preservation discussed in this document. The design process must realize that some of the long term preservation activities may conflict with the goals of rapid production and dissemination of products to consumers. The designers and architects of such systems should document where compromises have been made.

A major purpose of this reference model is to facilitate a much wider understanding of what is required to preserve and access information for the Long Term. To avoid confusion with simple "bit storage" functions, the reference model defines an Open Archival Information System (OAIS) which performs a Long-term information preservation and access function. An OAIS archive is one that intends to preserve information for access and use by one or

more Designated Communities, and it meets the requirements given in Section 3. It includes archives that have to keep up with steady input streams of information as well as those that experience primarily aperiodic inputs. It includes archives that provide a wide variety of sophisticated access services as well as those that support only the simplest types of requests. For the remainder of this document, the term archive is understood to refer to an OAIS, or OAIS archive, unless the context makes it clear otherwise (e.g., traditional archives).

The OAIS model recognizes the already highly distributed nature of digital information holdings and the need for local implementations of effective policies and procedures supporting information preservation. This allows, in principle, a wide variety of organizational arrangements, including various roles for traditional archives, in achieving this preservation. It is expected that organizations attempting to preserve information will find that using OAIS terms and concepts will assist them in achieving their information preservation goals.

2.1 OAIS ENVIRONMENT

The simple model shown in Figure 2-1 gives the environment surrounding an OAIS

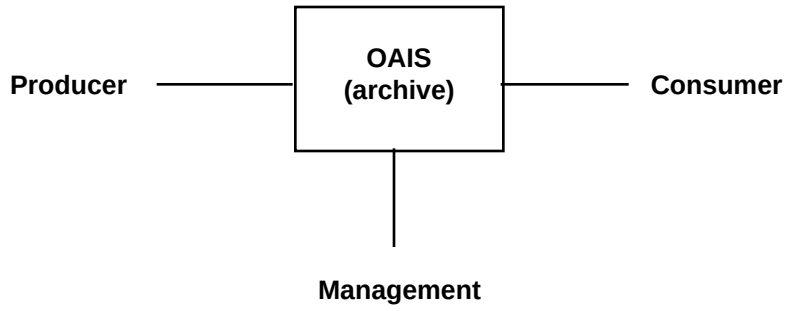


Figure 2-1. Environment Model of an OAIS

Outside the OAIS are **Producers**, **Consumers**, and **Management**.

- Producer is the role played by those persons, or client systems, which provide the information to be preserved.
- Management is the role played by those who set overall OAIS policy as one component in a broader policy domain. In other words, Management control of the OAIS is only one of Management's responsibilities. Management is not involved in day-to-day archive operations. The responsibility of managing the OAIS on a day-to-day basis is included within the OAIS in an administrative functional entity which will be described in Section 4.1.
- Consumer is the role played by those persons, or client systems, which interact with OAIS services to find and acquire, preserved information of interest. A special class of Consumers is the **Designated Community.** The Designated Community is the set of Consumers who should be able to understand the preserved information.

2.1.1 INTERACTIONS BETWEEN OAIS ARCHIVES

Other OAIS archives are not shown explicitly. Such archives may establish particular agreements among themselves consistent with Management and OAIS needs. Other archives may interact with a particular archive for a variety of reasons and with varying degrees of formalism for any pre-arranged agreements. One OAIS may take the role of Producer to another OAIS; an example is when the responsibility for preserving a type of information is to be moved to this other archive. One OAIS may take the role of Consumer to another OAIS; an example is when the first OAIS decides to rely on the other OAIS for a type of information it seldom needs and chooses not to preserve locally. Such reliance should have some formal basis that includes the requirement for communication between the archives of any policy changes that might affect this reliance. The range of possible interactions between OAIS archive is discussed in Section 6 .

2.2 OAIS INFORMATION

2.2.1 INFORMATION DEFINITION

A clear definition of information is central to the ability of an OAIS to preserve it. While formal modeling of information is given in Section 4, some key concepts are given in this section.

A person, or system, can be said to have a **Knowledge Base**, which allows them to understand received information. For example, a person who has a Knowledge Base that includes an understanding of English will be able to read, and understand, an English text.

Information is defined as any type of knowledge that can be exchanged, and this information is always expressed (i.e., represented) by some type of data. For example, the information in a hardcopy book is typically expressed by the observable characters (the data) which, when they are combined with a knowledge of the language used (the Knowledge Base), are converted to more meaningful information. If the recipient does not already include English in its Knowledge Base, then the English text (the data) need to be accompanied with an English dictionary and grammar information (i.e., **Representation Information**) in a form that is understandable using the recipient's Knowledge Base.

Similarly, the information stored within a CD-ROM file is expressed by the bits (the data) it contains which, when they are combined with the Representation Information for those bits, are converted to more meaningful information as long as the Representation Information is understandable using the recipients Knowledge Base. For example, assume the bits represent an ASCII table of numbers giving the coordinates of a location on the Earth measured in degrees latitude and East longitude. The Representation Information will typically include the definition of ASCII together with descriptions of the format of the numbers and their locations in the file, their definitions as latitude and longitude, and the definition of their units as degrees. It may go on to include additional meaning that is assigned to the table. In general, it can be said that "Data interpreted using its Representation Information yields

Information" and this is shown schematically in Figure 2-2.

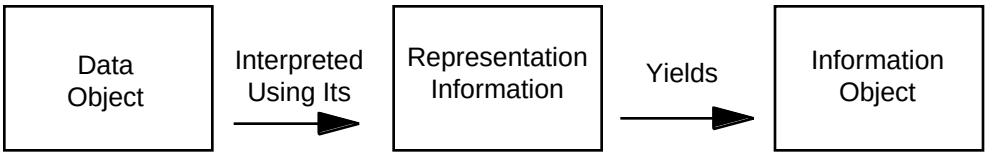


Figure 2-2. Obtaining Information from Data

In order for this **Information Object** to be successfully preserved, it is critical for an OAIS to clearly identify and understand the **Data Object** and its associated Representation Information. For digital information, this means the OAIS must clearly identify the bits and the Representation Information that applies to those bits. This required transparency to the bit level is a distinguishing feature of digital information preservation and it runs counter to information hiding approaches. This presents a significant challenge to the preservation of digital information.

As a further complication, the recursive nature of Representation Information, which typically is composed of its own data and other Representation Information, typically leads to a network of Representation Information objects. Since a key purpose of an OAIS is to preserve information for a Designated Community, the OAIS must understand the Knowledge Base of its Designated Community to understand the minimum Representation Information that must be maintained. The OAIS should then make a decision between maintaining the minimum representation data needed for its Designated Community and a larger amount of representation data which may allow understanding by a larger Consumer community with a less specialized Knowledge Base. Over time, evolution of the Designated Communities' Knowledge Base may require updates to the Representation Information to ensure continued understanding.

As a practical matter, software is used to access the Information Object and it will incorporate some understanding of the network of Representation Information objects involved. However this software should not be used as an excuse to avoid identifying and gathering the Representation Information that defines the Information Object because it is harder to preserve working software than to preserve information in digital or hardcopy forms.

2.2.2 INFORMATION PACKAGE DEFINITION

The definition of an Information Object is applicable to all the information types discussed in this and the following sections. In other words, they all have associated Representation Information although this is usually not shown explicitly.

Every submission of information to an OAIS by a Producer, and every dissemination of information to a Consumer, occurs as one or more discrete transmissions. Therefore it is convenient to define the concept of an Information Package (IP).

An **Information Package (IP)** is a conceptual container of two types of information called **Content Information** and **Preservation Description Information (PDI).** The Content Information and PDI are viewed as being encapsulated and identifiable by the **Packaging Information**. The resulting package is viewed as being discoverable by virtue of the **Descriptive Information**.

These Information Package relationships are shown schematically in Figure 2-3.

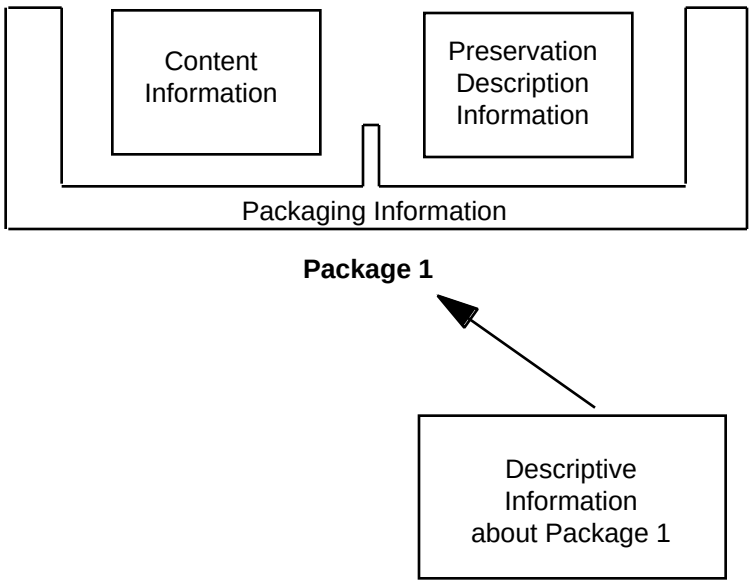


Figure 2-3. Information Package Concepts and Relationships

The Content Information is that information which is the primary target of preservation. It consists of the primary Digital Object (i.e., bits) and its associated Representation Information needed to make the Digital Object understandable to the Designated Community. For example, it may be an image is provided as the bit content of a CD-ROM file together with other files, on the same CD-ROM, that contain Representation Information.

Only after the Content Information has been clearly defined can an assessment of the Preservation Description Information be made. The Preservation Description Information applies to the Content Information and is needed to preserve the Content Information, to ensure it is clearly identified, and to understand the environment in which the Content Information was created. The Preservation Description Information is divided into four types of preserving information called Provenance, Context, Reference, and Fixity. Briefly, they are the following:

- Provenance describes the source of the Content Information, who has had custody of it since its origination, and what its history (including processing history) has been.
- Context describes how the Content Information relates to other information outside

the Information Package. For example, it would describe why the Content Information was produced and it may include a description of how it relates to another Content Information object that is available.

- Reference provides one or more identifiers, or systems of identifiers, by which the Content Information may be uniquely identified. Examples include an ISBN number for a book or a set of attributes that distinguish one Content Information from another.
- Fixity provides a wrapper, or protective shield, that protects the Content Information from undocumented alteration. For example, it may involve a check sum over the Content Information of a digital Information Package.

The Packaging Information is that information which, either actually or logically, binds and identifies relates the Content Information and PDI. For example, if the Content Information and PDI are identified as being the content of specific files on a CD-ROM, then the Packaging Information would include the ISO-9660 volume/file structure on the CD-ROM as well as thenames and directory information of the files on CD-ROM disk.

The Descriptive Information is that information which is used to discover which package has the Content Information of interest. Depending on the setting, this may be no more than a descriptive title of the Information Package that appears in some message, or it may be a full set of attributes that are searchable in a catalog service.

2.2.3 INFORMATION PACKAGE VARIANTS

It is necessary to distinguish between an Information Package that is preserved by an OAIS and the Information Packages that are submitted to, and disseminated from, an OAIS. These variant packages are needed to reflect the reality that some submissions to an OAIS will have insufficient Representation Information or PDI to meet final OAIS preservation requirements. In addition, these may be organized very differently from the way the OAIS organizes the information it is preserving. Finally, the OAIS may provide information to Consumers that does not include all the Representation Information or all the PDI with the associated Content Information being disseminated. These variants are referred to as the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP).

The **Submission Information Package (SIP)** is that package that is sent to an OAIS by a Producer. Its form and detailed content is typically negotiated between the Producer and the OAIS. Most SIPs will have some Content Information and some PDI, but it may require several SIPs to provide a complete set of Content Information and associated PDI to form an AIP. A single SIP may contain information that is to be included in several AIPs. The Packaging Information will always be present in some form.

Within the OAIS one or more SIPs is transformed into one or more **Archival Information Packages (AIP)** for preservation. The AIP has a complete set of PDI for the associated Content Information . The AIP may also contain a collection of other AIPs and this is discussed and modeled in Section 4. The Packaging Information of the AIP will conform to

OAIS internal standards, and it may vary as it is managed by the OAIS.

In response to a request, the OAIS provides all or a part of an AIP to a Consumer in the form of a **Dissemination Information Package (DIP)**. The DIP may also include collections of AIPs, and it may or may not have complete PDI. The Packaging Information will always be present in some form so that the Consumer can clearly distinguish the information requested. The Packaging Information may take several forms depending on the dissemination media and Consumer requirements.

2.3 OAIS HIGH LEVEL EXTERNAL INTERACTIONS

The following sections present a high level view of the interaction between the entities identified in the OAIS environment. Figure 2-5 is a data flow diagram that represents the operational OAIS archive external data flows. This diagram concentrates on the flow of information among Producers, Consumers and the OAIS and does not include flows that involve Management. These flows are dealt with further in Section 4.

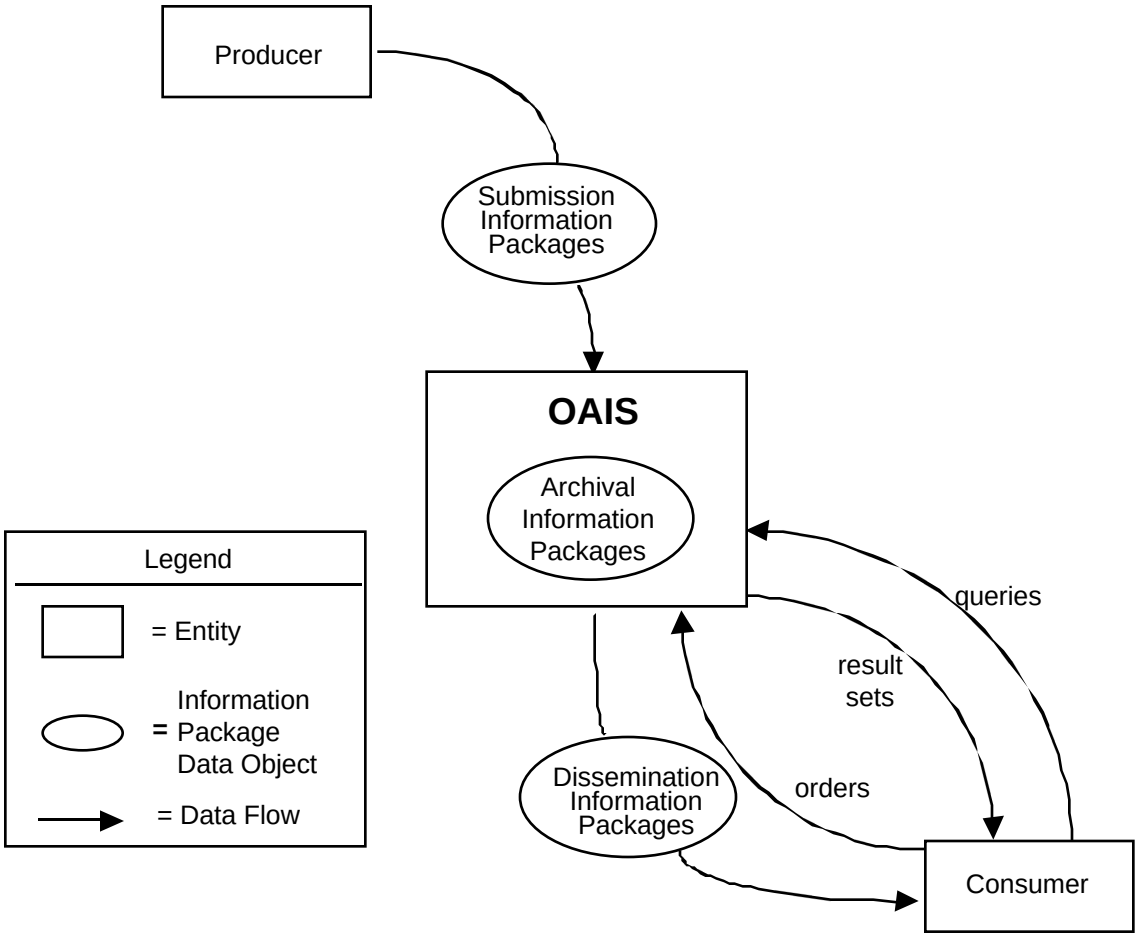


Figure 2-5. OAIS Archive External Data Flows

2.3.1 MANAGEMENT INTERACTION

Management provides the OAIS with its charter and scope. The charter may be developed by the archive but it is important that Management formally endorse archive activities. The scope determines the breadth of both the Producer and Consumer groups served by the archive.

Some examples of typical interactions between the OAIS and Management include:

- Management is often the primary source of funding for an OAIS and may provide guidelines for resource utilization (personnel, equipment, facilities).
- Management will generally conduct some regular review process to evaluate OAIS performance and progress toward long-term goals.
- Management determines or at least endorses pricing policies, as applicable, for OAIS services.
- Management participates in conflict resolution involving Producers, Consumers and OAIS internal administration.

Effective Management should also provide support for the OAIS by establishing procedures that assure OAIS utilization within its sphere of influence. For example, management policies should require that all funded activities within its sphere of influence submit data products to the archive and also adhere to archive standards and procedures.

2.3.2 PRODUCER INTERACTION

The first contact between the OAIS and the Producer is a request that the OAIS preserve the data products created by the Producer. This contact may be initiated by the OAIS, the Producer or Management. The Producer establishes a **Submission Agreement** with the OAIS, which identifies the SIPs to be submitted and may span any length of time for this submission. Some Submission Agreements will reflect a mandatory requirement to provide information to the OAIS, while others will reflect a voluntary offering of information. Even in the case where no formal Submission Agreement such as a WWW site, a virtual Submission Agreement may exist specifying the file formats and the general subject matter the site will accept.

Within the Submission Agreement, one or more **Data Submission Sessions**, are specified. There may be significant time gaps between the **Data Submission Sessions**. A Data Submission Session will contain one or more SIPs and may be a delivered set of media or a single telecommunications session. The Data Submission Session content is based on a data model negotiated between the OAIS and the Producer in the Submission Agreement. This data model identifies the logical components of the SIP (e.g., the Content Information, PDI, Packaging Information, and Descriptive Information) that are to be provided and how (and whether) they are represented in each Data Submission Session. All data deliveries within a

Submission Agreement are recognized as belonging to that Submission Agreement and will generally have a consistent data model which is specified in the Submission Agreement. For example, a Data Submission Session consists of a set of Content Information corresponding to a set of observations which are carried by a set of files on a CD-ROM. The Preservation Description Information is split among two other files. All of these files need Representation Information which must be provided in some way. The CD-ROM and its directory/file structure are the Packaging Information, which provides encapsulation and identification of the Content Information and PDI in the Data Submission Session.. The Submission Agreement indicates how the Representation Information for each file is to be provided, how the CD-ROM is to be recognized, how the Packaging Information will be used to identify and encapsulate the SIP Content Information and PDI, and how frequently Data Submission Sessions (e.g. one per month for two years) will arrive. It also gives other needed information such as access restrictions to the data.

Each SIP in a Data Submission Session is expected to meet minimum OAIS requirements for completeness. However, in some cases multiple SIPs may need to be received before an acceptable AIP can be formed and fully ingested within the OAIS. In other cases, a single SIP may contain data to be included many AIPs.. A Submission Agreement also includes, or references, the procedures and protocols by which an OAIS will either verify the arrival and completeness of a Data Submission Session with the Producer or question the Producer on the contents of the Data Submission Session.

2.3.3 CONSUMER INTERACTION

There are many types of interactions between the Consumer and the OAIS. These interactions include questions to a help desk, requests for literature, catalog searches, orders and order status requests. The ordering process is of special interest to the OAIS RM since it deals with the flow of archive holdings between the OAIS and the Consumer.

The Consumer establishes an **Order Agreement** for the OAIS for information expected to be received on the basis of some triggering event. This event may be periodic such as a monthly distribution of any AIPs ingested by the OAIS from a specific Producer, it may be a unique event such as the ingestion of a specific AIP or it may simply be the receipt of the Order Agreement. The Order Agreement may span any length of time and under it one or more **Data Dissemination Sessions** may take place. A Data Dissemination Session may involve the transfer of a set of media or a single telecommunications session. The Order Agreement identifies one or more AIPs of interest, how those AIPs are to be transformed and mapped into Dissemination Information Package (DIPs) and how those DIPs will be packaged in a Data Dissemination Session. The Order Agreement will also specify other needed information such as delivery information (e.g., name or mailing address), and any pricing agreements as applicable. There are two common order types initiated by Consumers, the **Event Based Order** and the **Adhoc Order**.

In the case of an Event Based Order, the Consumer establishes an Order Agreement with the OAIS for information expected to be received on the basis of some triggering event. This event may be periodic, such as a monthly distribution of any AIPs ingested by the OAIS

from a specific Producer, or it may be a unique event such as the ingestion of a specific AIP. The Order Agreement will also specify other needed information such as the trigger event for new Data Dissemination Sessions, the criteria for selecting the OAIS holdings to be included in each new Data Dissemination Session.

In the case of an Adhoc Order: The Consumer establishes an Order Agreement with the OAIS for information available from the archive. If the Consumer does not know a priori what specific holdings of the OAIS are of interest, the Consumer will establish a **Search Session** with the OAIS. During this Search Session the Consumer will use the OAIS **Finding Aids** that operate on **Descriptive Information,** or in some cases on the AIPs themselves, to identify and investigate potential holdings of interest. This may be accomplished by the submission of queries and the return of result sets to the Consumer. This searching process tends to be iterative with a Consumer first identifying broad criteria and then refining these criteria based on previous search results. Once the Consumer identifies the OAIS AIPs of interest he may provide an Order Agreement that documents the identifiers of the AIPs he wishes to acquire and how he will acquire them to the OAIS . If the AIPs are available, an Adhoc Order will be placed. However if the AIPs desired are not yet available, a Subscription Order may be placed.

The Order Agreement does not have to be a formal document. In general an OAIS will have a general pricing policy and maintain an information base of the electronic and physical mailing addresses of its users. In this case, the process of developing an Order Agreement may be no more than the completion of a World Wide Web form to specify the AIPs of interest.

3 OAIS RESPONSIBILITIES

Section 3.1 identifies the responsibilities that must be discharged by an OAIS. Section 3.2 expands on these responsibilities by giving detailed examples, although not all of these will be applicable to all OAISs.

3.1 MANDATORY RESPONSIBILITIES

This section establishes mandatory responsibilities that an organization must discharge in order to operate an OAIS archive. The OAIS must:

- Negotiate and accept appropriate information from information producers.

- Obtain sufficient control of the information provided to the level needed to ensure Long-term preservation.

- Determine (dependently or independently) which communities need to be able to understand the information provided.

- Ensure the information to be preserved is independently understandable to the Designated Communities. In other words, the communities should be able to understand the information without needing the assistance of the experts who produced the information.

- Follow documented policies and procedures which ensures the information is preserved against all reasonable contingencies and enables the information to be disseminated as authenticated copies of the original or as traceable to the original.

- Make the preserved information available to the Designated Communities..

In the following sections, each of these responsibilities is explored in greater depth for clarification.

3.2 DETAILED DISCUSSIONS OF RESPONSIBLITIES

This section expands on the responsibilities of Section 3.1 to provide clarification. Not all of these detailed responsibilities will be applicable to all OAISs.

3.2.1 NEGOTIATES AND ACCEPTS INFORMATION

An organization operating an OAIS will have established some criteria that aids in determining the types of information that it is willing to, or it is required to, accept. These criteria may include, among others, subject matter, information source, degree of uniqueness or originality, and the nature of the techniques used to represent the information (e.g., physical media, digital media, format). The information may, in general, be submitted using

a wide variety of common and not-so-common forms, such as books, documents, maps, data sets, and moon rocks using a variety of communication paths including networks, mail, and special delivery.

The OAIS negotiates with the Producer to ensure it acquires appropriate Content Information and associated PDI for its mission and the Designated Communities. It needs to extract, or otherwise obtain, sufficient Descriptive Information to assist its Designated Communities in finding the Content Information of interest. It also needs to ensure that the information meets all OAIS internal standards.

3.2.2 OBTAINS SUFFICIENT CONTROL FOR PRESERVATION

In general, the OAIS will accept the Content Information as either a custodian or as the new legal owner. When acting as a custodian, the OAIS may need to involve the actual owner(s) in some migration and access decisions depending on the authority it has been granted to act independently. When it is the legal owner, it already has the independence to do what is needed to preserve the information and make it accessible.

The OAIS must assume sufficient control over the Content Information and Preservation Description Information so that it is able to preserve it for the Long Term. There is no issue with the AIP's Packaging Information because, by definition, this is under internal OAIS control. The problems of assuming sufficient control of the Content Information and Preservation Description Information, when they are largely digital, are addressed in three related categories as follows:

- Copyright implications
- Authority to modify Representation Information
- Agreements with external organizations

Copyright implications: These issues occur when the OAIS acts as a custodian.. An OAIS will need to understand the copyright concepts and applicable laws prior to accepting copyright materials into the OAIS It can establish guidelines for ingestion of information and rules for dissemination and duplication of the information when necessary. It is beyond the scope of this document to provide details of national and international copyright laws.

Authority to modify Representation Information: Although the Fixity information within the Preservation Description Information of an AIP is ensuring that the Content Information related bits are not being altered, there will come a time when Content Information bits are not in a form that is convenient for the designated Consumer community. The Content Information bits may be fully documented in available hardcopy forms, so technically the information has not been lost, but practically the information has become inaccessible. The OAIS needs the authority to migrate the Content Information to new representation forms. If it is acting as a custodian, it may need to seek additional permission to make such changes. If the information is copyrighted, hopefully the OAIS already has already negotiated permission to make the changes needed to meet preservation objectives. It may need to

employ subject matter experts, from outside the OAIS, to help ensure that information is not lost. Ideally, when this situation arises, both the original AIPs (fully described) and new AIPs will be retained. Digital Migration issues are addressed more fully in Section 5.

Agreements with external organizations: An OAIS may establish a variety of agreements with other organizations to assist in its preservation objectives. For example, it may establish an agreement with another OAIS so that it does not have to preserve all the common Representation Information objects related to its Content Information objects. . Agreements with other organizations need to be monitored to be sure they are being followed and remain useful.

3.2.3 DETERMINES DESIGNATED CONSUMER COMMUNITIES

The submission, or planned submission, of Content Information and associated PDI requires a determination as to who the expected consumers, or Designated Community, of this information will be. This is necessary in order to determine if the information, as represented, will be understandable to that community. For example, an archive may decide that certain Content Information should be understandable to the general public and therefore this becomes the Designated Community.

For some scientific information, the Designated Community of consumers might be described as those with a first year graduate level education in a related scientific discipline. This is a more difficult case as it is less clear what degree of specialized scientific terminology might actually be acceptable. The producers of such specialized information are often familiar with a narrowly recognized set of terminology, so it is especially critical to clearly define the Designated Community for their information and to make the effort to ensure this community can understand the information.

The possible evolution of the definition of the Designated Community also needs consideration. Information originally intended for a narrowly defined community may need to be made more widely understandable at some future date. For example, information originally intended to be understandable to a particular scientific community may need to be made understandable to the general public. This is likely to mean adding explanations in support of the Representation Information and the Preservation Description Information and it can become increasingly difficult to obtain this information over time. Selecting a broader definition of the Designated Community (e.g. general public) when the information is first proposed for Long Term Preservation can reduce this concern and also improve the likelihood that the information will be understandable to all in the original community.

3.2.4 ENSURES INFORMATION IS INDEPENDENTLY UNDERSTANDABLE

The degree to which Content Information and its associated PDI conveys information to a Designated Community is, in general, quite subjective. Nevertheless, it is essential that an archive make this determination in order to maximize information preservation. Digital Content Information and PDI needs adequate Representation Information to be understandable to its Designated Community. Typically there are multiple Representation

Information objects involved and this is dealt with in section 4.2

For example, consider Content Information from a digital set of observations of rainfall, temperature, pressure, wind velocities, and other parameters measured all over the world for a year. This type of information is very extensive, is not usually in a form intended for direct human browsing or reading, but it is in a form appropriate to searching and manipulation by application software. Such content may only be understandable to the original producers unless there is adequate documentation of the meaning of the various fields and their inter-relationships, and how the values relate back to the original instrumentation that made the observations. In such specialized fields extra effort is needed to ensure that the Content Information and the Preservation Description Information are understandable to a Designated Community. If the archive does not have this level of expertise in-house, it may need to have outside community representatives review the information for long-term understandability. Otherwise some of the information may be understandable to only a few specialists and be lost when they are no longer available.

Even when a set of information has been determined to be understandable to a particular Designated Community, over time the Knowledge Base of this community may evolve to the point that important aspects of the information may no longer be readily understandable. At this point it may be necessary for the OAIS to enhance the associated Representation Information so that it is again readily understandable to the Designated Community.

As another example, a manuscript's Content Information may be written in English and therefore its content may be generally understandable to a wide audience. However, unless the purpose for which it was created is clearly documented, much of its meaning may be lost. This 'purpose' information is part of its Context and must be provided in the Preservation Description Information.

Digital Content Information needs software for efficient access. However, maintaining Content Information-specific software over the long term has not yet been proven cost effective due to the narrow application of such software. The danger of information loss is great when such software is relied upon for information preservation and understanding because it may cease to function under only small changes to the hardware and software environment. This may not be recognized unless there is a vigorous, ongoing, testing program.

3.2.5 FOLLOWS ESTABLISHED PRESERVATION POLICIES AND PROCEDURES

It is essential for an OAIS to have documented policies and procedures for preserving its AIPs, and it needs to follow those procedures. The appropriate policies and procedures will depend, at minimum, on the nature of the AIPs and any 'backup' relationships the archive may have with other archives. For example, migrations which alter any Content Information or PDI will need to be carefully monitored and the appropriate PDI fully updated. This attention to detail, while also ensuring against processing errors, requires that strong policies and procedures be in place and that they be executed.

The Designated Communities need to be monitored to be sure the Content Information is still understandable to them. The Designated Communities may loose their familiarity with some terminology, and their definition may be broadened to include other members with different backgrounds. For example, a periodic review with participants representing the Designated Communities could assist in this process.

A long-term technology usage plan, updated as technology evolves, is essential to avoid being caught with very costly system maintenance, emergency system replacements, and costly data representation transformations.

3.2.6 MAKES THE INFORMATION AVAILABLE

By definition, an OAIS makes its AIPs visible and available to its Designated Communities. Multiple views of its holdings, supported by various search aids that may cut across collections of AIPs, may be provided. Some AIPs may only exist as the output of algorithms operating on other AIPs. They appear as DIPs that, upon dissemination, need to include documentation on how they were derived from other AIPs. The expectations of OAIS Consumers regarding access services will vary widely among archives and over time as technology evolves. Pressures for ever more effective access must be balanced with the requirements for preservation under the available resource constraints.

Some AIPs may have restricted access and therefore may only be disseminated to Consumers who meet access restrictions. The OAIS needs to have published policies on access and restrictions so that the rights of all parties are protected.

In general, DIPs may be distributed by all varieties of communication paths, including networks and physical media.

4 DETAILED MODELS

The purpose of this section is to provide a more detailed model view of the functional entities of the OAIS and the information handled by the OAIS. This aids OAIS designers of future systems and provides a more precise set of terms and concepts for discussion of current systems.

4.1 FUNCTIONAL MODEL

The OAIS of Figure 2-1 is separated in Figure 4-1 into five functional entities and related interfaces. Only major information flows are shown. The lines connecting entities identify communication paths over which information flows in both directions. The lines to Administration are dashed only to reduce diagram clutter.

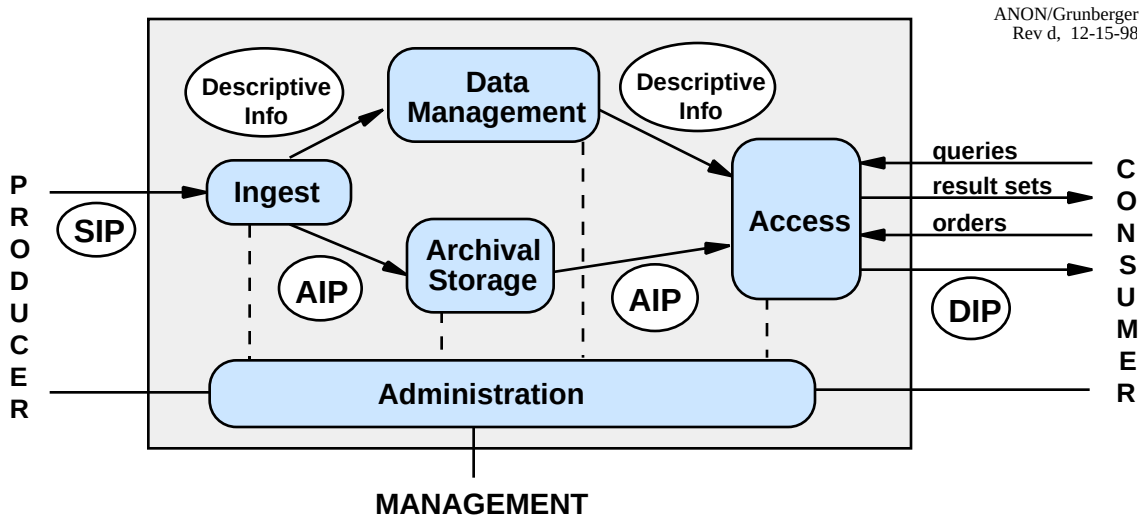


Figure 4-1. OAIS Functional Entities

The role provided by each of the entities in Figure 4-1 is described briefly as follows:

Ingest: This entity provides the services and functions to accept Submission Information Packages (SIPs) from Producers and prepare the contents for storage and management within the archive. Ingest functions include receiving SIPs, performing quality assurance on SIPs, generating an Archival Information Package (AIP) which complies with the archive's data formatting and documentation standards, extracting Descriptive Information from the AIPs for inclusion in the archive database, and coordinating updates to Archival Storage and Data Management.

Archival Storage: This entity provides the services and functions for the storage, maintenance and retrieval of AIPs. Archival Storage functions include receiving AIPs from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the

media on which archive holdings are stored, performing routine and special error checking, providing disaster recovery capabilities, and providing AIPs to Access to fulfill orders.

Data Management: This entity provides the services and functions for populating, maintaining, and accessing both Descriptive Information which identifies and documents archive holdings and administrative data used to manage the archive. Data Management functions include administering the archive database functions (maintaining schema and view definitions, and referential integrity), performing database updates (loading new descriptive information or archive administrative data), performing queries on the data management data to generate result sets, and producing reports from these result sets.

Administration: This entity manages the overall operation of the archive system. Administration functions include soliciting and negotiating submission agreements with Producers, auditing submissions to ensure that they meet archive standards, maintaining configuration management of system hardware and software, evaluating the contents of the archive and periodically requesting archival information updates, providing system engineering functions to monitor and improve archive operations, developing and maintaining archive standards and policies, providing customer support, monitoring changes in the Designated Communities, interacting with Management, and activating stored requests.

Access: This entity supports Consumers in determining the existence, description, location and availability of information stored in the OAIS and allowing Consumers to request and receive information products. Access functions include communicating with Consumers to receive requests, preparing finding aids to support user access to the archive collection, applying controls to limit access to specially protected information, coordinating the execution of requests to successful completion, generating responses (Dissemination Information Packages, result sets, reports) and delivering the responses to Consumers.

In addition to the entities described above, there are various **Common Services** assumed to be available. These services are considered to constitute another functional entity in this model.

4.1.1 DETAILED DESCRIPTION OF FUNCTIONAL ENTITIES

In the following subsections, specific flows of information among the entities are identified **in bold type** the first time they appear in the text. All identifiers include a suffix in braces. For example, "**confirmation of receipt {2b}"** is a flow identified in Section 4.1.1.2. The **"2"** in the suffix matches the last digit of the section number, and the **"b"** is an alphabetical index of the sequence of appearance in the text. These suffixes will be used in the data flow diagrams of Section 4.1.2, Figures 4-7 and 4-8, to provide references to the text.

The detailed functional descriptions of the following subsections are accompanied by diagrams, Figures 4-2 through 4-6, which depict only the major data flows within and among the entities. Omitted for clarity are minor flows such as acknowledgment notices, and also data flow suffixes.

4.1.1.1 Common Services

Modern, distributed computing applications assume a number of supporting services such as inter-process communication, name services, temporary storage allocation, exception handling, security, backup and directory services. Much excellent work has already been done in the area of open system environment reference models. The services described in this section are based on the services described in the IEEE POSIX OSE Reference Model (ref6), (IEEE 1003.0, 1995), the NIST Application Portability Profile (APP), (ref 7) NIST Special Publication 500-xxx, April 1995)

Operating system services provide the core services needed to operate and administer the application platform and provide an interface between application software and the platform. These services include the following:

- Kernel operations provide low-level services necessary to create and manage processes, execute programs, define and communicate signals, define and process system clock operations, manage files and directories, and control input-output processing to and from the external environment.
- Commands and utilities include mechanisms for operations at the operator level, such as comparing, printing, and displaying file contents; editing files; pattern searching; evaluating expressions; logging messages; moving files between directories; sorting data; executing command scripts; and accessing environment information.
- Realtime extension includes the application and operating system interfaces needed to support those application domains requiring deterministic execution, processing, and responsiveness. The extension defines the applications interface to basic system services for input/output, file system access, and process management.
- System management includes capabilities to define and manage user resource allocation and access (i.e., what resources are managed and the classes of access defined), configuration and performance management of devices, file systems, administrative processes (job accounting), queues, machine/platform profiles, authorization of resource usage, and system backup.
- Operating system security services specify the control of access to system data, functions, hardware, and software resources by users and user processes.

Network services provide the capabilities and mechanisms to support distributed applications requiring data access and applications interoperability in heterogeneous, networked environments. These services include the following:

- Data communication includes API and protocol specifications for reliable, transparent, end-to-end data transmission across communications networks.
- Transparent file access to available files located anywhere in a heterogeneous network.
- Personal/micro computer support for interoperability with systems based on other operating systems, particularly microcomputer operating systems, that may not be formally specified in a national or international standard.
- Remote Procedure Call services include specifications for extending the local procedure call to a distributed environment.

- Network security services include access, authentication, confidentiality, integrity, and non-repudiation controls and management of communications between senders and receivers of information in a network

Security services capabilities and mechanisms to protect sensitive information and treatments in the information system. The appropriate level of protection is determined based upon the value of the information to the application end-users and the perception of threads to it. These services include the following:

- Identification/authentication service confirms the identities of requesters for use of information system resources. In addition, authentication can apply to providers of data. The authentication service may occur at the initiation of a session or during a session.
- Access control service prevents the unauthorised use of information system resources. This service also prevents the use of a resource in an unauthorised way. This service may be applied to various aspects of access to a resource (e.g., access to communications to the resource, the reading, writing, or deletion of an information/data resource, the execution of a processing resource) or to all accesses to a resource.
- Data integrity service ensures that data is not altered or destroyed in an unauthorised manner. This service applies to data in permanent data stores and to data in communications messages.
- Data confidentiality service ensures that data is not made available or disclosed to unauthorised individuals or computer processes. This service will be applied to devices that permit human interaction with the information system. In addition, this service will ensure that observation of usage patterns of communications resources will not be possible.
- Non-repudiation service ensures that entities engaging in an information exchange cannot deny being involved in it. This service may take one or both of two forms. First, the recipient of data is provided with proof of the origin of the data. This protects against any attempt by the sender to falsely deny sending the data or its contents. Second, the sender of data is provided with proof of delivery of data. This protects against any subsequent attempt by the recipient to falsely deny receiving the data or its contents.

4.1.1.2 Ingest

The functions of the Ingest entity are illustrated in Figure 4-2 and detailed in the text that follows.

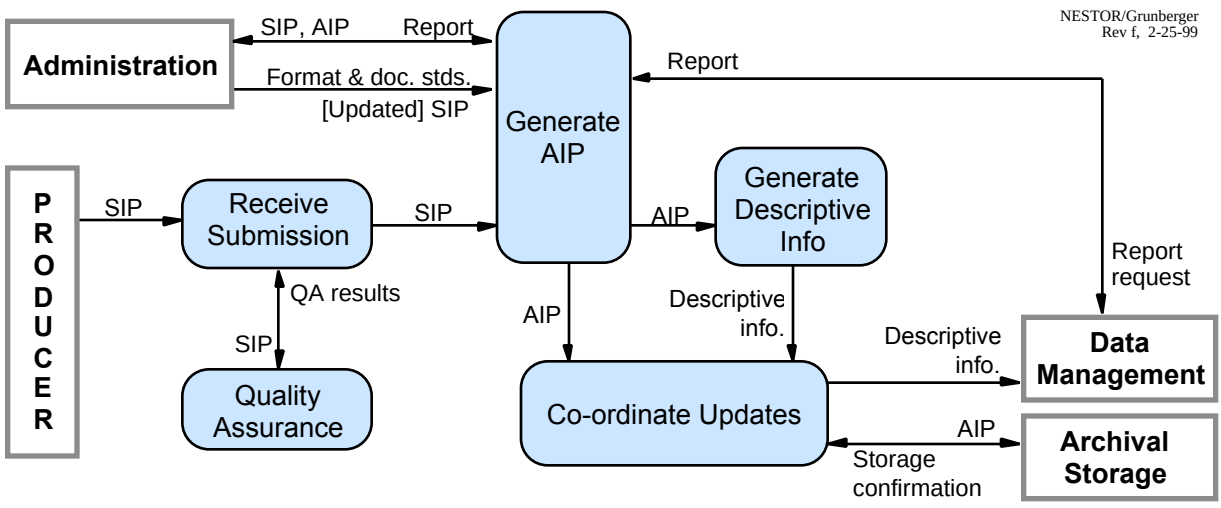


Figure 4-2. Functions of Ingest

The **Receive Submission** function provides the appropriate storage capability or devices to receive a **SIP {2a}** from the Producer. Digital SIPs may be delivered via electronic transfer (e.g. FTP), loaded from media submitted to the archive, or simply mounted (e.g. CD-ROM) on the archive file system for access. Non-digital SIPs would likely be delivered by conventional shipping procedures. The Receive Submission function may represent a legal transfer of custody for the Content Information in the SIP, and may require that special access controls be placed on the contents. This function provides a **confirmation of receipt {2b}** of a SIP to the Producer, which may include a request to **resubmit a SIP {2c}** in the case of errors resulting from the SIP submission.

The **Quality Assurance** function validates the successful transfer of the SIP to the staging area. For digital submissions, these mechanisms might include cyclic redundancy codes (CRCs) or checksums associated with each data file, or the use of system log files to record and identify any file transfer or media read/write errors.

The **Generate AIP** function transforms one or more SIPs into one or more AIPs that conforms to the archive's **data formatting and documentation standards {2d}**. This may involve file format conversions, data representation conversions or reorganization of the content information in the SIPs. The Generate AIP function may issue **report requests {2e}** to Data Management to obtain **reports {2f}** of information needed by the Generate AIP function to produce the Descriptive Information that completes the AIP. This function sends **SIPs or AIPs for audit {2g}** to the Audit Submission function in Administration, and receives back an **audit report {2h}**.

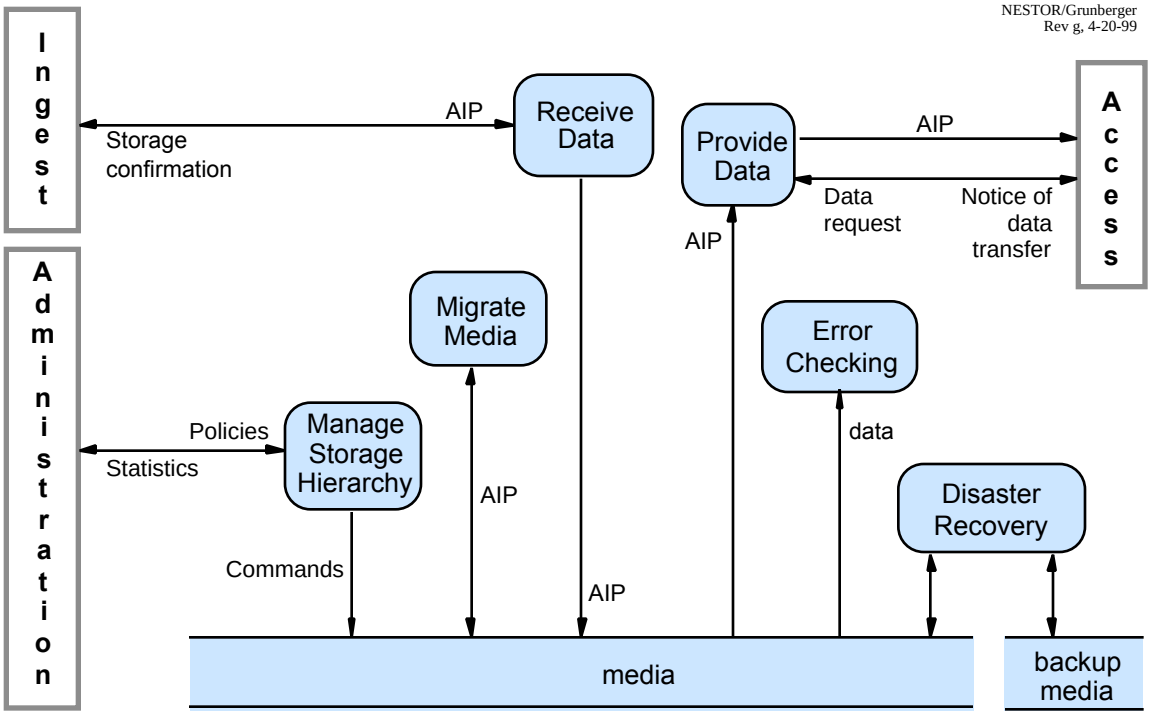
The **Generate Descriptive Information** function extracts Descriptive Information from the AIPs and collects Descriptive Information from other sources to provide to Data Management. This includes metadata to support searching and retrieving AIPs (e.g. who,

what, when, where, why) and could also include special browse products (thumbnails, images) to be used as finding aids.

The **Coordinate Updates** function is responsible for transferring the AIPs to Archival Storage and the Descriptive Information to Data Management. Transfer of the **AIP {2i}** includes a **storage request {2j}** and may represent an electronic, physical, or a virtual (i.e. data stays in place) transfer. After the transfer is completed and verified, Archival Storage returns a **storage confirmation {2h}** indicating the storage identification information for the AIP. The Coordinate Updates function also incorporates the storage identification information into the **Descriptive Information {2m}** for the AIP and transfers it to the Data Management entity along with a **database update request {2n}**. In return, Data Management provides a **database update response {2p}** indicating the status of the update.

4.1.1.3 Archival Storage

The functions of the Archival Storage entity are illustrated in Figure 4-3 and detailed in the text that follows.



SUBSCRIBER @EditionMgr @EditionClient @09720130 \a

Figure 4-3. Functions of Archival Storage

The **Receive Data** function receives a storage request {2g} and an AIP {2f} from Ingest and moves the data to permanent storage within the archive. The transfer request may need to indicate the anticipated frequency of utilization of the data objects comprising the AIP to

allow the appropriate storage devices or media to be selected for storing the AIP. This function will select the media type, prepare the devices or volumes and perform the physical transfer to the Archival Storage volumes. On completion of the transfer this function sends a storage confirmation {2h} message to Ingest including the storage identification of the AIPs.

The **Manage Storage Hierarchy** function positions the contents of the AIPs on the appropriate media based on storage **management policies {3a}**, operational statistics, or directions from Ingest via the storage request {2g}. It will also conform to any special levels of service required for the AIP or any special security measures that are required and ensure the appropriate level of protection for the AIP. These include on-line, off-line or near-line storage, required throughput rate, maximum allowed bit error rate, or special handling or backup procedures. This function also provides **operational statistics {3b}** to Administration summarizing the inventory of media on-hand, available storage capacity in the various tiers of the storage hierarchy, and usage statistics.

The **Migrate Media** function provides the capability to reproduce the AIPs over time. Within the Migrate Media function the Content Information and Preservation Description Information (PDI) must not be altered. However, the data constituting the Packaging Information may be changed as long as it continues to perform the same function. The migration strategy must select a storage medium, taking into consideration the expected and actual rates of errors encountered in various media types, their performance, and their costs of ownership. If media-dependent attributes (e.g. tape block sizes, CD-ROM volume information) have been included as part of the Content Information, a way must be found to preserve this information when migrating to higher capacity media with different storage architectures. Anticipating the terminology of Section 5, this function may perform "Refreshment", "Replication", and "Repackaging", but not "Transformation". Refer to Section 5 for a detailed description of migration issues.

The **Error Checking** function provides statistically acceptable assurance that no components of the AIP are corrupted during any internal Archival Storage data transfer or transformation. This function requires that all hardware and software within the archive provide notification of potential errors and that these errors are routed to standard error logs that are checked by the Archival Storage staff. The PDI Fixity Information provides some assurance that the Content Information has not been altered as the AIP is moved and accessed. Similar information is needed to protect the PDI itself. A standard mechanism for tracking and verifying the validity of all data objects within the archive may also be used. For example, cyclical redundancy checks (CRCs) could be maintained for every individual data file. A higher level of service such as Reed-Solomon coding to support combined error detection and correction could also be provided. The storage facility procedures should provide for random verification of the integrity of data objects using CRCs or some other error checking mechanism.

The **Disaster Recovery** function provides a mechanism for duplicating the digital contents of the archive collection and storing the duplicate in a physically separate facility. This function is normally accomplished by copying the archive contents to some form of removable storage media (e.g. digital linear tape, compact disk - recordable) but may also be

performed via hardware transport or network data transfers. The details of **disaster recovery policies {3e}** are specified by Administration.

The **Provide Data** function provides copies of stored AIPs to Access. This function receives a **data request {3f}** that identifies the requested **AIP(s) {3g}** and provides them on the requested media type or transfers them to a staging area. This function also sends a **notice of data transfer {3h}** to Access upon completion of an order.

4.1.1.4 Data Management

The functions of the Data Management entity are illustrated in Figure 4-4 and detailed in the text that follows.

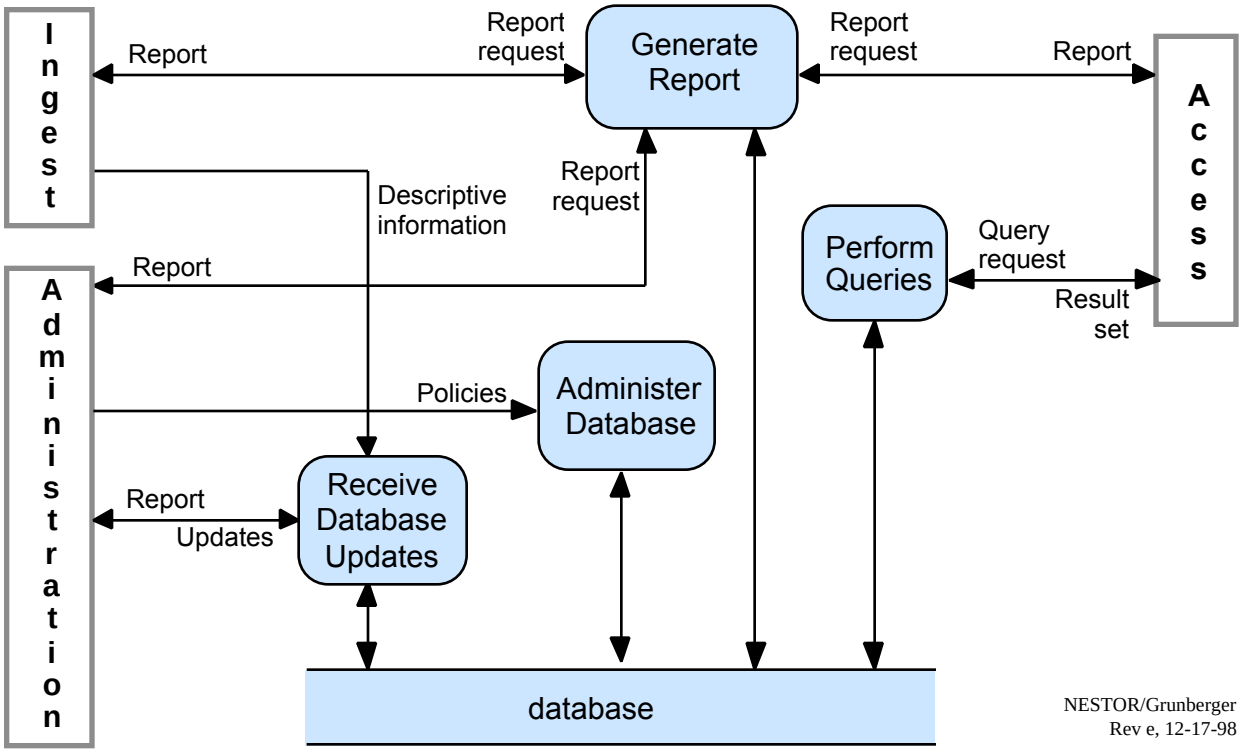


Figure 4-4. Functions of Data Management

The **Administer Database** function is responsible for maintaining the integrity of the Data Management database, which contains both Descriptive Information and system information. Descriptive Information identifies and describes the archive holdings, and system information is used to support archive operations. The Administer Database function is responsible for creating any schema or table definitions required to support Data Management functions; for providing the capability to create, maintain and access customized user views of the contents of this storage; and for providing internal validation

(e.g. referential integrity) of the contents of the database. The Administer Database is carried out in accordance with **policies {4h}** received from Administration.

The **Perform Queries** function receives a **query request {4a}** from Access and executes the query to generate a **result set {4b}** that is transmitted to the requester.

The **Generate Report** function receives a **report request {4c,** 2d**}** from Ingest, Access or Administration and executes any queries or other processes necessary to generate the **report {4d,** 2e**}** that it supplies the to the requester. Typical reports might include summaries of archive holdings by category or usage statistics for accesses to archive holdings.

The **Receive Database Updates** function adds, modifies or deletes information in the Data Management persistent storage. The main sources of updates are Ingest which provides Descriptive Information {2i} for the new AIPs, and Administration which provides **system updates {4e}** and **review updates {4f}**. Ingest transactions consist of Descriptive Information which identifies new AIPs stored in the archive. System updates include all system-related information (operational statistics, Consumer information, and request status). Review updates are generated by periodic reviewing and updating of information values (e.g. contact names, and addresses). The Receive Database Updates function provides regular reports to Administration summarizing the **status of updates {4g}** to the database, and also sends a **database update response {2k}** to Ingest.

4.1.1.5 Administration

The functions of the Administration entity are illustrated in Figure 4-5 and detailed in the text that follows.

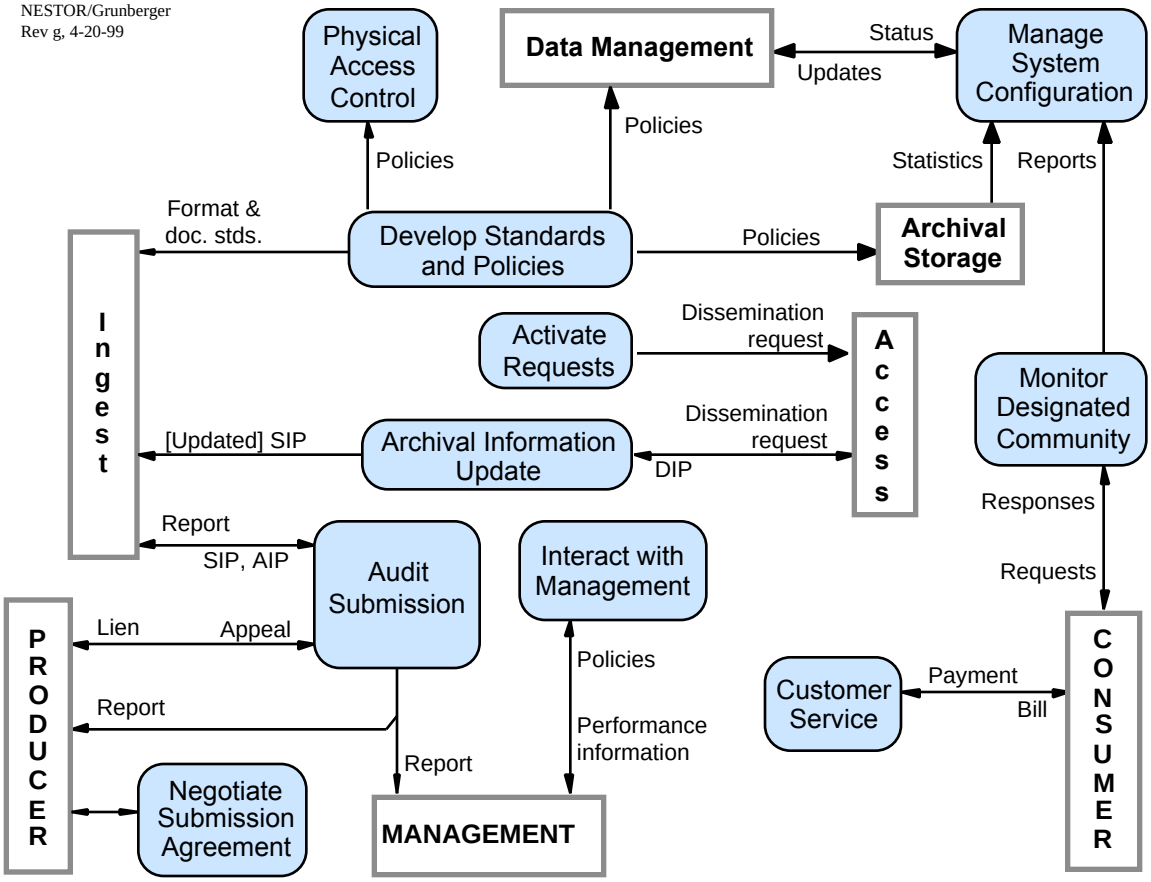


Figure 4-5. Functions of Administration

The Negotiate Submission Agreement function solicits desirable archival information {5a} for the OAIS and negotiates Submission Agreements {5b, 5c} with Producers. This function also negotiates a data submission schedule {5d, 5e} with the Producer. It maintains a calendar of expected Data Submission Sessions that will be needed to transfer one or more SIPs to the OAIS and the resource requirements to support their ingestion. The data submission formats and procedures must be clearly documented in the archive's data submission policies {5f} and the deliverables must be identified by the Producer in the Submission Agreement.

The **Manage System Configuration** function provides system engineering for the archive system to systematically control changes to the configuration. This function maintains integrity and tractability of the configuration during all phases of the system life cycle. It also audits system operations, system performance, and system usage. It continuously monitors the functionality of the entire archive system and prepares recommendations and plans for system evolution.

The **Archival Information Update** function provides a mechanism for updating the contents of the archive. It may also provide for updates to AIPs stored in Archival Storage or Descriptive Information stored in Data Management by sending an **order {5g}** to Access,

updating the contents of the resulting **DIPs {5h}** and resubmitting them as **SIPs {5i}** to Ingest.

The **Physical Access Control** function provides mechanisms to restrict or allow physical access (doors, locks, guards) to elements of the archive as determined by archive policies.

The **Develop Standards and Policies** function is responsible for developing and maintaining the archive system data standards. These standards include format standards, documentation standards and the procedures to be followed during the ingestion process. It will also develop storage management policies {3a} (for the Archival Storage hierarchy), **migration policies {5j}** to assure that archive storage formats do not become obsolete, and database administration policies {4h}. It will also determine security policies for the contents of the archive including those affecting Physical Access Control.

The **Audit Submission** function will verify that submissions meet the specifications of the Submission Agreement. This function is carried out by the archive data engineers and may also involve an outside committee (e.g., science and technical review). The audit process must verify that the quality of the data meets the requirements of the archive and the review committee. It must verify that there is adequate Representation Information and PDI to ensure the Content Information is understandable and independently usable to the Designated Community. The formality of the review will vary depending on internal archive policies. The Audit process may determine that some portions of the SIP are not appropriate for inclusion in the archive and must be resubmitted or excluded. After the audit process is completed any **liens {5k}** are reported to the Producer who will then resubmit {2c} the SIP to Ingest or **appeal {5m}** the decision to Administration. After the audit is completed a **final ingest report {5n}** is prepared and provided to the Producer and to Management. Audit methods potentially include sampling, periodic review, and peer review.

The **Interact with Management** function receives and carries out Management **policies {5p}**. These policies include such things as the OAIS charter, scope, resource utilization guidelines, and pricing policies. It also receives status information from various elements of the archive and provides OAIS **performance information {5q}** to Management.

The **Activate Requests** function maintains a record of event-driven requests and periodically compares it to the contents of the archive to determine if all needed data is available. If needed data is available, this function generates an **order {5r}** that is sent to Access. This function can also generate orders [5r} on a periodic basis where the length of the period is defined by the Consumers or on the occurrence of an event (e.g., a database update).

The **Customer Service** function will also create, maintain and delete Consumer accounts and will **bill {5s}** and collect **payment {5t}** from Consumers for the utilization of archive system resources.

The **Monitor Designated Community** function interacts with archive Consumers, Producers and Management to track changes in their service requirements. Such requirements might include data formats, media choices, preferences for software packages, and mechanisms for

communicating with the archive. This function may be accomplished via a periodic formal review process, via community workshops where feedback is solicited or by individual interactions. As part of its activities, it sends out **information requests {5u}** to Consumers, receives **information responses {5v}** back, and provides reports to the Manage System Configuration function.

4.1.1.6 Access

The functions of the Access entity are illustrated in Figure 4-6 and detailed in the text that follows.

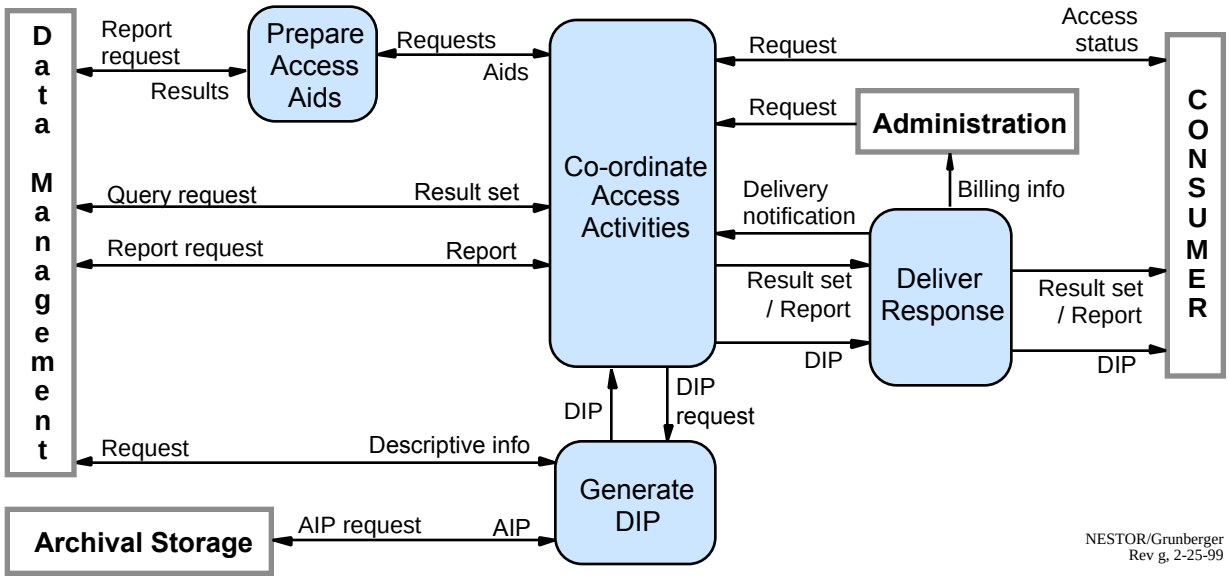


Figure 4-6. Functions of Access

The **Coordinate Access Activities** function provides a single user interface to the information holdings of the archive. This interface will normally be via computer network or dial-up link to an on-line service, but might also be implemented in the form of a walk-in facility, printed catalog ordering service, or fax-back type service. The Coordinate Access Activities function provides access to a suite of finding aids to assist the users in determining the holdings of the archive. These aids are generated by the Prepare Access Aids function in response to requests by Coordinate Access Activities. Three categories of consumer requests are distinguished: **query requests {6b}** which are executed in Data Management and return immediate **result sets {6c}** for presentation to the user; **report requests {6d}** which may require a number of queries and produce a formatted **reports {6e}** for delivery to the Consumer; and **orders {6f}** which may access either or both Data Management and Archival Storage to prepare a formal **Dissemination Information Package (DIP) {6g}** for on- or off-line delivery. An order may be an ad-hoc request that is executed only once or a subscription

request which will be maintained by the Activate Requests function in Administration and may result in periodic deliveries of requested items. Other special request types are allowed but are not detailed. This function will determine if resources are available to perform a request, assure that the user is authorized to access and receive the requested items and notify the Consumer that a request has been accepted or rejected (possibly with an estimate of request cost and an option to cancel the request). It will then transfer the request to data management or to the Generate DIP function for execution. This function also provides **assistance {6h, 6i}** to OAIS Consumers including providing status of orders and other consumer support activities.

The **Prepare Access Aids** function provides tools and products that provide an overview of information products available in the archive system. Access aids include special versions of products which can be quickly viewed such as thumbnails images and abstracts of documents or special interfaces (such as graphical selection menus) to identify and select available data. This function also generates requests for **specialized queries {6a}** produce new representations of the data objects to extend the retrieval capabilities of the archive (e.g., data mining). [DMS note: This last sentence and data flow 6a in Figure 4-7 needs clarification of what is intended.]

The **Generate DIP** function accepts a dissemination request , retrieves the AIP {3g} from Archival Storage and moves a copy of the data to a staging area for further processing. This function also transmits a report or query request {2c} to Data Management to generate the **Descriptive Information {6j}** needed for the DIP. If special processing is required, the Generate DIP function accesses data objects in staging storage and applies the requested processes. The types of operations which may be carried out include statistical functions, sub-sampling in temporal or spatial dimensions, conversions between different data types or output formats, and other specialized processing (e.g. image processing). This function places the completed DIP response in the staging area and notifies the Coordinate Access Activities function that the DIP is ready for delivery.

The **Deliver Response** function handles both on-line and off-line deliveries of responses (DIPs, result sets, reports and assistance) to Consumers. For on-line delivery, it accepts a response from Coordinate Access Activities and prepares it for on-line distribution in real time via communication links. It identifies the intended recipient, determines the transmission procedure requested, places the response in the staging area to be transmitted, and supports the on-line transmission of the response. For off-line delivery it retrieves the response from the Coordinate Access Activities function, prepares packing lists and other shipping records, and then ships the response. When the response has been shipped, a notice of processed order is returned to the Coordinate Access Activities function and **billing information {6k}** is submitted to Administration.

4.1.2 DATA FLOW DIAGRAMS

The flow of data items among the OAIS functional entities is diagrammed in this section. Figure 4-7 shows the more significant data flows. To avoid complication of this figure, the Administration data flows, which are generally background activities, are isolated to an

Administration context diagram, Figure 4-8. Data flows associated with Common Services are implicit in the illustrated functions, and are therefore not shown.

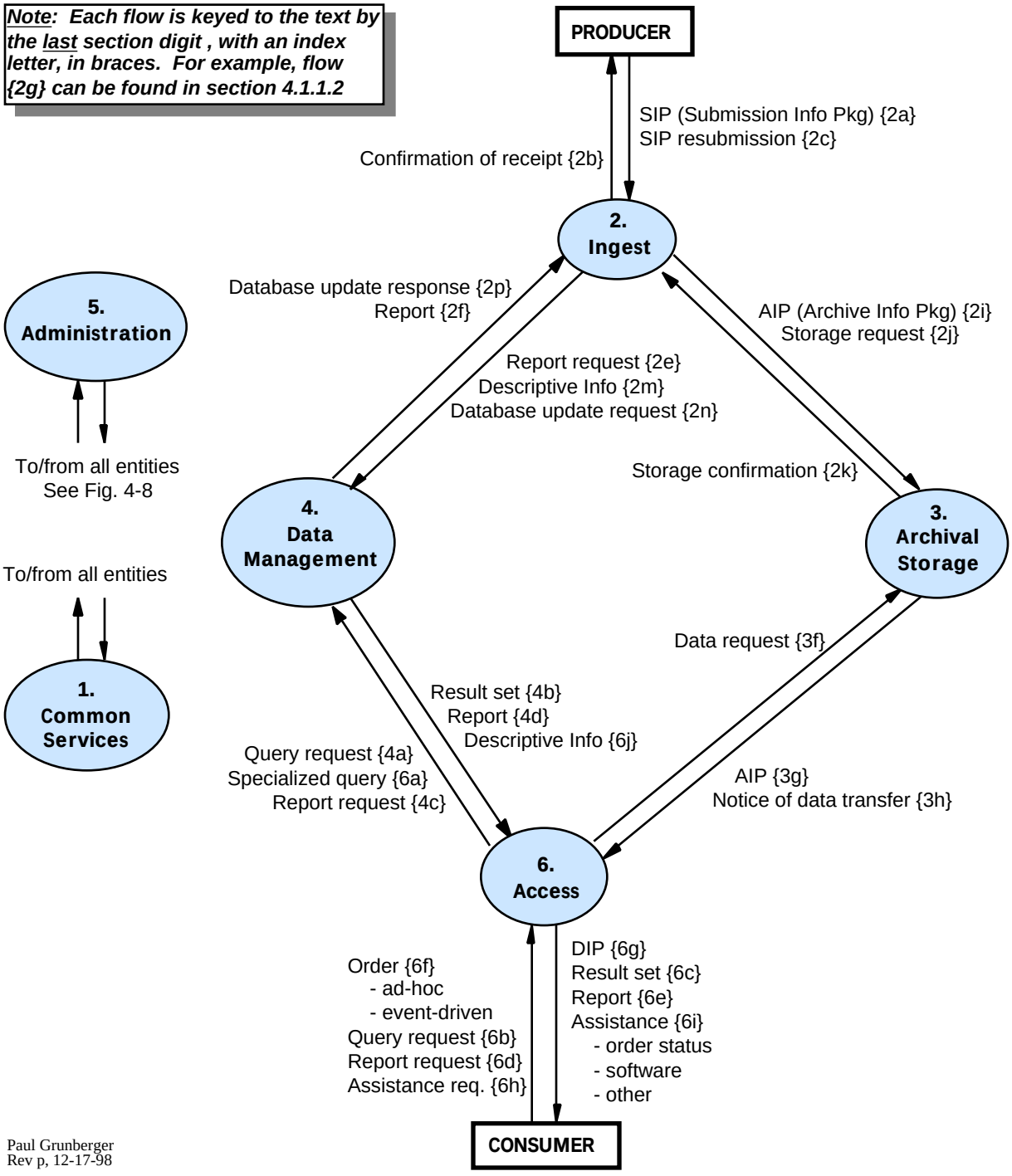


Figure 4-7. OAIS Data Flow Diagram

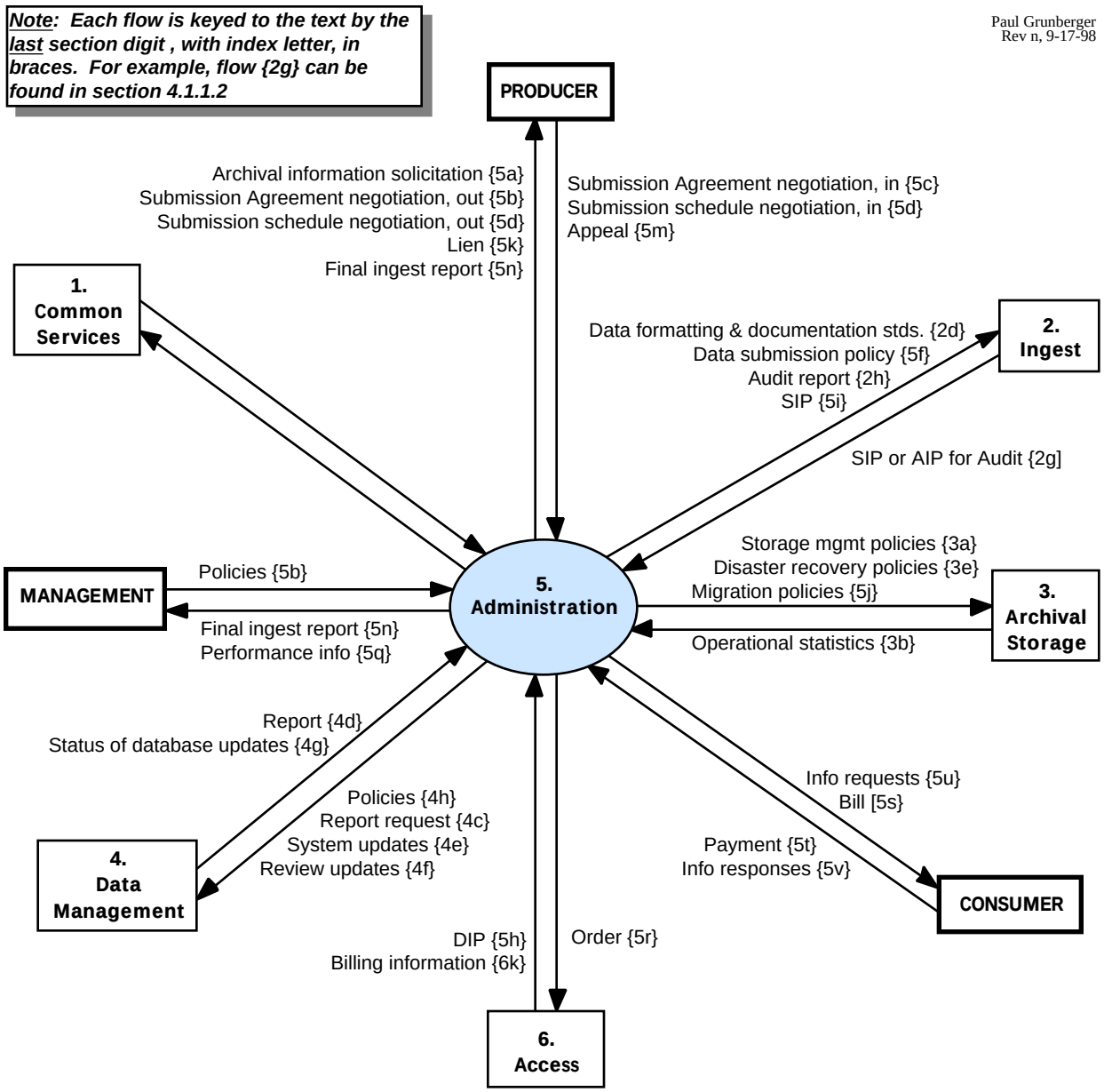


Figure 4-8. Administration Context Diagram

4.2 INFORMATION MODEL

This section builds on the concepts presented in section 2 to further describe the types of information that are exchanged and managed within the OAIS. This section also defines the specific information objects that are used within the OAIS to preserve and access the information entrusted to the archive. This more detailed model of OAIS related information

objects is intended to aid the architect or designer of future OAIS systems. The objects discussed in this section are conceptual and should not be taken to imply any specific implementations.

As discussed in Section 2, the primary goal of an OAIS is to preserve information for a designated community over an indefinite period of time. In order to preserve this information an OAIS must store significantly more than the contents of the object it is expected to preserve. This section analyzes those information requirements to describe the object classes of data associated with an OAIS. This section uses Universal Modeling Language (UML) object model diagrams to illustrate the concepts discussed in the text. An overview of the notation used and critical object modeling concepts is presented in Annex D of this document.

Section 4.2.1 provides a model of the information required for effective long-term preservation of information. Section 4.2.2 describes the conceptual objects and containers that represent the contents of an OAIS.

4.2.1 LOGICAL MODEL FOR ARCHIVAL INFORMATION

4.2.1.1 Information Object

A basic concept of the OAIS Reference Model is the concept of information being a combination of data and Representation Information. The UML diagram in figure 4-9 illustrates this concept. The **Information Object** is composed of a Data Object that is either physical or digital and the Representation Information that allows for the full interpretation of the data into meaningful information. This model is valid for all the types of information in an OAIS.

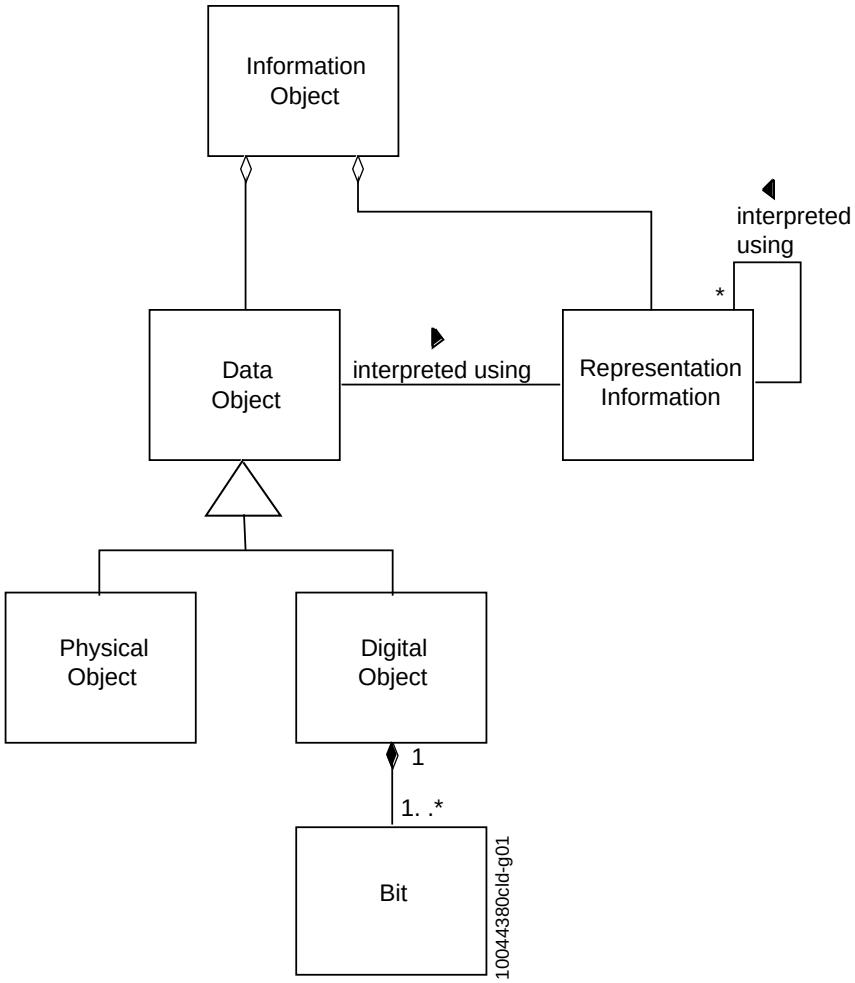


Figure 4-9. Information Object

4.2.1.2 Data Object

The Data Object may be expressed as either a physical object (e.g., a moon rock) together with some Representation Information, or it may be expressed as a digital object (i.e., a sequence of bits) together with the Representation Information giving meaning to those bits.

4.2.1.3 Representation Information

The Representation Information accompanying a physical object like a moon rock may give additional meaning, as a result of some analysis, to the physically observable attributes of the rock. This information may have been developed over time and the results, if provided, would be part of the Information Object.

The Representation Information accompanying a digital object, or sequence of bits, is used to provide additional meaning. It typically maps the bits into commonly recognized data types such as character, integer, and real and into groups of these data types. It associates these

with higher level meanings that can have complex inter-relationships that are also described.

The remainder of this section focuses on the Representation Information object when the Data Object is specialized as a Digital Object.

4.2.1.3.1 Representation Information Types

The Digital Object, as shown in Figure 4-9, is itself composed of one or more bit sequences. The purpose of the Representation Information object is to convert the bit sequences into more meaningful information. It does this by describing the format, or data structure concepts, which are to be applied to the bit sequences and that in turn result in more meaningful values such as characters, numbers, pixels, arrays, tables, etc. These common computer data types, aggregations of these data types, and mapping rules which map from the underlying datatypes to the higher level concepts needed to understand the Digital Object are referred to as the **Structure Information** of the Representation Information object. These structures are commonly identified by name or by relative position within the associated bit sequences.

The Representation Information provided by the Structure Information is seldom sufficient. Even in the case where the Digital Object is interpreted as a sequence of text characters, and described as such in the Structure Information, the additional information as to which language was being expressed should be provided. This type of additional, required information is referred to as the **Semantic Information**. When dealing with scientific data, for example, the information in the Semantic Information can be quite varied and complex. It will include special meanings associated with all the elements of the Structural Information, operations that may be performed on each datatype, and their inter-relationships. Figure 4-10 emphasizes the fact that Representation Information must contain both Structure Information and Semantic Information.

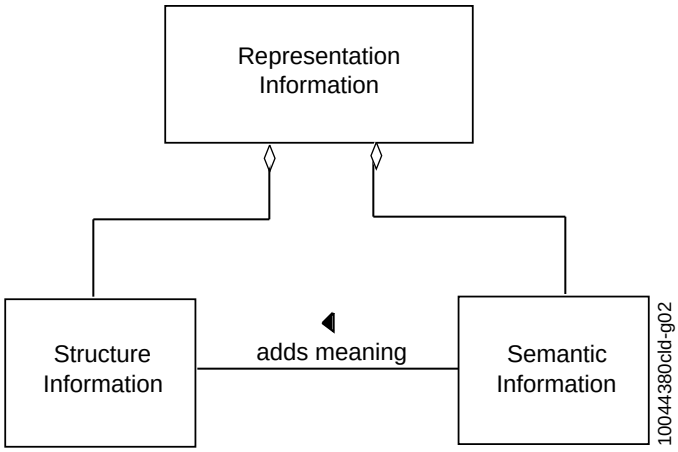


Figure 4-10. Representation Information Object

4.2.1.3.2 Representation Networks

Representation Information, which is itself an information object, may be expressed in physical forms (e.g., a paper document) or in digital forms. When the Representation Information is in digital form, additional Representation Information is needed to understand the bits of the Representation Information. In principle, this recursion continues until physical forms are encountered. For example, Representation Information expressed in ASCII needs the additional Representation Information for ASCII, which may be a physical document giving the ASCII standard. Because each Representation can be composed of multiple components, each with is own Representation; the result can be described as a **Representation Network**. In practice, the recursive chain of the Representation Network may also be broken when there is widely available software that understands a particular representation, such as ASCII display software. Though this is a common practice, it is also dangerous.

Software modules that have a built in understanding of some representation net leaf node, or a collection of such nodes, are often used to terminate the recursive chain in a Representation Network. The very minimum services for such software is to be able to present information, from a part of the digital object, transformed by its understanding of the relevant part of the representation information, to a human or other application. Examples of this type of software include word processor's supporting complex document format representations and scientific visualization systems supporting representations for expressing time series or multidimensional arrays. The software uses its knowledge of the underlying Representation Information to provide these services. This is useful as long as the software executes properly. However for indefinite long-term information preservation, a full and understandable description of the Representation Information is essential.

One problem with using software to terminate a Representation Network is that the software can give a false view of the underlying Content Information and its associated Representation Information. A software module can provide services other than those originally intended from the representation information alone. For example, the services may include the ability to edit the underlying digital object or to perform transformations, which go well beyond simply providing the representation information in new forms. In another example, the software may give a view about relationships within the information that are not contained in the standard (or original) representation information. In essence, this software has expanded on the original representation information by expanding on the assumed relationships among fields of the data object. As a concrete example, a table of numbers may have representation information that says, in essence, 'here is a set of pairs of temperature versus altitude, as measure by instrument x over time period y. Software to display this information might display it as a smooth curve assuming a standard interpolation between the given points.

A software module, in relation to the representation information of the relevant Content Information, may also have the problem that it does not present all the information of the underlying representation. It may have been designed to do a specific job other than

presenting all information, or it may have misunderstood the representation information. Again, it may be difficult to tell from the software alone, what information is missing from presentations to humans or applications.

These situations, which are particularly prevalent with the use of proprietary formats, push in the direction of attempting to preserve the operating software module as a proxy for preserving the information the module is working with. While preserving the experience of working with a particular software module may be a worthy goal in some situations. It may introduce substantial ambiguity as to what is the Content Information that is being preserved.

There are also no assurances software that is common now will be available and operational in the future. The subject of preserving operating software to terminate the Representation Network will be discussed more fully in the next section.

Figure 4-11 illustrates that Representation Information is built from:

- preexisting standards called Referenced Representation Information that define primitive datatypes,
- mapping rules that map those primitive datatype into the more complex datatype concept used by the Data Object,
- Operations and relationships, that may be applied to the newly formed datatype.

The diagram also indicates the recursive nature of Representation Information as an Information Object that requires other Representation Information to enable interpretation.

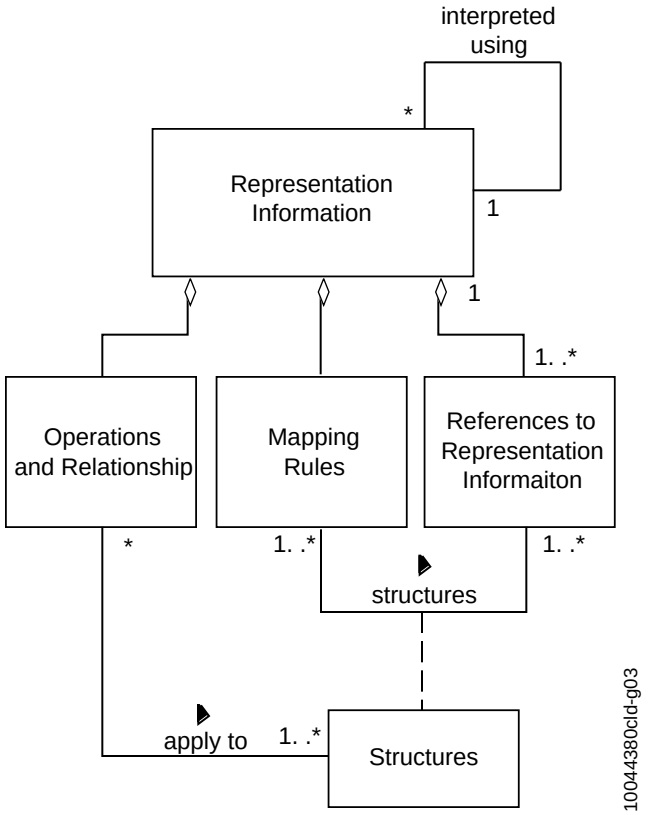


Figure 4-11. Representation Network Object Model

An example of a Representation Network for a multimedia document is shown as figure 4-12. The representation for the syntax of the content object is an international standard such as HTML for expressing on-line documents. In this hypothetical standard for multi-media, the statement is made that real numbers are expressed in the ISO floating point standard, characters in ASCII and images in JPEG. In order to understand the document a consumer would need all of these underlying standards in addition to mapping rules to combine the underlying standards into the concepts required to be expressed by the document. In this example each of the underlying standards is assumed to be an ASCII file so there is no complicated Representation Network for understanding any of the representations. However, if the underlying standards were encoded in a word-processing format it would create an extension to the Representation Network for understanding the underlying standards.

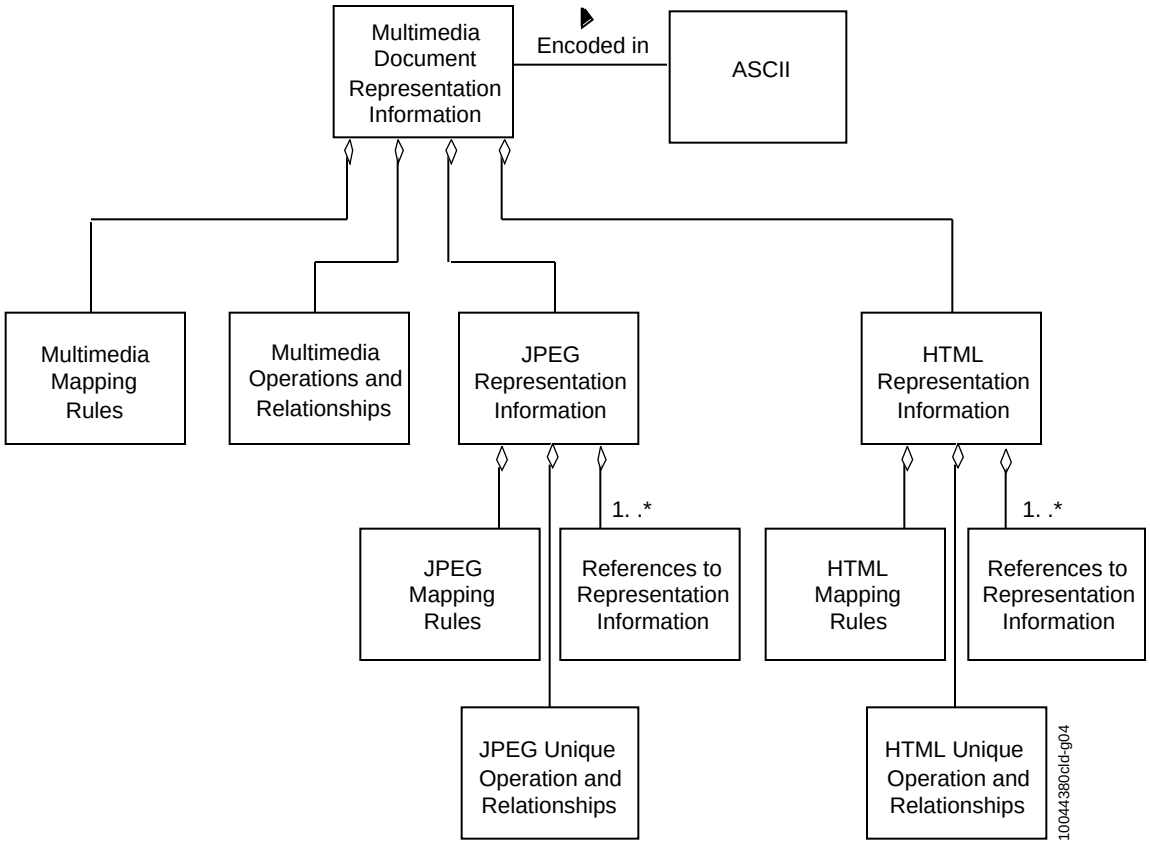


Figure 4-12. Example Representation Network

As illustrated in the previous discussion, the representation requirements for long-term preservation are much more stringent than for contemporary data processing and distribution. These issues are more fully discussed in the next section

4.2.1.4 Taxonomy of Information Object Classes used by OAIS

There are many types of information involved in the long-term preservation of information in an OAIS. Each of these types can be viewed as a complete Information Object in that it contains a data object and adequate Representation Information to understand the data. This section builds on the discussions in section 2.2 about the types of supporting information needed to enable long-term preservation and the discussion in the previous section on the role of Representation Information. The information modeling in this section discusses several types of Information Objects that are used in the OAIS. The objects are categorized by their content and function in the operation of an OAIS into Content Information objects, Preservation Description Information objects, Packaging Information objects, and Descriptive Information objects. The following sections discuss the contents of each of the types of Information Object. Figure 4-14 shows a taxonomy of those Information Objects used within the OAIS.

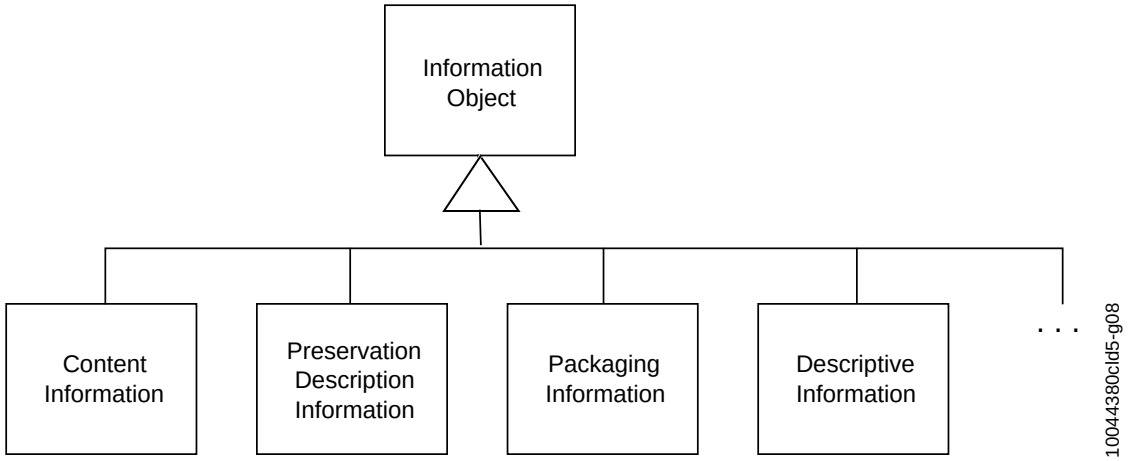


Figure 4-14. Information Object Taxonomy

4.2.1.4.1 Content Information

The **Content Information** is the primary information of interest. Deciding what is the Content Information may not be obvious and may need to be negotiated with the Producer. The Content Information can be viewed as a primary Data Object together with its Representation Information as shown in Figure 4-9. The Data Object in the Content Information may be either a Digital Object or a Physical Object (e.g., a physical sample, microfilm). Any Information Object may serve as Content Information. The special thing about an instance of Content Information is that it is the information that an archive is tasked to preserve.

The Representation Information for the primary Digital Object (both semantic and syntactic) is needed to fully transform the bits into the Content Information. In principal, this even extends to the inclusion of definitions (e.g., dictionary and grammar) of any natural language (e.g., English) used in expressing the Content Information. Over long time periods the meaning of natural language expressions can evolve significantly in both general and in specific discipline usage.

As a practical matter, the OAIS needs to have enough Representation Information associated with the bits of the Data Object in the Content Information that it feels confident that the members of the Designated Community, can enter the Representation Network with enough knowledge to begin accurately interpreting the Representation Information. This is a significant risk area for an OAIS, particularly for those with an expert Designated Community, because jargon and apparently widely understood terms may be short-lived. In such cases extra care needs to be exercised to ensure that the natural evolution of the Designated Community Knowledge Base does not effectively cause information loss from the Content Information.

The Representation Information can also be viewed as including software that supports the presentation of the Content Information to the Consumer. Examples of this type of software include word processor's supporting complex document format representations and scientific visualization systems supporting representations for expressing time series or multidimensional arrays. The software uses its knowledge of the underlying Representation Information to provide these services.

Often required information will be embedded in the software packages used by the Designated Community to present and analyze the Content Information. A reason for preserving working software arises from a convenience factor. Even with a complete set of representation information, practical access to all or part of a digital object requires the use of software. Thus a software module that provides useful access to digital data may be preserved in a working state as a matter of convenience.

This is not difficult to do as long as the environment, which supports the software module, is readily available. This environment consists of some underlying hardware and an operating system, various utilities that effectively augment the operating system and storage and display devices and their drivers. A change to any of these may cause the software module to no longer function, to function incorrectly, or to be unable to present results to the application or human user. The complexity of these interactions is what traditionally makes the preservation of working software such an arduous task.

In summary, the use of software to terminate representation networks is attractive from the point of view of minimizing the resources need to ingest data and provide current users with access to data. However the reliance on working software can provide major problems for Long-term Preservation when that software ceases to function. Indefinite long-term information preservation requires a full and understandable description of the Representation Information. Section 6.2 discusses the techniques that can be used to preserve software over time and the risks associated with this technique.

An important function of the OAIS is deciding which parts of the Content Information are the primary Digital Object and what parts are the Representation Information. This aspect is critical to a clear understanding of what is being preserved. The identification of the Content Information with its Representation Information objects can be addressed by a series of steps, as follows:

1. Identify the bits comprising the primary Digital Object of the Content Information

2. Identify one or more Representation Information objects that together address all the bits and convert them into more meaningful information.

3. For each Representation Information object identified in Step 2, examine each to identify if it is a Referencing Representation Information object. If it is, identify the Representation Information objects it incorporates by reference. Repeat this step until no additional Representation Information objects are identified.

4. For each new Representation Information object identified in Steps 2 and 3 above, that is held as a Digital Object, repeat Steps 2 through 4 until no new Representation Objects are identified.

5. The Content Information consists of the primary Digital Object and each of the Representation Information objects identified in Steps 2 through 4.

As an example of this practice, consider an electronic file containing a sequence of values obtained from a sensor looking at the Earth's environment. There is a second file, encoded using ASCII, which provides information on how to understand the first file. It describes how to interpret the bits of the first file to obtain meaningful numbers, what these numbers mean in terms of the physics of the observation being conducted, the date and time period over which the observations were made, an average value for the observed values, and who made the observations. These two files are submitted to an OAIS for preservation.

Assume that the OAIS determines that the Content Information to be preserved is the observed bits together with their values as numbers and the physics meaning of these numbers. This information is conveyed by the bit sequence within the first file together with the Representation Information from the second file which, is needed to transform the first file's bits into meaningful physical values. Note that neither the first file's underlying media nor the particular file system carrying the bits is part of the Content Information in this example. Only part of the second file's content is considered a part of the Content Information and this is the part that enables the transformation of the bits from the first file into meaningful physical values. In fact this second file does not carry all the Representation Information needed to make this transformation because the following additional information is needed:

- Information that the second file is encoded in ASCII so that it can be read as meaningful characters;

- Information on how the characters are used to express the transformations from bits to numbers to meaningful physics values.

This information, typically referred to as a combination of format information and data dictionary information, may also include instrument calibration values and information on how the calibrations are to be applied. All this information may be widely understandable once the ASCII characters are visible because it has all been expressed in English (or some other natural language), or some of it may be in more structured forms that will need additional Representation Information to be understood.

Therefore the Representation Information of the second file needs additional Representation Information, and this information may need additional Representation Information, etc., forming a linked set of Representations of Representations. This is a good example of the complex Representation Net.

Recall that in the example above, there was a determination that the Content Information

consisted of the observed sensor values and their meanings. This is by no means the only determination that could have been made. It could just as easily have been determined that the primary Digital Object of the desired Content Information was the bit sequences within the first file together with the all the bit sequences within the second file. The fact that some of these latter bit sequences are used to interpret the first files bit sequences is just an example of a set of bits that is somewhat self-describing. It is irrelevant that some of the bits in the second file are the basis for information on the date and time period over which the observations were made, the average value for the observed values, and who made the observations. Once it has been determined that all these bits constitute the primary Digital Object of the Content Information, then the Representation Information is that information needed to turn them into meaningful information. How extensive this meaning is to be carried and how far the Representation Network needs to be carried are local issues for the OAIS and its related Producer and Consumer communities.

4.2.1.4.2 Preservation Description Information

In addition to Content Information, the Archival Information must include information that will allow the understanding of the Content Information over an indefinite period of time. The specific set of Information Objects, which are required for this function, is called collectively called **Preservation Description Information** (PDI). The PDI must include information that is necessary to adequately preserve the particular Content Information with which it is associated. It is specifically focused on describing the past and present states of the Content Information, ensuring it is uniquely identifiable, and ensuring it has not been unknowingly altered.

This information is typical for all types of archives and has been classified in the context of traditional archives. However, the class definitions must be extended for digital archives. The following definitions are based on the categories discussed in the paper "Preserving Digital Information" (REF. 2) The relationship between the concepts in OAIS RM and the Preserving Digital Information paper are discussed in Annex B of this document. Table 4-1 gives illustrative examples of this information for various popular Content Information types.

- **Reference Information**: This information identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Content Information. It also provides those identifiers that allow outside systems to refer, unambiguously, to this particular Content Information. Examples of these systems include taxonomic systems, reference systems and registration systems. In the OAIS Reference Model most if not all of this information is replicated in Package Descriptions, which enable Consumers to access Content Information of interest.
- **Context Information**: This information documents the relationships of the Content Information to its environment. This includes why the Content Information was created, and how it relates to other Content Information objects existing elsewhere. The Context Information differs from the PDI definition in that it does not include the information used in associating logical information with physical media. This type of information is assigned to the Packaging Information in the OAIS Reference Model.
- **Provenance Information**: This information documents the history of the Content

Information. This tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. This gives future users some assurance as to the likely reliability of the Content Information. Provenance can be viewed as a special type of context information.

- **Fixity Information**: This information documents the authentication mechanisms, and it provides any authentication keys used to ensure that the particular Content Information object has not been altered in an undocumented manner. Fixity Information includes special encoding and error correction schemes that may be specific to a single Content Object or a family of Content Objects. Fixity Information does not include the integrity preserving mechanisms provided by the OAIS underlying services, error protection supplied by the media and device drivers used by Archival Storage although the Fixity Information may specify minimum quality of service requirements for these mechanisms

Content Information Type	Reference	Provenance	Context	Fixity
Space Science Data	Object Identifier Journal Reference Mission, instrument, and title attribute set	Instrument Description Processing History Sensor Description Instrument Instrument mode Decommutation map Software Interface Spec.	Calibration history Related data sets Mission Funding history	CRC Checksum Reed-Solomon coding
Bibliographic Information	ISBN Title Author	Printing history Copyright Position in series References	Related References Dewy Decimal System Publishing Date Publisher	Author Digital signature Cover
Software Package	Name Author/Originator Version number Serial Number	Revision History License holder Registration Copyright	Help file User Guide Related Software Language	Certificate Checksum Encryption CRC

Table 4-1. Examples of PDI Types

The OAIS needs to explicitly decide what the exact definition of Content Information is in order to be able to ensure that it also has the PDI needed to preserve the Content Information. Once the Content Information has been determined, it is possible to assess the Preservation Description Information.

4.2.1.4.3 Packaging Information

The **Packaging Information** is that information which, either actually or logically, binds or relates the components of the package into an identifiable entity on specific media. . For example, if the Content Information and PDI are identified as being the content of specific

files on a CD-ROM, then the Packaging Information may include the ISO-9660 volume/file structure on the CD-ROM. These choices are the subject of local archive definitions or conventions. The Packaging Information does not necessarily need to be preserved by an OAIS since it does not contribute to the Content Information or the PDI. However there are cases where the OAIS may be required to reproduce the original submission exactly. In this case the Content Information is defined to include all the bits submitted.

The OAIS should also avoid holding PDI or Content Information only in the naming conventions of directory or file name structures. These structures are most likely to be used as Packaging Information. Packaging Information is not preserved by Migration. Any information saved in file names or directory structures may be lost when the Packaging Information is altered. The subject of Packaging Information is an important consideration to the Migration of Information in an OAIS to newer media. This subject is addressed in detail in section 5 of this document.

4.2.1.4.4 Descriptive Information

The Information Objects described previously in this section provide the information necessary to enable the long-term preservation function of the archive. In addition to preserving information, the OAIS must provide adequate features to allow consumers to locate information of potential interest, analyze that information, and order desired information. This is accomplished through a specialization of the Information Object called Descriptive Information which contain the data that serves as the input to documents or applications called **Access Aids.** The Descriptive Information is generally derived from the Content Information and PDI. The Description Data can be viewed as an index to enable efficient access to the associated Information Package via associated Access Aids. **Access Aids** are documents or applications that can be used to locate, analyze, retrieve, or order information from the OAIS.

4.2.2 LOGICAL MODEL OF INFORMATION IN AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)

The previous section defines the types of Information Objects that are needed by an OAIS to enable the long term preservation of information and effective access to the preserved information by the designated community. This section uses those Information Objects descriptions to model the conceptual information structures required to accomplish these functions. The models presented in this section are not intended to imply a concrete implementation but rather to highlight the relationship among the types of information needed in the archival process.

4.2.2.1 Information Package

The conceptual structure for supporting long term preservation of information is the Information Package. An Information Package is a container that contains two types of information objects, the Content Information and the Preservation Description Information

(PDI); the information package can be associated with two other types of information objects, Packaging Information and Package Descriptions. There are several types of Information Packages that are used within the archival process. These Information Packages may be used to structure and store the OAIS holdings, to transport the required information from the Producer to the OAIS, or to transport requested information between the OAIS and Consumers. There are differing information requirements for each of these functions. The UML diagram in figure 4-15 illustrates the conceptual view of an Information Package. This UML diagram shows that an IP contains 0 or 1 Content Information objects, 0 or more PDI objects and is associated with exactly one piece of Packaging Information, which identifies and delimits the Information Package. The Information Package is also associated with 0 or more Package Descriptions that described the Content Object to enable efficient access.

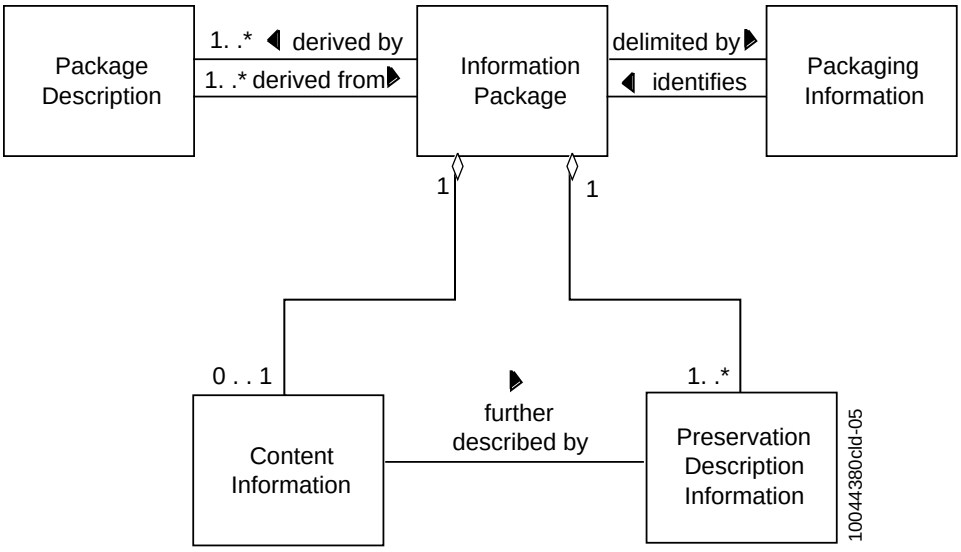


Figure 4-15. Information Package Contents

4.2.2.2 Types of Information Packages

There are three subtypes of the Information Package identified in section 2.2, the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP). The definitions of these package types in section 2 are based on the function of the archival process which uses the package and the translation from one package to another as it passes through the archival process. This taxonomy of Information Package types is shown in Figure 4-16.

It is necessary to distinguish between an Information Package that is preserved by an OAIS and the Information Packages that are submitted to, and disseminated from, an OAIS. These variant packages are needed to reflect the reality that some submissions to an OAIS will have insufficient Representation Information or PDI to meet final OAIS preservation requirements. In addition, they may be organized very differently from the way the OAIS

organizes the information it is preserving. Finally, the OAIS may provide information to Consumers that does not include all the Representation Information or all the PDI with the associated Content Information being disseminated. These variants are referred to as the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP). Although they are all information packages, they differ in mandatory content and the multiplicity of the associations among contained classes.

The **Submission Information Package (SIP)** is that package that is sent to an OAIS by a Producer. Its form and detailed content is typically negotiated between the Producer and the OAIS. Most SIPs will have some Content Information and some PDI, but it may require several SIPs to provide a complete set of Content Information and associated PDI. The Content Information and the PDI both have associated Representation Information, and if there are multiple SIPs involved that use the same Representation Information, it is likely that such Representation Information will only be provided once to the OAIS. As another variation, since some types of PDI will apply to multiple SIPs from the same source, such PDI may be provided in a separate SIP that is without Content Information. The Packaging Information will always be present in some form.

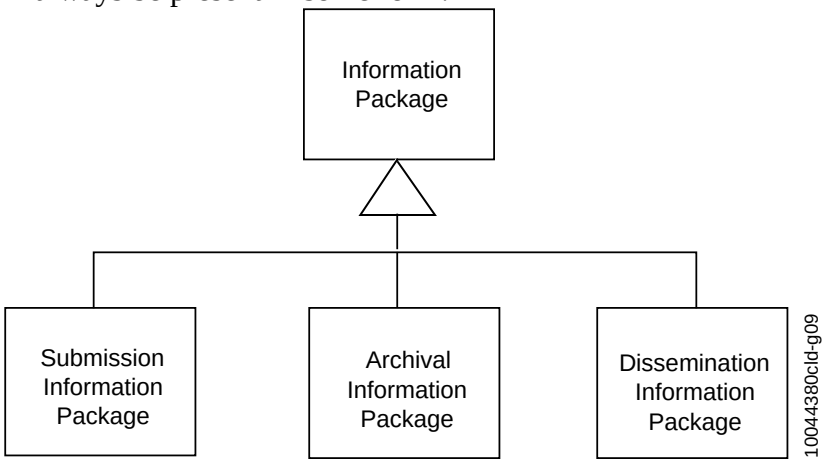


Figure 4-16. Information Package Taxonomy

The Descriptive Information associated with a SIP is likely to be provided prior to submitting the SIP to the OAIS, but it may be provided at any time. It may be no more than a text description with a name or title, carried by the Packaging Information, by which the SIP may be recognized.

Within the OAIS one or more SIPs are transformed into one or more **Archival Information Packages (AIP)** for preservation. The AIP has a complete set of PDI for the associated Content Information. The AIP may also contain a collection of other AIPs and this is discussed and modeled in Section 4. The Packaging Information of the AIP will conform to OAIS internal standards, and it may vary as it is managed by the OAIS. The Descriptive Information associated with an AIP may be extensive and will be managed by the OAIS so Consumers can find and order the Content Information of interest.

In response to an Order, the OAIS provides all or a part of an AIP to a Consumer in the form

of a **Dissemination Information Package (DIP)**. The DIP may also include collections of AIPs, and it may or may not have complete PDI. The Packaging Information will always be present in some form so that the Consumer can clearly distinguish the information Ordered. The Packaging Information may take several forms depending on the dissemination media and Consumer requirements. The Descriptive Information associated with a DIP may be provided with the transfer of the DIP, or it may be provided at any time before or after the transfer. Its purpose is to give the Consumer enough information to recognize the DIP from among possible similar packages. It may be no more than a text description with a name or title, as carried by the Packaging Information, by which the DIP may be recognized.

Though the implementation of the AIP may vary from archive to archive, the specification of the AIP as a container that contains all the needed information to allow long-term preservation and access to archive holdings remains valid. The information model for the AIP presented in the section 4.2.2.3 should be used as a reference to establish the types of information required to enable long-term preservation and access.

The exact information contents of the SIP and DIP and their relationship to the corresponding AIP are dependent on the agreements between the archive and its Producers and Consumers. The model for both of these packages is the same as for the Information Package shown in Figure 4-15 both in mandatory content and the multiplicity of the associations among contained classes. The subject of transformations between SIP and AIP and between AIP and DIP is further discussed in Section 4.3.

4.2.2.3 The Archival Information Package

An **Archival Information Package** (AIP) which is modeled in Figure 4-17 is a specialization of the Information Package. The AIP is defined to provide a concise way of referring to a set of information that has, in principle, all the qualities needed for permanent, or indefinite-long term, preservation of a designated Information Object. The AIP is itself an information object that is a container of other information objects. Within the AIP is the designated information object and it is called the Content Information.

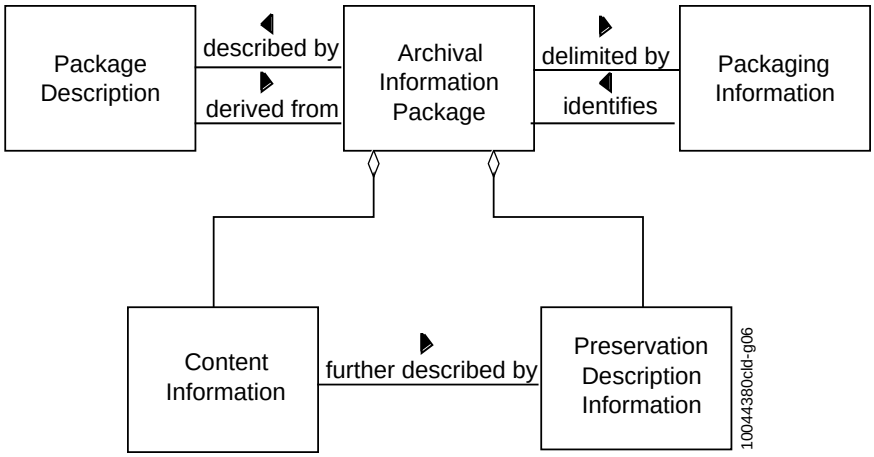


Figure 4-17. Archival Information Package (AIP)

Also within the AIP is an information object called the **Preservation Description Information (PDI)**. The PDI contains additional information about the Content Information and is needed to make the Content Information meaningful for the indefinite long-term.

The Preservation Description Information requirements in an AIP are much more stringent than the requirements for Preservation Description Information in the general Information Package. While no PDI objects are mandatory in an Information Package, all classes of PDI information must be present in an AIP. This is illustrated in Figure 4-18.
The contents of each type of PDI are left to the discretion of the individual archive.

For example, in some OAIS holdings a statement that the creator of the Content Information is unknown may be adequate Provenance Information while in other OAIS holdings it may be mandatory that more complete provenance be researched.

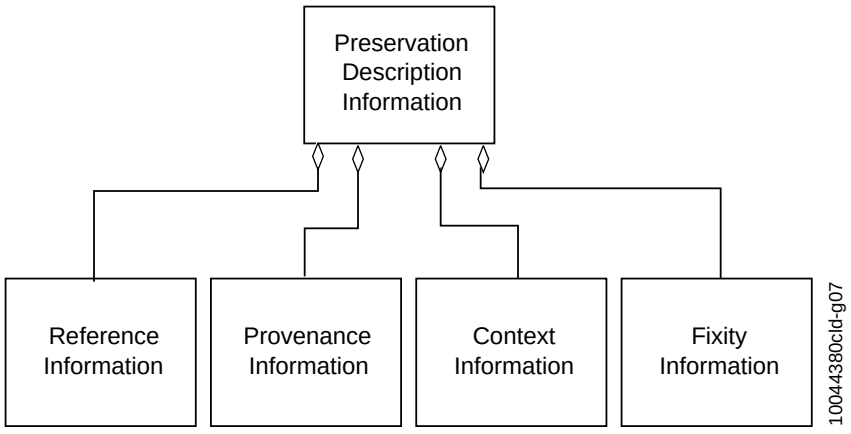


Figure 4-18. PDI Preservation Description Information

The AIP is delimited and identified by the **Packaging Information**. The Packaging Information may actually be present as a structure on the media that contains the AIP or, it may be virtual in that it is contained in the OAIS Archival Storage function. However, the delimitation and internal identification functions must be well defined in an OAIS.

Each AIP is associated with a structured form of Descriptive Information called the **Package Description**, which enables the consumer to locate information of potential interest, analyze that information, and order desired information. The information needed for one Access Aid is called an **Associated Description.** A single Package Description may contain several Associated Descriptions depending on the number of different Access Aids that can locate, visualize, retrieve or order the associated Content Information and PDI. Figure 4-19 is a UML diagram that models the Package Description and Access Aids.

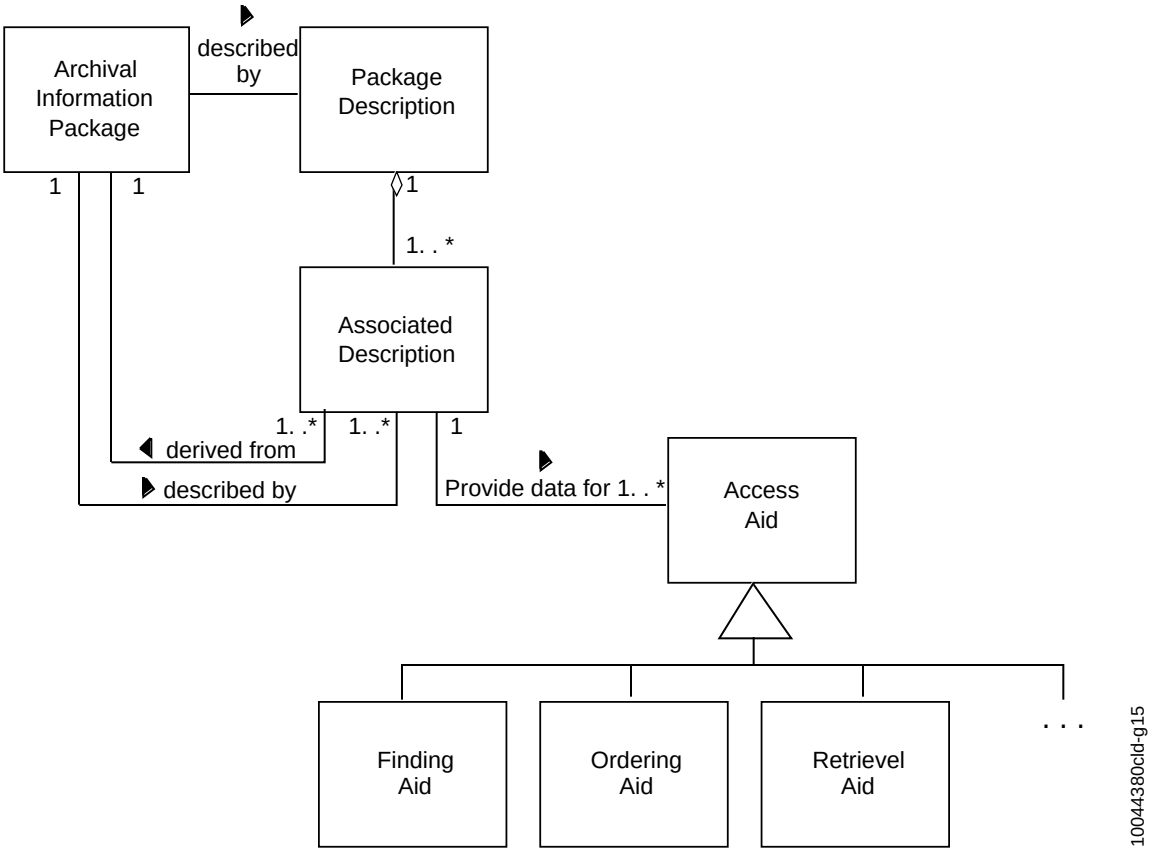


Figure 4-19. Package Description

The Package Description must contain one Associated Description that supplies data for a Retrieval Aid that allows authorized users to retrieve the Content Information and PDI described by the Package Description. This Retrieval Aid is generally part the Archival

Storage functional area and translates from the unique identifier assigned by the OAIS to identify the AIP into the set of operations and filenames needed to retrieve the AIP from the file management system used in Archival Storage and returns the Content Information and PDI for the requested AIP. In most current archives, only internal archive processes and operations personnel and functions are authorized to use this Access Aid. However, as technology advances increase the processing power of the archive and the bandwidth between the archive and the user, such access methods as “content based queries” and “data mining” may provide the user with direct read only access to the content Information.

The Package Description may also contain any number of Associated Descriptions, each of which contains data for one or more Access Aids. Two important subtypes of Access Aid are **Finding Aid** and **Ordering Aid**.

A **Finding Aid** is an application that assists the consumer in locating information of interest. A single AIP may have a number of Associated Descriptions that describe the Content Information using different technologies

An **Ordering Aid** is an application that assists the consumer to discover the cost of and order AIPs of interest. The Ordering Aids also allow users to specify transformations to be applied to the AIPs prior to dissemination. These transformations can include data object transformations such as subsetting, subsampling or format transformations. The transformations can also involve modifying the PDI in the AIP prior to dissemination.

The Package Description is not required for the long-term preservation of the Content Information but is needed to provide visibility and access into the contents of an archive. The contents of the Package Description are highly dependent on the structure of the Content Information and PDI it describes. The uses and types of product Descriptions in an OAIS are further defined in the section 4.2.3.1.

Figure 4-20 gives a detailed view of the Archival Information Package by expanding the PDI and the Content Information. All the "contains " relationships discussed in this section are logical containment relationships. This type of containment relationship may be physical or may be accomplished via a pointer to another object in storage.

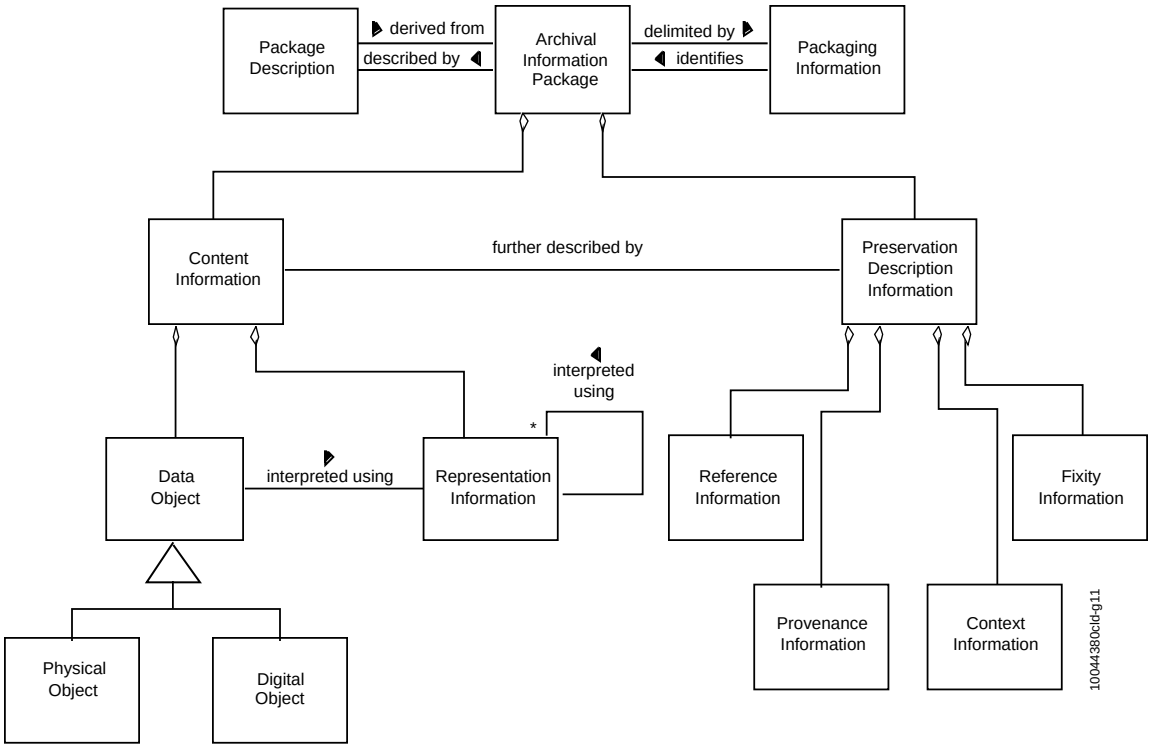


Figure 4 -20. Archival Information Package (Detailed View)

4.2.2.4 Specialization of the AIP and Package Descriptions

Two specializations of the AIP are discussed in this section, the **Archival Information Unit (AIU)** and the **Archive Information Collection (AIC)**. Figure 4-21 is an UML diagram illustrating this specialization. Both AIU and AIC are subtypes of the AIP and as such contain constructs to enable both long-term preservation and consumer access. The AIU represents the type used for the preservation function of a single content atomic object. The AIC organizes a set of AIPs (AIUs and other AICs) along a thematic hierarchy, which can support flexible and efficient access by the consumer community. Conceptually all the AIPs organized by an AIC are contained in the Content Information of that AIC. The differences between AIUs and AICs is the complexity of their Content Information and their associated Package Descriptions and Packaging Information. This reference model considers the differences in the Content Information and associated Packaging and Description functionality between AIU and AIC to be adequately complex and linked to justify the definition of separate classes.

From an Access viewpoint, new subsetting and manipulation capabilities are beginning to blur the distinction between AICs and AIUs. Content objects which used to be viewed as atomic, can now be viewed as containing a large variation of contents based on the subsetting parameters chosen. In a more extreme example, the Content Information of an AIU may not exist as a physical entity. The Content Information could consist of several input files (or

pointers to the AIUs containing these data files) and an algorithm which uses these files to create the data object of interest.

From an information preservation viewpoint the distinction between AIU and AIC remains clear. An AIU is viewed as having a single content information object that is described by exactly one set of PDI. An AIC Content Information is viewed as a collection of other AICs and AIUs, each of which has its own PDI. In addition, the AIC has its own PDI that describes the collection criteria and process.

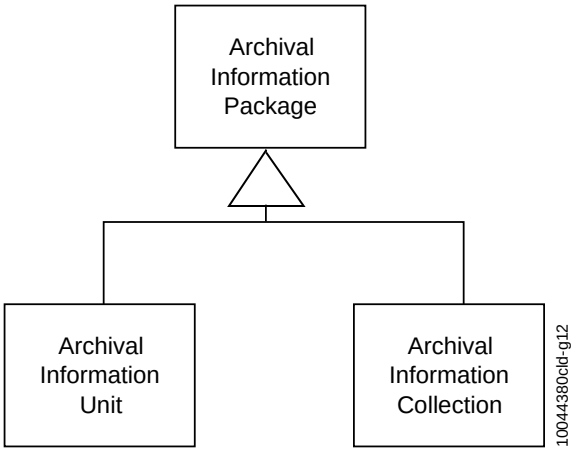


Figure 4-21. Archival Specialization of the AIP

There are two specializations of the Package Description, the Unit Description and the Collection Description. Figure 4-22 is a UML diagram illustrating this specialization. The difference in these two classes is based on the functionality needed to effectively access the contents of an atomic AIU versus the functionality needed to effectively access AIPs that are contained in an AIC.

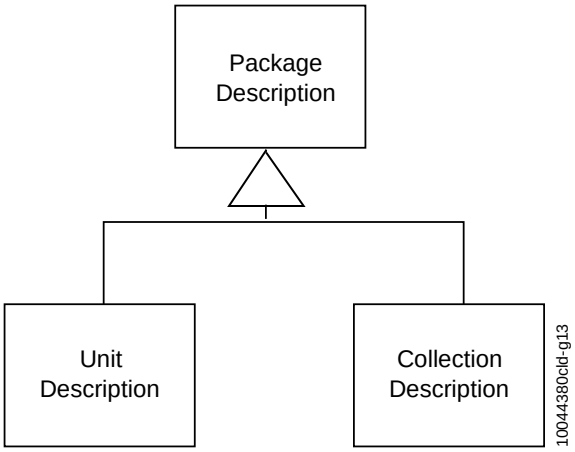


Figure 4-22. Archival Specialization of the Package Description

To aid in the understanding of these constructs, he next two sections of this document will use an example of a company setting up an OAIS of digital versions of movies. This example will focus on the information content of constructs in an AIP. Section 4.3 will illustrate more of the details of the information transformations and dataflows in an OAIS.

4.2.4.1. Archival Information Unit

The AIUs can be viewed as the "atoms" of information that the archive is tasked to store. A single AIU contains exactly one Content Information object (which may consist of multiple files) and exactly one set of PDI. When an information object is ingest into the OAIS a **Unit Description**, which is a subtype of a Package Description is created by extracting information from the Content Information and the PDI and adding OAIS specific information such as a unique identifier .The AIU is illustrated in Figure 4-23.

In the example of a digital movie OAIS, the AIU for a single movie can be viewed as three objects, one containing a digital encoding of the movie in a proprietary format, one containing the Representation Information needed to understand the proprietary format (these two objects form the Content Information), and the other containing facts about the movie such as date of creation, featured actors, director, producer, sequels, movie studio, and a checksum to ensure the integrity of the digital movie (PDI). Since the OAIS reference model is implementation independent, each of these objects could be implemented as one file or multiple files. This type of implementation dependent information is contained in the Packaging Information. When a movie is ingested into the OAIS a Unit Description for an Ordering Aid can be created by extracting information from the Content Information and the PDI and appending it to the unique ordering information.

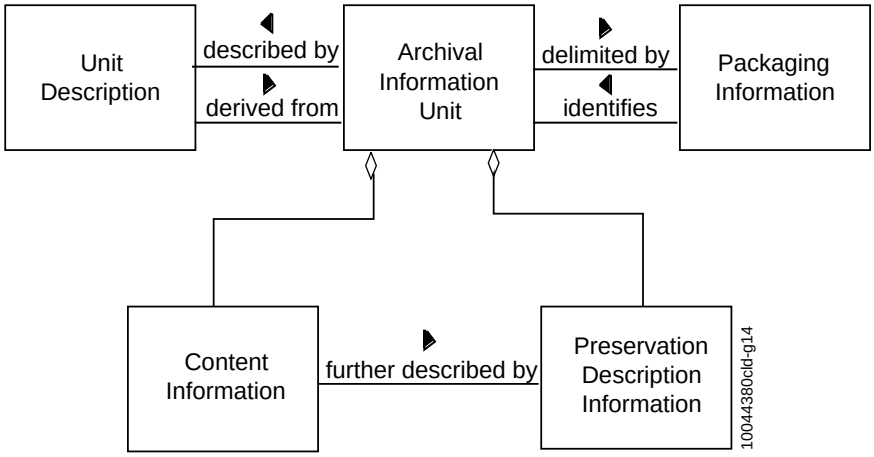


Figure 4-23. Archival Information Unit (AIU)

4.2.4.2 Unit Description

The Unit Description is a specialization of the Package Description that always contains a set of **Associated Descriptions** each of which describe the AIU Content Information from the

point of view of a single **Access Aid**. Figure 4-24 is an UML diagram that illustrates the Unit Description contents.

All Unit Descriptions must supply an Associated Description for a **Retrieval Aid** which enables authorized users to retrieve the AIU described by the Unit Description from Archival Storage. This description includes the unique identifier assigned to the AIP by Archival Storage during the Ingest Process.

An important type of Access Aid is the **Finding Aid, which** is an application that assists the consumer in locating information of interest. A single AIU may have a number of Associated Descriptions that describe the Content Information using different technologies. Additionally as new description extraction and display technologies become available an archive may want to update the Unit Description associated with each of its AIUs to add a new Associated Description that utilizes the new technology to better describe the AIUs.

In the digital movie OAIS example, initially, there may be one Associated Description that is a free text description of a movie, another that is a five minute clip and another that is a row in a relational database that is used by movie collectors to locate movies of interest. After the archive has been operational for a period of time a technique for supplying compressed digital movies may be developed based on recording every tenth frame. The archivist may decide to create an additional type of Associated Description that is populated using the results of this new technique. If desired, the user can run each of the AIUs contained in the archive though this compression technique and create a new Associated Description for each movie in the archive or simply include this Associated Description for new AIUs as they are ingested into the OAIS.

Another important class of Associated Descriptions supply data for **Ordering Aids** that allow the consumer to discover the cost of and order AIUs of interest. The Ordering Aids also allow users to specify transformations to be applied to the AIUs prior to dissemination. These transformations can include data object transformations such as subsetting, subsampling or format transformations. The transformations can also involve modifying the PDI in the AIU prior to dissemination.

For example, the digital movie OAIS could allow a user to order a digital movie as a VHS tape, a laser disc or an MPEG object delivered on-line. Each of these would involve a format transformation and, in theory, an update to the PDI information in the AIP to create accurate PDI for the DIP.

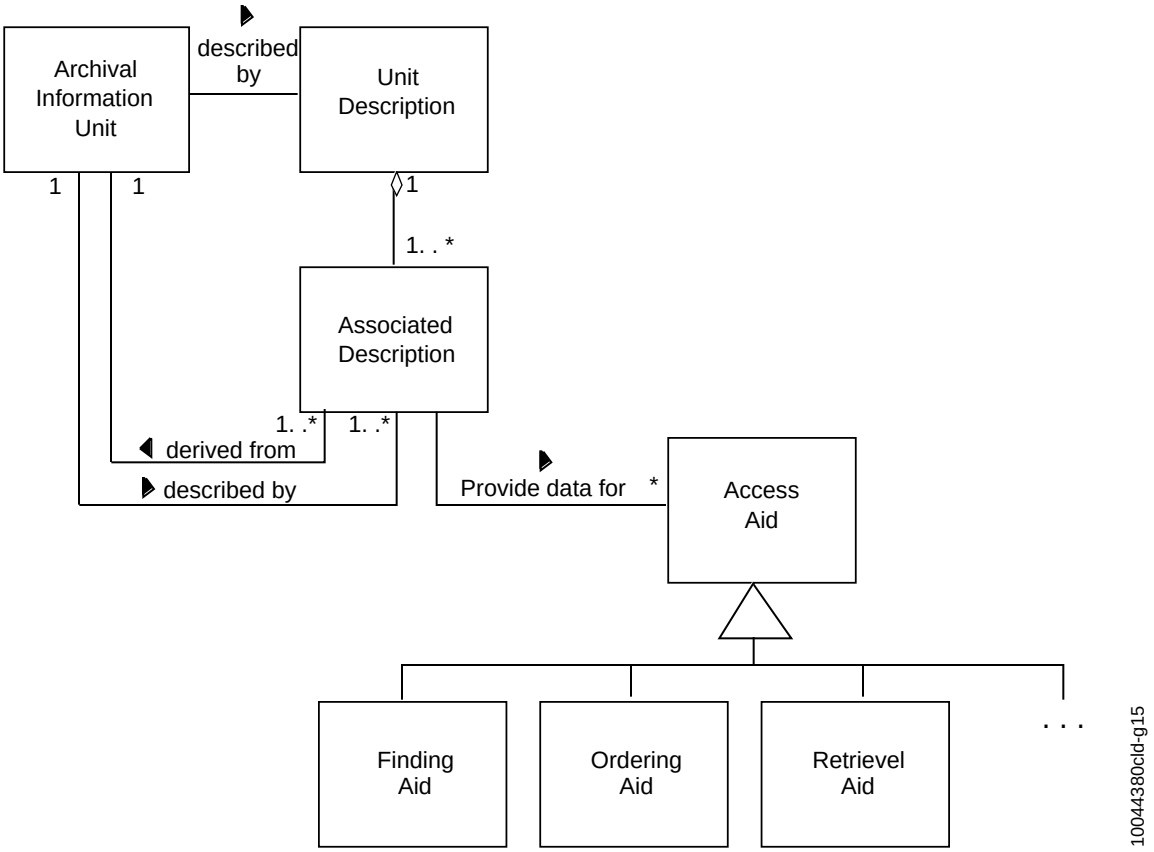


Figure 4-24 Unit Description

4.2.4.3 Archival Information Collections

The AIU and its associated Unit Description provide the information necessary for a consumer to locate and order AIUs of interest. However, it can be impossible for a consumer to sort through the millions of Unit Descriptions contained in a large archive. This problem is addressed here.

The content information of an AIC is composed of complete AIPs each of which have their own Content Information, PDI, and associated Packaging Information and Package Descriptions. These AIPs are then aggregated into Archive Information Collections (AIC) using criteria determined by the archivist. Generally AICs are based on the AIUs of interest having common themes or origins and a common set of Associate Descriptions. At a minimum all OAIS can be viewed has having at least one AIC which contains all the AIPs held by the OAIS.

For example the digital movies OAIS may have AICs based on the subject area of the movie such as mystery, science fiction, or horror. In addition the archive may have AICs based on other factors such as director or lead actor.

.

A logical model of an AIC is shown in Figure 4-25. As in the previous sections all the containment relationships are logical containment and may be physical or may be accomplished via a pointer to another object in storage. For example, the Content Information of an AIC can be created either by creating physical collections of the contained AIPs or by pointing to the contained AIPs. A single AIP can belong to any number of AICs.

For example, a pattern recognition technique might be created for digital movies and the digital movie OAIS might offer a service to search its archives for large structures such as the pyramids or a New York skyline. Note that this type of service is very processing intensive, involving potentially large numbers of AIUs to be transferred from Archival Storage to Access and then running the appropriate process to analyze the Content Information from each AIU. If the results are generally useful, the archivist could summarize the results of this "content based query" into an Associated Description of a new AIC that contains movies with large structures. This technique is frequently referred to as data mining.

An important feature of the AIC as shown in Figure 4-25 is the fact that an AIC is a complete AIP which contains PDI which provides further information about the AIC such as provenance on when and why it was created, context to related AICs, and the desired level of security/fixity information. This is in addition to the PDI contained in member AIPs. This type of information is often necessary for a Consumer to have confidence in the reliability of an AIC. In the above example, the usefulness of the AIC of movies with large structures is to some extent based on the algorithm used and the Provenance of when the AIC was created or last updated.

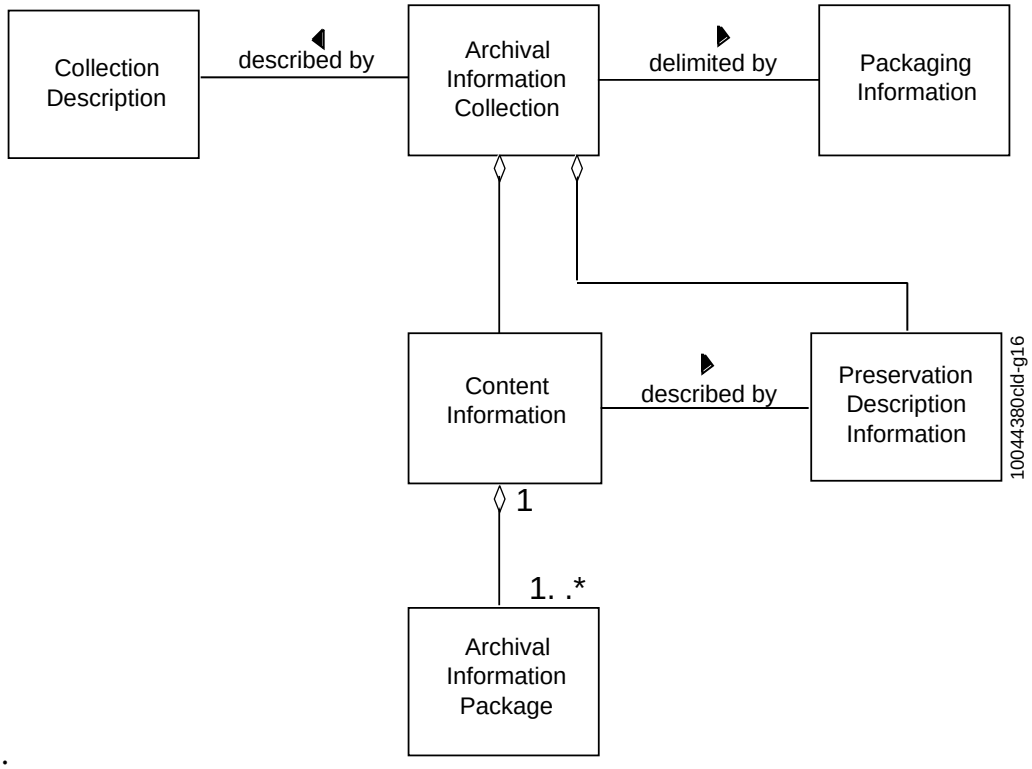


Figure 4-25. Archive Information Collections Logical View

4.2.4.4 Collection Descriptions

The **Collection Description** is a subtype of the Package Description that has added structures to better handle the complex content information of an AIC. The Collection Description, which is modeled in Figure 4-26, contains the information classes that are contained in the Unit Description.

There are two types of Associated Descriptions in a Collection Description:

· Associated Description that describe the collection as a whole (called Overview Description in Figure 4-26)

· Associated Description that separately describe each member of the collection (called Member Description in Figure 4-26)

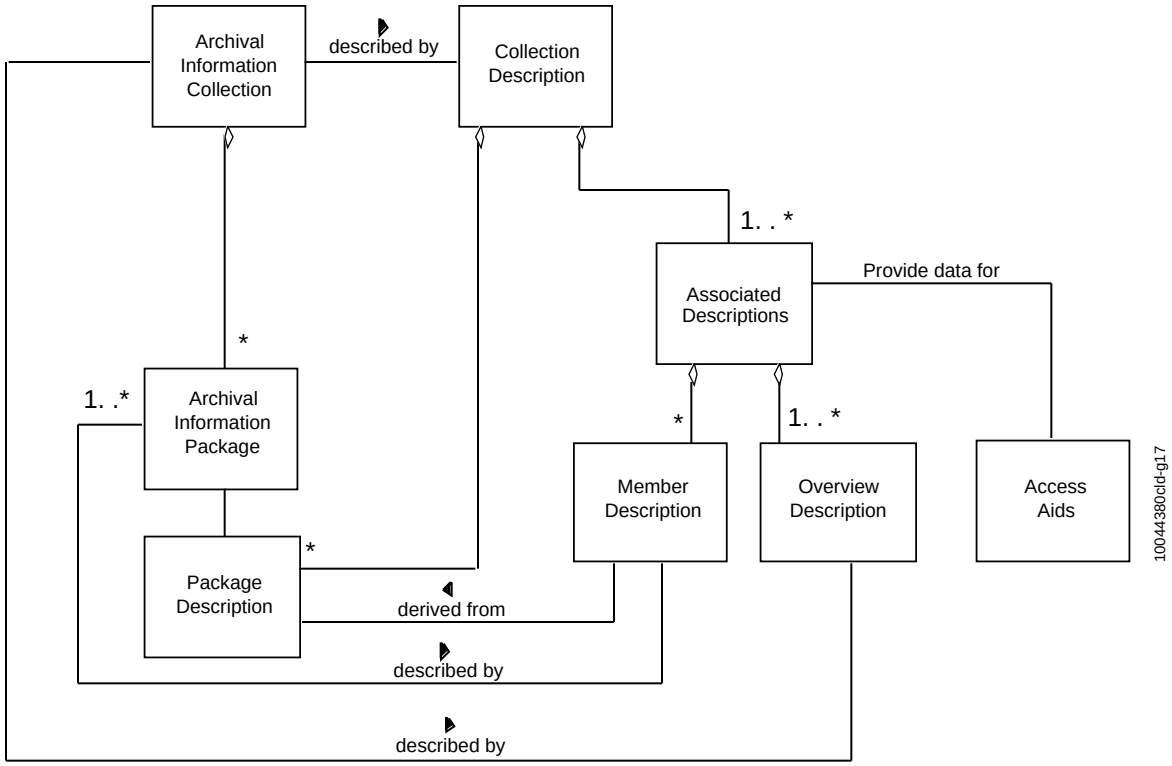


Figure 4-26. Collection Descriptions

The required Associated Description in a Collection Description provides information for Ordering Aids that provide a user with access to the entire set of Content Information of the associated AIC and the PDI for the AIC, but not necessarily to the individual AIPs contained in the AIC. The Collection Description may contain the Package Descriptions of the AIPs contained in the AIC. This containment relationship is logical in that the AIC may either include the Package Descriptions of member Information Packages directly or, more commonly, use pointers to the Package Descriptions of the member Information Packages.

This list of the Package Descriptions for contained AIPs in an AIC could provide Access Aids with a method to Retrieve or Order individual members of the AIC.

It also allows alternative concepts for the implementation of Finding Aids that enable the Consumer to locate AIPs of interest that are contained in an AIC. The Associated Descriptions that provide data for these Finding Aids could be implemented either in a centralized fashion searching an Associated Description in the Collection Description or in a distributed fashion by searching the Associated Description of each member Package Description.

Another important benefit of the Collection Descriptions is the ability to define new **Access Collections**. An Access Collection may be based on new data mining results or to reflect current phenomena or areas of interest that may not be of permanent interest.
One examples of an Access Collection in digital movies OAIS might be a new arrivals collection or a twenty most popular titles collection which is updated periodically. Another example of an Access Collection is a collection based on the results of a pattern recognition algorithm that has not been verified.

To create an Access Collection, an archive would create a Collection Description that did not have an associated AIC. The Collection Description could have a customized Associated Member Description that documented the newly mined description data for each member AIP. A specialized finding aid could use this new Associated Member Description in conjunction with existing Associated Descriptions in the Package Description information of each member AIP to locate AIPs of interest to the user. The Package Descriptions of contained AIPs would also supply data for an Ordering Aid, which would allow the Consumer to order the Information Packages of interest to the consumer.

If an OAIS decides that a Access Collection is valuable enough to be preserved for the long-term, it can store the required Content Information and PDI in Archival Storage thus creating a new AIC.

Another important application of Access Collections is the concept of locating some members of a collection which have been scheduled for ingest at a future time. In this case, the Associated Descriptions supporting a Finding Aid would allow future AIPs to be located. However, the Associated Description for the Ordering Aid and/or the Retrieval Aid would contain the information that this product was not currently available and allow the user to enter an Event Based Order which would be triggered the AIP of interest became available.

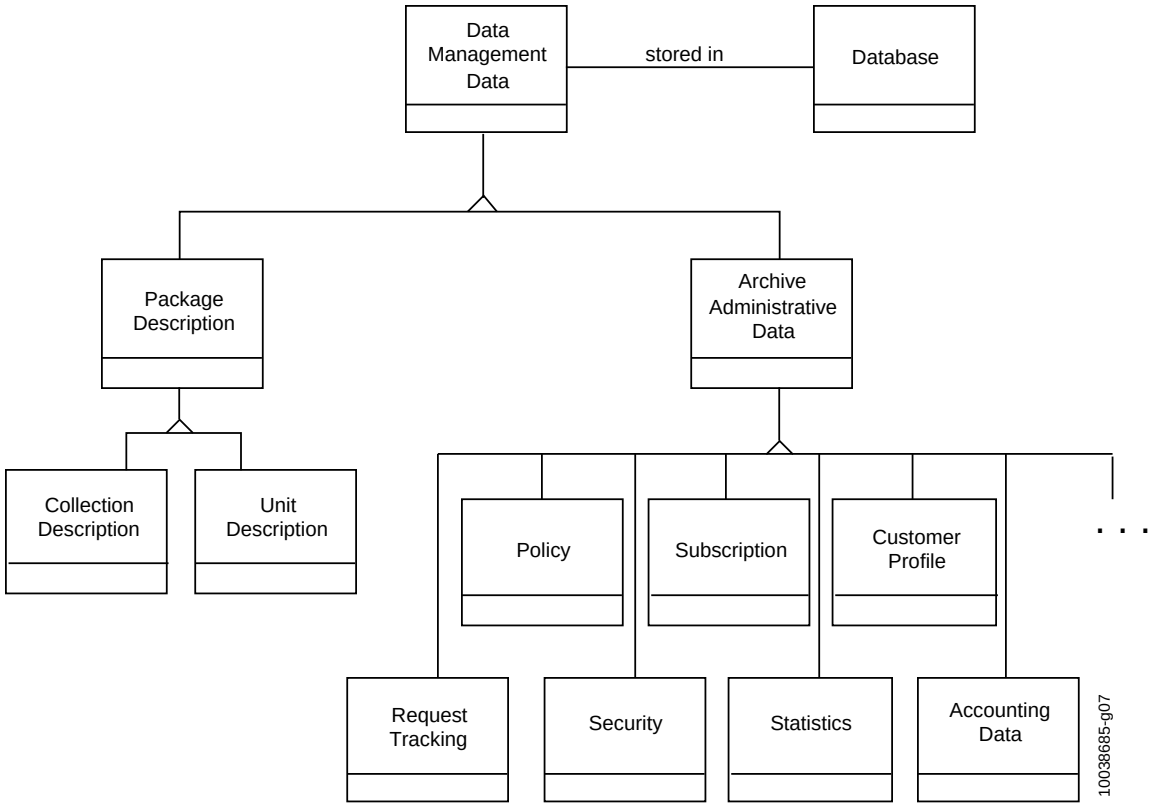
4.2.3 DATA MANAGEMENT INFORMATION

Currently, Package Descriptions are stored in persistent storage such as database management systems to enable easy, flexible access and update to the contained Associated Descriptions. In addition to the Package Descriptions discussed in the previous sections, all the information needed for the operation of an archive could be stored in databases as persistent data classes. Figure 4-27 illustrates the various types of "data management information" within the OAIS. The Archive Administration Information represents the entire

range of information required for the day-to-day operation of the archive. These information include:

- Policy information which provides pricing information and availability constraints for ordering archived information
- Request tracking information that records the progress of each user transaction with an archive. The request tracking process can be very complicated, involving database events and triggers, or as simple as a flat file tracking Order Requests.
- Security information that includes user names and any passwords or other mechanisms needed to authenticate the identity and privileges of archive users.
- Subscription information which provides the information needed to support repeating or future requests
- Statistical information needed by archive administration and Management to determine future policies and performance tuning for more effective archive operation. Examples of these statistics include the number of times an AIP was ordered over a time period and the average time between receiving an order request and shipping the requested holding.
- Customer profile information which enables the archive to maintain facts such as user name and address to avoid the user having to reenter these facts each time he enters a request.
- Accounting information that includes the data necessary for the operation of the archive as a business. The accounting data include payroll data, accounts payable data and accounts receivable data.

These classes are intended as examples rather than an exhaustive list of the data required for archive administration. These classes are conceptual and individual OAIS implementations may vary significantly. For example, individual OAIS may choose to combine the Customer related information types such as Security and Customer Profile into a single database.



4-27. Data Management Information

4.3 INFORMATION PACKAGE TRANSFORMATIONS

The previous portions of this section have discussed the functional architecture of an OAIS and an information architecture to represent the Information Packages and associated Package Descriptions and Packaging Information. This section looks at the transformations, both logical and physical) of the Information Package and its associated objects as they follow a lifecycle from the Producer to the OAIS and from the OAIS to the Consumer.

Figure 4-19 presents a high-level data flow diagram that depicts the principle data flows involved in OAIS operations. These flows do not include administrative flows such as accounting and billing.

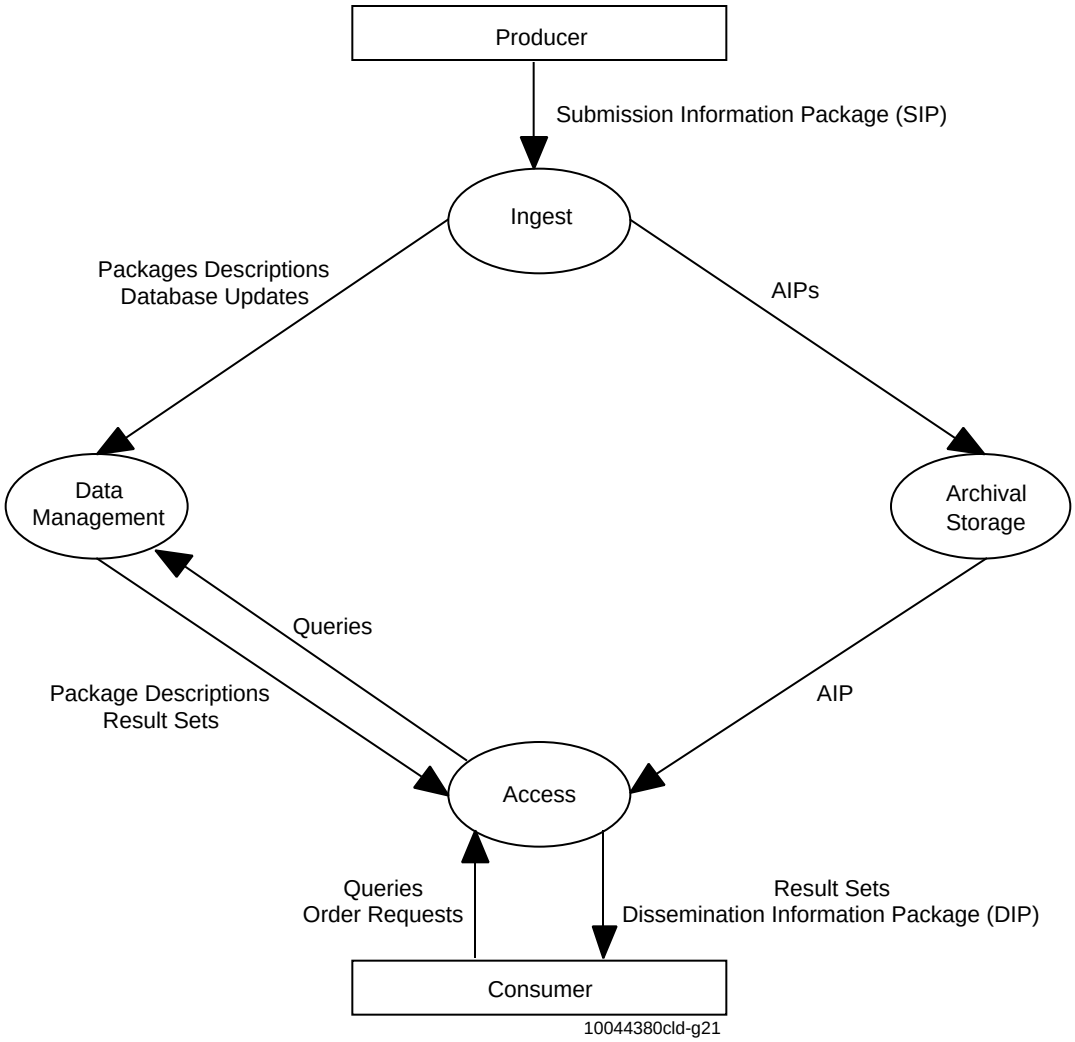


Figure 4-22. High Level Data Flows in an OAIS

4.3.1 DATA TRANSFORMATIONS IN THE PRODUCER ENTITY

The data within the data producer entity are private and may be in any format the producer desired. However, when the decision is made to store the data in an OAIS, the scientific investigators who are responsible for the data meet with archivists to negotiate a Submission Agreement as discussed in Section 2.2.3 of this document. This agreement defines information such as the content, format, and scheduled arrival times of the Submission Information Package (SIP). The SIP is an Information Package that is provided to the OAIS by the producer. The SIP consists of the Content Information plus the data that is necessary to assure that those data can be maintained by the OAIS and that the data can be interpreted and used by Consumers who withdraw them from the OAIS at some time far in the future. These SIPs are periodically transferred to the OAIS in a Data Submission Session. The number of Data Submission Sessions between an OAIS and a Producer can range from a single session in the transfer a final data product to multiple sessions a day in the case of active OAIS which store data for experiments which are still in process. The Data Submission Session can be logically viewed as sets of content data objects and description objects, although physically the description or metadata can be included in the digital objects (i.e., self-describing objects) or divided into many separate descriptive items. In addition to the logical view of data (the SIP), the specification of a data delivery session must also include the mapping of the objects to the media on which they are delivered. This mapping includes the encoding of the object and description and the allocation of logical objects to files.

4.3.2. DATA TRANSFORMATIONS IN THE INGEST FUNCTIONAL AREA

Once the SIP are within the OAIS, their form and content may change. An OAIS is not always required to retain the information submitted to it in precisely the same format as in the SIP. Indeed, preserving the original information exactly as submitted may not be desirable. For example, the computer medium on which our images are recorded may become obsolete, and the images may need to be copied to a more modern medium. In addition, some types of information such as the unique identifier used to locate the Information Package within the OAIS will not be available to the producer and must be input during the Ingest process to the OAIS.

The mapping between SIPs and AIPs is not one-to-one. Here are some examples:

- One SIP – One AIP — A government agency is ready to archive its electronic records from the previous fiscal year. All of the year's records are placed onto magnetic tapes that are submitted as one SIP. The archive stores the tapes together as a single AIP.
- Many SIPs - One AIP — A satellite sensor makes observations of the Earth over a period of one year. Every week all of the latest sensor data are submitted to the archive as a SIP. The archive has a single AIP containing all of the sensor's observations for the year. Ingest merges the Content Information from each weekly SIP into a specified file/files in Ingest persistent storage. The PDI data for the AIP is sent after the last sensor data for the

year has been received. After all of the weekly SIPs and the SIP containing the PDI have arrived, Ingest processes the AIP

- One SIP - Many AIPs — A company submits financial records to an archive as one SIP. The archive chooses to store this information as two AIPs: one that contains public information and the other that contains sensitive information. This makes it easier for the archive to manage access to the information.
- Many SIPs - Many AIPs — An oil and gas company collects information on its wells. Every year it submits SIPs containing all of the well status information for one well to an archive. The archive maintains one AIP for each oil or gas field and breaks out the information on each well to the proper AIP based upon its geographic coordinates.
- One SIP - No AIPs— An investigator (either within or outside of an archive) creates a new algorithm for detecting hurricanes in images. He runs this algorithm over all the images contained in an archive. This data is combined into either a new Associated Description or a set of Package Descriptor updates which is input as a SIP.

The ingest process transforms the SIPs received in the Data Submission Session into a set of AIPs and Package Descriptors which can be stored and accepted by the Archival Storage and Data Management functional entities. The complexity of this ingest process can vary greatly from OAIS to OAIS or from producer to producer within an OAIS. The simplest form of the process involves removing the Content Information, PDI and Package Descriptors from the producer transfer media and queuing them for storage by the Archival Storage and Data Management functional. In more complex cases, the PDI and Package Descriptors may have to be extracted from the Content Information or input by OAIS personnel during the ingest function; the encoding of the information objects or their allocation to files may have to be changed; in the most extreme case, the granularity of the Content Information may be changed, and the OAIS must generate new PDI and Package Descriptors reflecting the newly generated information objects. When many SIPs are required for the creation of one AIP, the Ingest functional area will provide staging storage for the SIPs until all the SIPs required for the AIP arrive.

In addition, the Ingest functional entity will classify incoming information objects and determine in what existing collection or collections each object belongs and will create messages to update the appropriate Collections Descriptions after the AIPs are stored in Archival Storage. The OAIS and external organizations may provide additional Associated Descriptions and finding aids that allow alternative access paths to the information objects of interest. Researchers will develop new and fundamentally different access patterns to information objects. It is important that an OAIS's Ingest and internal data models are sufficiently flexible to incorporate these new descriptions so the general user community can benefit from the research efforts. A good example of this type of new associated description are a phenomenology database in Earth Observation, which allows users to obtain data for a desired event such as a hurricane, or volcano eruption from many instruments with a single query. It is important to note that such finding aids may become obsolete unless the data they require are preserved as parts of the AIPs they access.

It is expected that the Ingest functional entity will coordinate the updates between Data Management and Archival Storage and provide appropriate coordination and error recovery.

The AIP should first be stored in Archival Storage. The confirmation of that operation will include a unique identification to retrieve that AIP from Storage. This identifier should be merged into the Package Description prior to the addition of the Collection Description to Data Management

4.3.3 DATA TRANSFORMATIONS IN THE ARCHIVAL STORAGE AND DATA MANAGEMENT FUNCTIONAL AREAS

The Archival Storage functional entity takes the AIPs produced by the Ingest process and merge it into the permanent archive holdings. The Data Management functional entity takes the Package Descriptions produced by Ingest and augments the existing Collection Descriptions to include their contents. The logical model of the ingested data should already be mapped into the logical model of the archives holdings, so the major transformation that occurs in this step is mapping the acquisition session from the ingest physical data model, which will tend to be on staging storage, to the permanent storage of the OAIS, which could range from database management systems (DBMS) to hierarchical file management systems (HFMS), or any mixture of the above.

The internal view of the OAIS is the permanent representation of the archived data, so all encoding and mappings must be well documented and understood. The process of transferring the ingest objects is frequently by a software process such as an HFMS driver or a DBMS. In this case, it is the responsibility of the OAIS to maintain an active copy of the software or careful documentation of the internal formats so the data can be transferred to other systems in the future without loss of information.

4.3.4 DATA FLOWS AND TRANSFORMATIONS IN THE ACCESS FUNCTIONAL AREA

When a Consumer wishes to use the data within the OAIS, he may use a Finding Aid to locate information of interest. These Finding Aids present Consumers with the logical view of the OAIS holdings so the Consumers can decide which AIPs they wish to acquire. At a minimum, the access view is the high-level logical view of the Collection Descriptions discussed in section 4.2.3. In most cases, the OAIS will have spent significant time and effort developing Associated Descriptions and Finding Aids such as catalogs that will aid the user in locating AIPs or AICs of interest. The consumer will establish a Search Session with the Access entity. During this Search Session, a Consumer will use the OAIS finding aids to identify and investigate potential holdings of interest. This searching process tends to be iterative, with a user first identifying broad criteria and then refining the criteria on the basis of previous search results. When the user has identified candidate objects of interest he may use more sophisticated Finding Aids such as browse image viewers or animation to further refine his result set.

Once the Consumer identifies the OAIS holdings he wishes to acquire, he must issue an order request to the OAIS to acquire the data. The data consumer produces a logical view of the desired AIUs and associated Unit Descriptions to be include included in the Dissemination

Information Package. At this point, the consumer issues an order request for this DIP that triggers the negotiation of a request agreement with the Consumer in which the physical details of the Data Dissemination Session such as media type and object format are specified. This process may involve no visible interaction between the consumer and the archive if adequate defaults exist. This order can also specify any transformations the Consumer wishes applied to the AIPs in creating the Dissemination Information Package (DIP).

The Access functional area then records the Order Request in the Administration functional area. When the conditions required to request dissemination based on a recorded Order are met (note for many Orders these conditions are met immediately) the Administration functional area sends a dissemination request to Access. Access then contacts the Storage and Data Management functional areas and requests the AIPs and associated Package Descriptors necessary to populate the DIP requested by the consumer. The Storage and Data Management functional areas create copies of the requested objects in staging storage.

Access then transforms this set of the AIPs and associated Package Descriptors into a set of DIPs and stores those DIPs onto physical distribution (either physical or communications) media to be delivered to the data consumer in a Data Dissemination Session. The complexity of this transformation process can differ greatly on the basis of the level of processing services offered by the OAIS and requested by the data consumer in his order. In the simplest case, the DIP contains duplicates of the AIPs and associated Package Descriptors of interest from storage and data management function. In more complex cases, the desired CI may have to be extracted from the information objects or inserted into self-describing information objects and the encoding of the information objects or their allocation to physical files may have to be changed. In the most extreme case, when the OAIS supports subsetting services, the granularity of the information objects may be changed, and the Dissemination process may generate DIPs and associated Package Descriptors reflecting the new granularity. The mapping between DIPs and AIPs is 1:1 if no transformation are requested, however the use of subsetting services and other product processing options could create many DIPs from a single AIP, or a single DIP based on combining many AIPs.

5.0 PRESERVATION PERSPECTIVES

This section addresses various practices that have been, or might be, used to preserve digital information and to preserve access services to digital information. It uses the functional and information modeling concepts of section 4.2 and applies them to these practices, and it extends the terminology to distinguish significant aspects of these practices. Section 5.1 addresses the preservation of digital information as it is migrated across media and across formats. Section 5.2 addresses the preservation of access services to digital information as technology changes and software is ported to new systems, wrapped to maintain consistent interfaces, or emulated to support legacy applications. Some key issues with various approaches are identified.

5.1 INFORMATION PRESERVATION

The fast-changing nature of the computer industry and the ephemeral nature of electronic data storage media are at odds with the key purpose of an OAIS: to preserve information over a long period of time. No matter how well an OAIS maintains its current holdings, it is likely over time it will need to migrate much of the stored information to different media or to a different hardware or software environment to remain accessible. Today's digital data storage media can typically be kept at most a few decades before the probability of irreversible loss of data becomes too high to ignore. Further, the rapid pace of technology evolution makes many systems much less cost-effective after only a few years.

Digital Migration is defined to be the transfer of digital information, while intending to preserve it, within the OAIS. It is distinguished from transfers in general by three attributes:

- a focus on the preservation of the full information content,
- a perspective that the new archival implementation of the information is a replacement for the old, and
- full control and responsibility over all aspects of the transfer resides with the OAIS.

This section addresses the Digital Migration of AIPs within an OAIS.

5.1.1 DIGITAL MIGRATION MOTIVATORS

Three major motivators are seen to drive Digital Migrations of AIPs within an OAIS. These are:

- **Media Decay:** Digital media, over time, become increasingly unreliable as secure preservers of bits. Even those that are used with some level of error correction eventually need to be replaced. The net result of media decay is that AIP information must be moved to newer media.
- **Improved Cost-Effectiveness:** The rapid pace of hardware and software evolution provides greatly increasing storage capacities and transfer bandwidths at reducing costs. It also drives the obsolescence of some media types well before they have time to decay. In addition, improved AIP packaging designs may be less dependent on underlying media

and supporting systems, and therefore simplify migration efforts, may be recognized. To remain cost-effective, an OAIS must take advantage of these technologies. Depending on the particular technologies involved, the AIP information may have to be moved to new media types not previously supported and it may have to revise its AIP implementations to take advantage of the new technologies.

- **New Consumer-Service Requirements:** The Consumers of an OAIS also experience the benefits of new technologies and consequently raise their expectations of the types and levels of service they expect from an OAIS. These increased services may require new forms of DIPs to service particular Designated Communities, which in turn may drive an OAIS to hold new forms of AIPs to reduce output conversions. Additionally, AIPs typically go through popularity swings and the OAIS may need to provide different levels of access performance to meet Consumer demands over time. This is likely to be satisfied by moving some AIPs to different media that provide increased or decreased levels of access performance. Finally, the Designated Community for a given AIP may be broadened, resulting in the need to revise AIP forms so as to be understandable and usable by this broader community. All of these can result in the migration of AIPs within an OAIS.

Digital Migrations are time consuming, costly, and expose the OAIS to greatly increased probabilities of information loss. Therefore an OAIS has a strong incentive to avoid migrations when possible, and otherwise to reduce manual involvement to save cost.

5.1.2 MIGRATION CONTEXT

Key functional and information modeling concepts from section 4, as they relate to migration perspectives, are summarized in Figure 5-1.

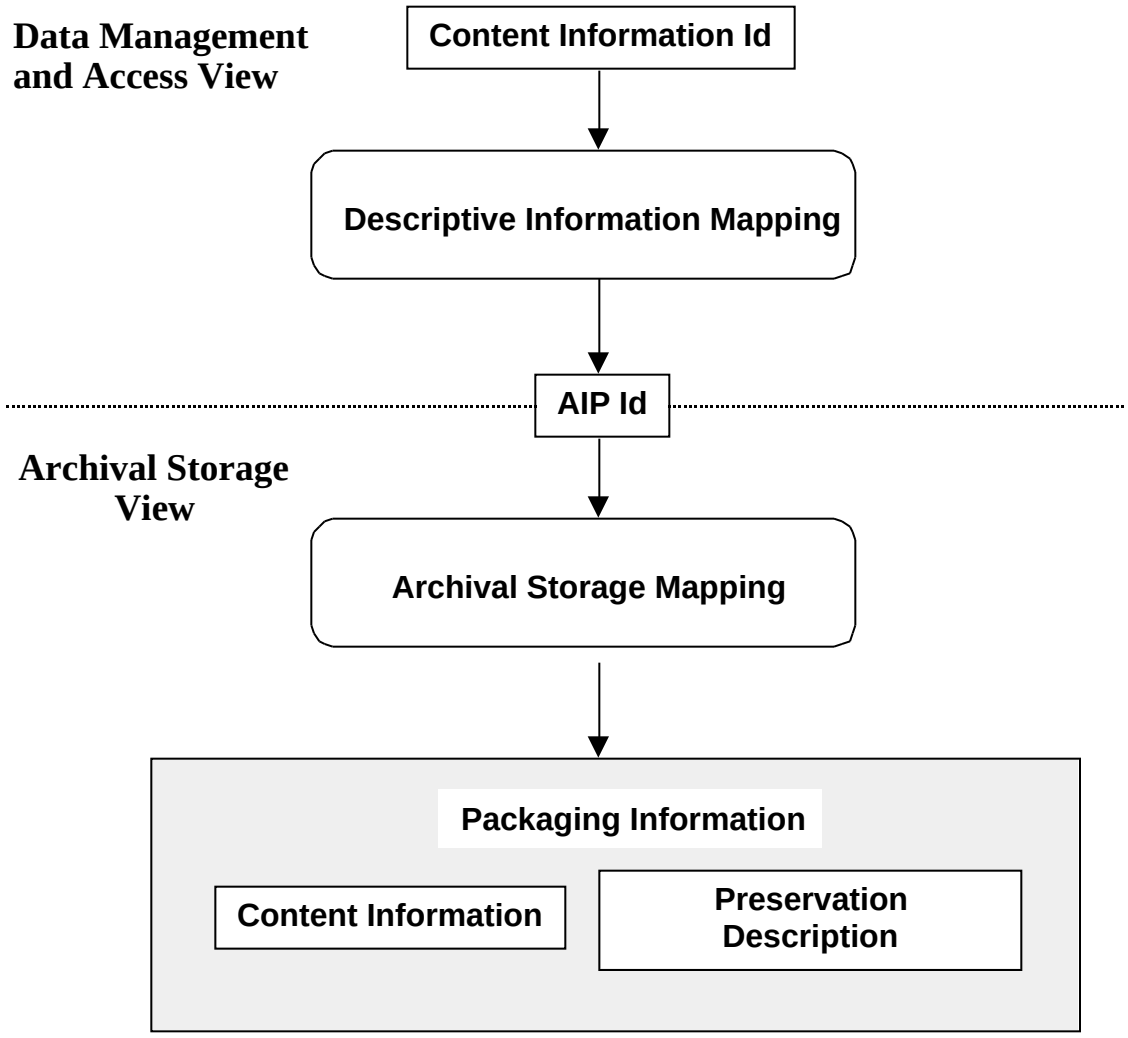


Figure 5-1: Conceptual View of Relationships Among Names and AIP Components

The OAIS Consumer interface in Access provides one or more Content Information Ids, with associated name spaces, to assist in identifying a particular Content Information object of interest. One or more of these Content Information Ids will be included in the PDI Reference Information associated with that Content Information object. The Descriptive Information in Data Management will map each of these Ids to the same AIP Id. The Access Function uses this information to obtain the AIP Id and gives it to Archival Storage to retrieve the associated AIP.

Within Archival Storage, the AIP Id is mapped to the location of AIP Packaging Information

by the Archival Storage mapping infrastructure. The AIP Packaging Information, in turn, logically delimits and identifies the Content Information and the PDI, and binds them into a single entity for preservation. For example, if the Content Information and PDI are determined to be the content of several files, the pointers to documents describing the representations of those files, and the documents themselves, then the Packaging Information would logically be defined as the implementation of the file system holding the file content bits, the data structure holding the pointers, the information which is used to distinguish the Content Information from the PDI, and an encapsulating data structure which identifies the files and other data structures as the components of the AIP Package. The associated Archival Storage mapping infrastructure might then be implemented as a data base which relates the AIP Id to the location of the encapsulating data structure.

The transfer of any part of the Content Information, PDI, or Packaging Information to the same or new media, with the intent that it replaces that part of the previous AIP, is considered to be a Digital Migration of the AIP. Note that a change to the Archival Storage mapping information only, which is outside of the AIP concept, is not considered to be a migration of the associated AIP, although such changes need to be carefully controlled to ensure that access to the AIP is maintained.

The ways in which AIPs are implemented will have a major influence on both the level of automation and the probability of information loss during migrations. Good AIP designs can both increase migration automation and reduce information loss probabilities. To better understand the impacts of these factors on AIP migrations it is useful to categorize migrations into several types and then to consider some issues associated with selected implementation approaches.

5.1.3 MIGRATION TYPES

Based on the models and concepts above, it is possible to identify four primary digital migration types. The primary types, ordered by increasing risk of information loss, are:

- **Refreshment:** A Digital Migration where a media instance, holding one or more AIPs or parts of AIPs, is replaced by a media instance of the same type by exactly copying all bits on the medium used to hold AIPs and to manage and access the medium. As a result, the existing Archival Storage mapping infrastructure, without alteration, is able to continue to locate and access the AIP.
- **Replication**: A Digital Migration where there is no change to the Packaging Information, the Content Information and the PDI. The bits used to convey these information objects are preserved in the transfer to the same or new media-type instance. Note that Refreshment is also a Replication, but Replication can occur without all the constraints of Refreshment.
- **Repackaging**: A Digital Migration where there is some change in the bits of the Packaging Information.
- **Transformation**: A Digital Migration where there is some change in the Content Information or PDI bits while attempting to preserve the full information content.

There is the smallest risk of information loss under Refreshment because none of the bits that are used to hold AIP information or to support finding and accessing AIPs are altered. There is also little risk of information loss under Replication because none of the bits representing AIP information have changed. However if a new media type is involved there will be some changes needed in the Archival Storage mapping infrastructure (see Figure 5-1). The risk is that something may go wrong in the process and some unintended changes to bits may take place. Repackaging recognizes that some bit changes will take place, but these are mostly confined to information used to delimit the Content Information and the PDI, and so generally do not alter the information carried by the Content Information or the PDI. There is the usual risk that something will go wrong and there are also cases where some interaction between Packaging Information and the Content Information or PDI cannot be avoided. This poses additional risk of information loss. Finally, Transformation poses the most risk because changes to the Content Information or PDI are made.

To understand more clearly what may be involved in these migration types it is necessary to look at possible implementation approaches. It will be seen that some migrations are a mixture of Repackaging and Transformation. It is also important to recall that, for any given AIP the OAIS must first clearly identify what constitutes the Content Information, and only then can the PDI be identified. Following this the Packaging Information can also be identified. Further, there is no single "correct" definition of what should be the Content Information as this must be determined by the OAIS for each AIP it constructs and stores. All these issues are discussed in more detail in the following sections using a series of implementation and migration scenarios.

5.1.3.1 Refreshment

A migration involves Refreshment when the effect is to replace a media instance with a copy that is sufficiently exact that all Archival Storage hardware and software continues to run as before. Consider the following scenario.

> The number of correctable bit errors on a CD-ROM disk has reached a dangerous point and the decision is made to replace it with an exact copy. Once the equivalence between the two has been checked, the new CD-ROM replaces the old CD-ROM and Refreshment has taken place. All AIP components on the CD-ROM are unaltered.

5.1.3.2 Replication

A migration involves Replication when there are no bit changes to the Packaging Information, the Content Information, and the PDI. Consider the following scenarios:

> The Content Information and PDI for an AIP are encapsulated into a standard packaging structure and held in the body of a single file. A Replication migration is easily achieved by simply copying the bit order in the file body to a new file on the same or other media. Changes to the Archival Storage mapping infrastructure may be needed to continue to locate the file, but no change in Packaging, Content

Information, or PDI has taken place. Replication, with this type of Packaging Information, affords ease of migration to new media types with maximum automation and little risk of information loss.

5.1.3.3 Repackaging

A migration involves Repackaging when there is some change to the Packaging Information during the transfer. The Packaging Information plays the critical role of delimiting and relating, at a minimum, the Content Information and PDI. If the Content Information and PDI are themselves composed of multiple components, the Packaging Information may be asked to delimit and relate these as well. These are implementation decisions that the OAIS needs to explicitly recognize. Consider the following scenario:

> All the Content Information and PDI bits for an AIP are contained within the body of three files on a CD-ROM. The Packaging Information consists of the bits used to implement the file and directory structure that provides access to these three files. The contents of the three files are moved to three new files on another media type, with a new directory and file implementation. Even if all the directory and file names have been preserved in the transfer, a Repackaging has taken place because the bits used to represent the Packaging Information have changed.

5.1.3.4 Transformation

Digital Migrations that require some changes to the Content Information or PDI are referred to as Transformations. These changes will be to some of the bits in the primary Digital Object of the Content Information or PDI with corresponding changes in the associated Representation Information. In all cases the intent is to provide maximum information preservation because the resulting AIP is intended to be a full replacement for the AIP that is undergoing Transformation. The new AIP qualifies as a new **Version** of the original AIP.

The Representation Information plays a key role in Transformations and the impacts of the changes on the Representation Information may be used to categorize the Transformations. A Representation Information object can be modeled as consisting of a base set of entities, a set of resulting entities, and mapping rules that define the resulting entities and their relationships in terms of the base entities. Using this model of a Representation Information object, two types of Transformation can be defined: **Reversible Transformation** and **Non-Reversible Transformation.**

A Reversible Transformation occurs when the new representation defines a set (or a subset) of resulting entities that are equivalent to the resulting entities defined by the original representation. This means that there is a one-to-one mapping back to the original representation and its set of base entities. An example is replacing a representation that uses the ASCII codes “A through Z” with a representation that uses the UNICODE UTF-16 codes for “A through Z”. The Transformation will result in the replacement of 7-bit codes with 16-bit codes in the AIP object undergoing change. The reverse Transformation can subsequently be performed by replacing the UNICODE UTF-16 codes for “A through Z”

with the ASCII codes for “A through Z” and the original AIP is recovered.

A Non-Reversible Transformation occurs when a Reversible Transformation can not be guaranteed. For example, replacing an IBM 7094 floating point value with an IEEE floating point value is a Non-Reversible Transformation because the resulting entities of these two representations are not semantically equivalent. One will have more precision than the other. However they may be sufficiently equivalent, depending on what the values they represent are being used for, to be effectively interchangeable. If this is the case, a Non-Reversible Transformation effectively preserves the information content. For complex formats, where the meanings and relationships among groups are significant, it may be difficult to establish that a Non-Reversible Transformation has adequately preserved the Content Information. Examples of Reversible and Non-Reversible Transformations are given in the scenarios that follow.

The following scenario identifies a Reversible Transformation that occurs when incorporating a loss-less compression function on the Content Information of an AIP.

> All the Content Information bits for an AIP are contained within the body of three files on a CD-ROM. The Packaging Information includes the bits used to implement the file and directory structure that provides access to these three files. The contents of the three files are transferred to a new CD-ROM and in the process they are compressed using a loss-less compression algorithm. This transfer is a Transformation because the compression process has altered the Content Information, and it is a Reversible Transformation because there is a decompression algorithm that will return the original file content bits. The relevant Representation Information components of the original Content Information needs to be updated to include the decompression algorithm, and the PDI information also needs to be updated, in forming this new AIP version.

The following scenario identifies a Non-Reversible Transformation that can occur when Content Information is migrated to a new format that can express a more varied data model than the original format.

> All the Content Information bits for an AIP are contained within the body of three files on a CD-ROM. The Packaging Information includes the bits used to implement the file and directory structure that provides access to these three files. The contents of the three files are transferred to a new CD-ROM and in the process the third file is altered because there are no longer readily available tools to make effective use of the third file’s content in its current form. The new format, which is in common use, employs a different data model from that of the original format and there are many ways in which the information may be mapped into the new format. This mapping must be carefully done to ensure there is no significant information loss to the Designated Community. This mapping must be included in the PDI, and of course the Representation Information describing the new format will replace that which was describing the previous format. The result is a new AIP version. This is a Transformation migration that is also a Non-Reversible Transformation when there is

no algorithm that will reproduce the original file from the new file.

The following scenario identifies a Reversible Transformation that includes Repackaging. It occurs when the Content Information contains an embedded file name that is a pointer to one of its components, and the Content Information is moved to a new media type with new names for the files.

The Content Information for an AIP is defined to be the body of three files on a CD-ROM. The first file contains an internal name that links the third file and specifies a relationship between them. The Packaging Information includes the directory and file structure that identifies the three files. During a migration to a new media type, these three files are put into a new directory and given new names. This constitutes a Repackaging migration because there is a new implementation of the directory and file structure, which is providing the packaging function. However, the internal name must also be updated in order to maintain the link between the first and third files. This update changes the Content Information and means that the migration is also a Transformation. If the internal name had been a universal identifier, it would not have needed changing. However the standardized framework supporting the universal identifier would contain the mapping information leading to the location of the third file and therefore would need updating. This approach is an advantage for an OAIS because it allows updates to be centralized and more easily managed. However, the required technology is more complex and there is no universal agreement on the identification technique to use.

The final scenario identifies a Non- Reversible Transformation that includes Repackaging. It occurs when the Content Information includes file names, directory structure, and associated file attributes. The Content Information is then migrated to a new media type carrying a different implementation of the directory and file name structures that support fewer file attributes.

The Content Information and PDI bits for an AIC are defined to be an aggregation of AIUs where each AIU consists of the body of three files on a CD-ROM together with their file names, file attributes, and directory names. The Packaging Information is the bits used to implement the file and directory structure that provides access to each of the three file instances, but does not include the actual file and directory names. There may be thousands of AIU instances on a single CD-ROM. The transfer of this AIC to a new media type that employs a new representation for the file and directory structure that has fewer file attributes may result in a Non- Reversible Transformation migration as well as a Repackaging migration. This is a Transformation because the Content Information that originally was stored in the file and directory structures must be re-distributed among the new file and directory structures and probably within the body of the files themselves. This is a Non-Reversible Transformation if there is a no algorithmic one-to-one mapping between the resulting file and directory structures and file contents, and the original file and directory structures. It is a Repackaging because there is a new implementation of the directory and file

structure, which was taken to be part of the packaging. The practice of encoding Content Information into a file or directory name increases the risk of information loss because evolution of a data management environment is facilitated by being able to update directory and file names as needed.

5.1.4 DISTINGUISHING AIP VERSIONS, EDITIONS AND DERIVED AIPS

Unless a Digital Migration involves Transformation, it is not considered to create a new AIP version and it is not required that its PDI be updated. In other words, the AIP version is considered to be independent of Refreshment, Replication, and Repackaging that does not affect the Content Information or PDI. This does not mean that the OAIS does not track such migrations; rather it is not required to update the PDI as part of such tracking.

A Digital Migration that involves Transformation results in a new version of the AIP as defined in Section 5.3.4. In this case, the PDI needs to be updated to identify the source AIP and its version, and to describe what was done and why. The new AIP is viewed as a replacement for the source AIP where the information has been preserved to the maximum extent practical.

An AIP may, in some environments, be subject to upgrading or improvement over time. This is not a Digital Migration in that the intent is not to preserve information, but to increase or improve it. This type of AIP change may be referred to as creating a new **Edition**. The new Edition is viewed as a replacement for the previous Edition, but it may be of historical interest to retain the previous Edition.

An OAIS may also find it convenient to provide an AIP that is derived from an existing AIP. It may do this by extracting some information, or by aggregating information from multiple AIPs, to better serve Consumers. This type of resulting AIP may be referred to as a **Derived AIP**. It does not replace any of the AIPs that it was derived from and it is not a result of a Digital Migration.

5.2 ACCESS SERVICE PRESERVATION

An OAIS may wish to preserve a range of Consumer access services in the face of changing technology. To delineate some access service preservation issues and provide terminology; this section addresses two very different scenarios. As a common starting point, it is assumed that a Consumer has located an AIU of interest and wishes to obtain the information contained therein.

The first service allows the Consumer, as a client, to access the AIU using an Application Programming Interface (API) supported by the OAIS. This interface delivers the bits of the Content Information's Digital Object and identifies locations for obtaining associated Representation Information and PDI. However as technology evolves, the OAIS moves to new hardware, new media, and new operating systems. If the OAIS wishes to maintain the same API for its Consumers, it will need to provide a 'wrapper' around part of its new infrastructure to match its services to the established API. The API will need to be

adequately documented and tested to ensure it correctly delivers the AIU Content Information. When the API is not too complex and is applicable across a wide range of the OAIS's AIUs, this wrapping approach is clearly feasible and may result in an acceptable cost/benefit ratio to the OAIS. The ”Layered Model of Information “ presented in Annex E of this document further describes some potentially standard APIs.

With the second service the OAIS provides not just an API, but a full computing environment that allows the Consumer to execute an OAIS preserved application specific to a range of the OAIS's AIUs. In other words, the Consumer can view the AIUs Content Information through the application's transformation and presentation capabilities. For example, there may be a desire to use a particular application that extracts data from an ISO 9660 CD-ROM and presents it as a multi-spectral image. This application runs under a particular operating system, requires a set of control information, requires use of a CD-ROM reading device, and presents the information to driver software for a particular display device. An OAIS may supply such an environment, including the application, when the environment is readily available. However as the OAIS moves to new computing environments, at some point the application will cease to function or will function incorrectly. As described in Section 4.2, it may not be obvious when the application runs but functions incorrectly. To surely identify such a situation, it would be necessary to record its correctly functioning output as data, along with adequate Representation Information and PDI so it could be preserved. This would need to be checked with the results obtained after moving to a new environment. This may be quite difficult if the application has many different modes of operation. Further, if the application's output is primarily sent to a display device, recording this stream does not guarantee that the display looks the same in the new environment and therefore the combination of application and environment may no longer be giving fully correct information to the Consumer.

The OAIS response to preserving an application execution service would likely depend on whether or not it had the source code for the application. If the OAIS had the source code and adequate documentation on the application, the expected approach would be to port the application to the new environment and attempt to test it adequately to ensure it was functioning correctly. Ideally all possible output values would have been recorded initially so they could be used as the basis for ensuring correct functioning following the port. However if the application applies to a narrow range of the OAIS holdings, this level of testing is likely to result in an unacceptable cost/benefit ratio for the OAIS.

Where the OAIS does not have the source code and must rely on an executable, the techniques described above could not be employed. The OAIS could consider emulating the original hardware platform or reverse engineering the application.

If the application provides a well-defined API for further access, the API could be adequately documented and tested to attempt a reverse engineering of the code. However if the Consumer interface is primarily one of display or other devices which affect human senses (e.g., sound), reverse engineering becomes nearly impossible. A recent report recommended emulation (see Ref. 5) as the key approach to digital information preservation. One advantage of hardware emulation is the claim that once a hardware platform is emulated

successfully all operating systems and applications that ran on the original platform can be run without modification on the new platform. However this does not take into account dependencies on input/output devices. Emulation has been used successfully when a very popular operating system is to be run on a hardware system for which it was not designed, such as running a version of Windows ™ on an Apple ™ machine. However even when strong market forces encourage this approach, not all applications will necessarily run correctly under the emulated environment. For example, it may not be possible to fully simulate all of the old hardware dependencies and timings, because of the constraints of the new hardware environment. Even determining fully the characteristics of such a complex environment, to determine what is to be emulated, may not be possible in practice. Further, when the application presents information to a human interface, determining that the information is still being presented correctly by some new device is problematical. Given these constraints, and the apparent lack of strong market forces requiring such an approach, hardware emulation appears to be both a major technical and economic risk.

The previous discussion implies several guidelines to assist an OAIS plan for application access. These guidelines include:

- Require source code and adequate documentation of key access applications
- Do not attempt to preserve access via an application for which you only have an executable for the application
- APIs should be well-defined and well-documented
- Do not use proprietary and privately held formats or applications to provide access to holdings which are to preserved for long periods of time

6 ARCHIVE INTEROPERABILITY

Users of multiple OAIS archives may have reasons to wish for some uniformity or cooperation among them. For example, consumers of several OAIS archives may wish to have

- common finding aids to aid in locating information over several OAIS archives,
- a common Package Descriptor schema for access
- a common DIP schema for dissemination, or
- a single global access site.

Producers may wish to have

- a common SIP schema for submission to different archives, or
- a single depository for all their products.

Managers may wish to have means for:

- cost reduction through sharing of expensive hardware
- increasing the uniformity and quality of interactions with several OAIS's.

Therefore, it may be advantageous for OAIS archives to cooperate to meet these wishes. The motivation might come from the archives themselves, or an authority that has some influence over them may impose it. In the former case, the archive might be motivated by the desire to keep consumers happy with their products, or to keep users happy with their quality of service, or simply by the need to compete with other archives in order to survive or grow. Situations like this can and have motivated agreements without the need for any explicit federation establishing an external authority. However, in cases where explicit federation is established, the external authority is represented in this Reference Model by Management.

The purpose of this chapter is to explore the degree of interaction and cooperation among archives. Section 6.1 focuses on technical levels interaction, while Section 6.2 discusses management issues concerning the tension between cooperation and autonomy.

6.1 TECHNICAL LEVELS OF INTERACTION BETWEEN OAIS ARCHIVES

OAIS associations can be categorized technically by both external and internal factors. External factors include characteristics of the Producer and Consumer communities. Internal factors could include common implementations of the information models presented in Section 4.2, or multi-archive sharing of one or more of the functional areas presented in Section 4.1.

This section defines four categories of archive association. The first three categories have successively higher degrees of interaction:

- *Independent* - A Local community with no management interaction and no knowledge by one archive of Standards implemented at another.

- *Cooperating* - A Local community with potential Global producers, common submission standards, and common dissemination standards, but no common access. One archive may make subscription requests for key data at the cooperating archive.
- *Federated* - Includes both Local and Global communities, and has both Local and Global access. The Local community has priority over the global community. Optionally, Global dissemination and Ingest are options.
- *Shared resources* - A Local community in which Management has entered into agreements with other archives to share resources, perhaps to reduce cost. This requires various standards internal to the archive (such as ingest-storage and access-storage interface standards), but does not alter the community's view of the archive.

The remainder of this section gives a more detailed view of these categories of association.

6.1.1 INDEPENDENT ARCHIVES

An independent archive is assumed to serve only a single designated community. The archive and the designated community must agree on the design the design of SIPs, DIPs, and finding aids. An independent archive may choose to design these structures based on formal or de-facto standards, which would allow cooperation with other archives that implement the same standards. However, the design decisions to use these standards are not based on the possibility of inter-operation with other archives, but rather on local requirements and cost savings.

The classification of an archive as independent is not based on its size or distributed functionality. An independent archive may occupy one site, or may be physically distributed over many sites. It may use many standards for a given internal element. However, if there is no concept of interaction with other archives, the archive is independent.

6.1.2 COOPERATING ARCHIVES

Cooperating archives are based on standards agreements among two or more archives. The simplest form of cooperation between archives is when one archive acts as a consumer of material from another archive. In this case the consuming archive must support the DIP format of the producing archive as a SIP format. Cooperating archives have related communities of interest, so they order and ingest data from other cooperating archives and possibly have common data producers. No common access, submission and dissemination standards are assumed. The only requirement for this architecture is that the cooperating groups support at least one common SIP and DIP format for inter-archive requests. The control mechanism for this sort of inter-operation can be subscription requests at each archive.

The following figures illustrate the concept of cooperating archives.

At a rudimentary level of archive interaction, Figure 6-1 represents a simple mutual information exchange agreement between archives. (Note: In this and the following figures,

the OAIS is represented as a “five-port device” following the arrangement of Figure 4-1. In each case, a two-archive federation is shown for simplicity, although the concept can be extended indefinitely.) The essential requirement for this federation is a set of mutual Submission Agreements, subscriptions, and user interface standards to allow DIPs from one archive to be ingested as SIPs by another. Therefore, it assumes that some pair-wise compatibility has between established between the archives. This does not necessarily require common access, dissemination and submission methods for all participants, although that might encourage more exchange. This level of agreement would also be useful when the holdings of one archive was consolidated/transferred into another archive due to Management issues.

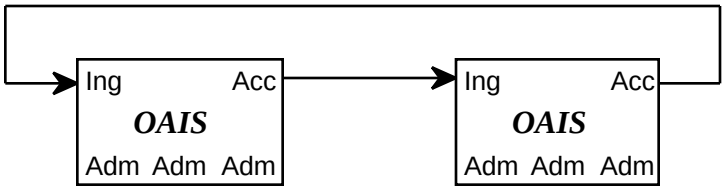


Figure 6-1. Cooperating Archives with Mutual Exchange Agreement

Figure 6-2 is an example of OAIS archives that have standardized their submission and dissemination methods for the benefit of users. No special external element is needed for this. Its disadvantage is that there is no formal mechanism for exchange of Description Information so Consumer must have separate Search Sessions to locate AIPs of interest.

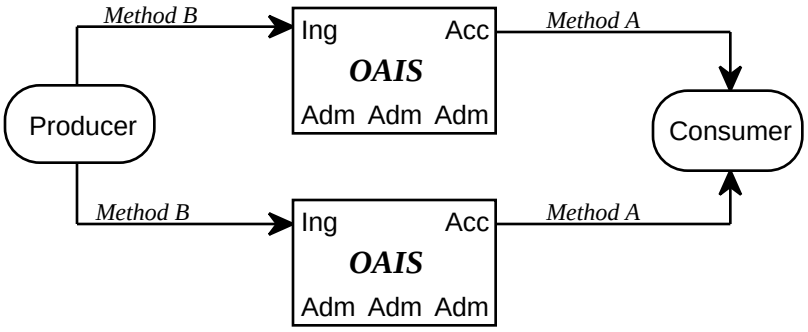


Figure 6-2. Cooperating Archives with Standard Ingest and Access Methods

6.1.3 FEDERATED ARCHIVES

Federated Archives are conceptually consumer-oriented. In addition to the local designated community of an archive, a Global community exists which has requirements for access to the holdings of the archive. However, the Local consumers are likely to have priority over the global consumers.

At the federated level of association, external elements can be introduced to improve inter-

operability. For example, Figure 6-3 illustrates a functional architecture to solve the Access problem described in Section 6.1.2, using an entity external to the Federated OAIS's. Here, two OAIS archives that have similar Designated Communities have decided to Federate to allow Consumers to locate Information Packages of interest from either OAIS with a single Search Session. The Common Catalog & Manager is the external (global) binding element that serves as a common access point for the information in both archives. DIPs containing the finding aids from each OAIS are ingest into the Common Catalog as is shown by the dotted lines in Figure 6-3. The Common Catalog may limit its activity to being a finding aid or it may include common dissemination of products from either or both OAIS's as shown by the dashed lines in the figure.

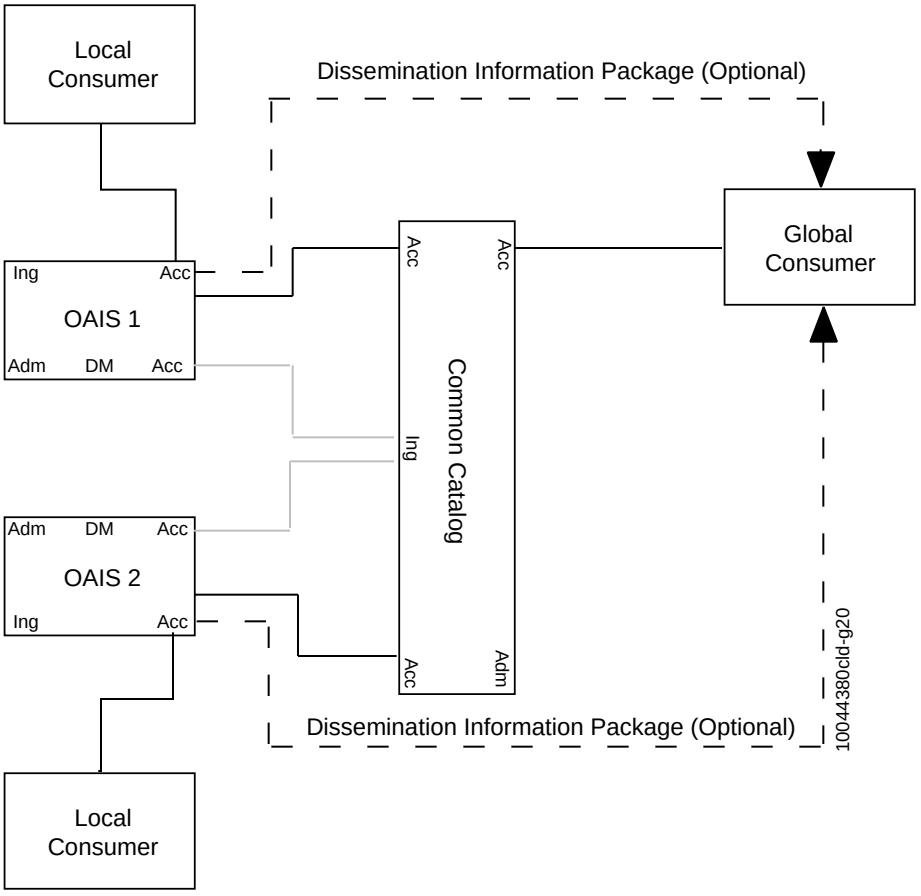


Figure 6-3. An OAIS Federation Employing a Common Catalog

Federated archives may be further classified into three levels of functionality.

· *Federated 1* - Global access is accomplished by the export of a standard-format Associated Description to a global site. The global site independently manages a set of descriptors from many archives and has finding aids to locate which archive owns a

collection of interest. An example of this level of federation is Yahoo on the World Wide Web. The consumer is given a combined view of the holdings of multiple sites, which is maintained centrally at the Yahoo site. To view details of the documents, the user must access the site that contains the actual document. This is made easier because most WWW sites and clients support a standard set of protocols.

- *Federated 2* - Global access is accomplished by having a global node that can distribute a query to multiple local archives. This means that the local Data Management entity must store an additional Associated Description in the global format or have a translator from the global queries to local queries. An option in this case is to establish a common DIP format to ease the load on consumers who may be ordering products from many archives. Examples of this are early versions of IMS V0 or CEOS IMS.

- *Federated 3* - This adds to the functionality of Federated 2 a standard ordering and dissemination mechanism, available through the global nodes. This is a fully functional, modern, federated system. Here, the global system may influence the Associated Descriptor schema designs in each local archive; it would be optimal to build new local archives based on the global schemas and finding aids to ensure high degrees of inter-operability. An example of this level of inter-operability is the ECS and ICS in the earth observation domain.

There are several major policy/technology issues that must be addressed when an OAIS joins a Federation or several independent OAIS decide to create a Federation. These issues include:

- **Unique AIP Names** for each AIP in the Federation. A responsibility of an OAIS is to uniquely identify each AIP it owns. When an OAIS joins a Federation, there is no assurance that some its current OAIS AIP identifiers are not already used by other members of the Federation. An example of a general solution to this problem is to form the AIP identifiers in the Federation by assigning a Unique ID for each OAIS in the Federation and concatenating it to each AIP preserved by that OAIS. This OAIS name could be formatted according to a standard that gives the Customer or other Federation members the information needed to establish a connection to the OAIS that contains the AIP Interest. An example of a standard that accomplishes this is the ISO X.500 Directory Services Naming.
- **Duplicate AIPs** in several different OAIS with different AIP names. This problem is caused by the fact the prior to Federation some OAIS will have duplicated Content Information from AIPs in other OAIS's to enable local user access. In this case a Global Consumer will see all the copies as separate, uniquely identified AIPs. Detailed examination of the PDI associated with the Content Information should allow the Consumer to locate the original, authoritative copy but the search process could be very frustrating to the user. A practical way to handle this is to have a field in the Associated Description of all AIPs that identifies whether they are the original or a copy. This technique is not effective if, prior to federation, two or more archives received the content information from the producer to archive. In this case the federated archive would view these duplicate AIPs as unique, original AIPs.
- **The Preservation of Federation Access to AIPs** when the OAIS owning the AIPs

terminates operations. Unfortunately, many archives will close while their holdings still are of value to the Federation community. The federation should have an agreement for each member OAIS, which states the OAISs the have the responsibility to take over the preservation of a closed OAIS's holdings.

- **User Authentication and Access Management** for global users. If an OAIS has a policy that restricts the access to some of its AIPs or charges for the dissemination of some Information Packages, there is a problem of how to identify and authenticate the user who is making requests through the central node. Each OAIS will have some implemented an Authentication and Access Management system for its Local Users and the infrastructure for this function will have to be extended to include Global Users. Some examples of techniques used for this in current systems are :
  - Default priorities where all members from Global Node share a common set of access constraints and the Global Node handles all the authentication to verify the Consumer as a legitimate user of the Global node. The authentication at member OAIS is done assuming that all requests from the global node are from a single user.
  - User Credential passing where the specific remote user can be authenticated by any of the Federation OAIS and the global node simply acts as an intermediary to carry the authentication dialog (This is difficult to do securely with existing technologies but X.509 certificates used in the WWW show promise).

There are many factors influencing the decision of which of these techniques should be used by a specific Federation. The major criterion is the granularity of the Access Constraints in the Federation. If there is little private data and no charge for data dissemination, a policy that determines a user's access constraints by the source he uses to discover and order AIPs is very reasonable. This involves little modification to the OAIS Authentication system, simply adding the Global Node as a Consumer. The Global Node will have to include mechanisms to identify Global Users to each of the Federated OAIS Authentication mechanisms.

If there are charges for disseminating archived information or significant private data that needs individual user authentication the proxy techniques will not be sufficient and User Credential passing techniques such as passwords and Certificates must be applied. The technologies to enable these supporting mechanisms are still evolving.

6.1.4 ARCHIVES WITH SHARED FUNCTIONAL AREAS

In this type of association, Management has entered into agreements with archives to share or integrate functional areas. The motive for this may be to share expensive resources such as hierarchical file management system for Archive Storage, peripheral device for Ingest or dissemination of Information Packages or super computers for complicated transformations between SIPs, AIPs or DIPs. This association is fundamentally different from the foregoing examples, in that we can no longer ignore the internal architecture of the archive.

Figure 6-4 illustrates the sharing of a common storage function, consisting of an Archival Storage entity and a Data Management entity, between two archives, OAIS 1 and OAIS 2. The access, dissemination and ingest facilities can be at any of the previously described levels of inter-operability. In fact, each archive can serve totally independent communities as implied in this Figure. However, for the common storage element to succeed, standards are

needed at the internal Ingest-storage and Access-storage interfaces. The CNES Long Term Storage Facility is an existing example of this architecture.

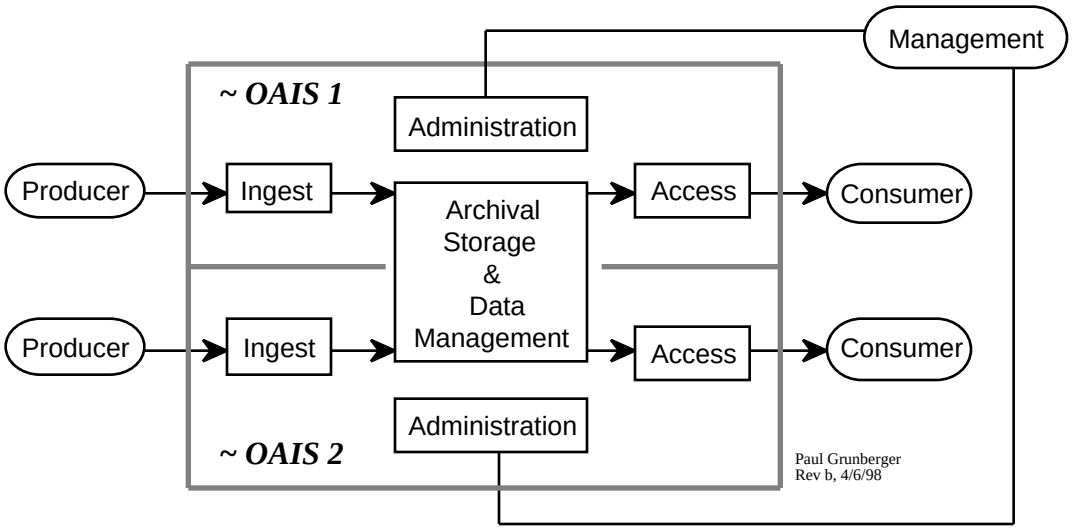


Figure 6-4. Archives with Shared Storage

6.2 MANAGEMENT ISSUES WITH FEDERATED ARCHIVES

The above examples show that the OAIS model is consistent with federation to accomplish specific objectives. However, it should also be considered that some of these objectives might be accomplished through voluntary action. This is an important dimension in association of systems, including archives, because it establishes the degree of autonomy for each system. At the heart of the autonomy issue is the ease with which an association may be altered by one of the participants. Some possible characterizations of autonomy levels might be:

- No interactions and therefore no association
- Associations that maintain your autonomy. You have to do certain things to participate, but you can leave the association without notice or impact to you. An example is participation in the Internet, including operation as a domain name server. You need to meet certain requirements to participate, including the maintenance of a site with certain characteristics, but you will in effect be expunged from the association if you simply stop conforming, However, there is no penalty for this. Therefore, you have maintained full autonomy the whole time because you are free, without penalty, to do what you want.
- Associations that bind you by contract. To change the nature of this association to, meet your desires, you will have to re-negotiate the contract. The amount of autonomy retained depends on how difficult it is to negotiate the changes. The difficulty may rise as more entities become a party to the contract.

So, the autonomy dimension is a key one for interacting archives, determining the ease with

which each can effect changes in the nature of the association and the impact/penalty to each for recovering full autonomy. This dimension is different from the degree of technical homogeneity the association implements or supports, but it is not totally independent. For example, a high degree of technical homogeneity can be achieved in a broad association where each participating entity is free to leave without penalty. However, the permanence of such an association can not be guaranteed, and may be improved by making it more difficult to re-negotiate the association, or by establishing penalties for recovering full autonomy. Also, a given degree of technical homogeneity might be achieved more rapidly and at lower cost when the contract is more binding.

ANNEX A. SCENARIOS OF EXISTING ARCHIVES

This annex is not part of the standard.

A.1 PLANETARY DATA SYSTEM ARCHIVE

I. DOMAIN

Domain and Consumers. The Planetary Data System is chartered to provide data archiving services, data access and expert help to the NASA-funded planetary science community. The PDS is a distributed system with a Central node at the Jet Propulsion Laboratory and discipline nodes (imaging, geosciences, atmospheres, planetary plasma interactions, small bodies, rings) located at universities around the country. The early focus has been on restoring historical mission data and has produced several hundred CD-ROM volumes containing about 80 per cent of the important planetary data archives. There has been an increased emphasis on providing access to the general public for educational outreach over the past several years.

Data Producers. Planetary data sets originate with NASA flight project data management and science teams (new data, some restorations), individual scientists (newly processed or value-added data) or via the PDS discipline nodes (restorations and value-added data). At least 50 per cent of the PDS resources have been devoted to restorations over the past seven years, with several more years of work needed to capture all historical data.

II. INGEST PROCESS AND INGEST INTERFACE

The PDS has developed a very formal interface with the major data producers (flight projects). This interface is documented in the Data Preparation Workbook and involves substantial interaction between node personnel, data engineers and project representatives. A Project Data Management Plan, signed by the PDS project manager provides the basic project data description and agreement to deliver to PDS. Since about 1993 all NASA announcements for Planetary investigations or analysis require that all data generated be delivered to PDS in conformance with PDS standards.

Submission Agreements

Projects provide a Project Data Management Plan. Sometimes a more specific document, the Archive and Transfer Plan, supplements the PDMP, providing extended product documentation and a schedule of deliveries.

Individual scientists can propose to be "data nodes" and receive funds from a PDS discipline node for preparing restored or value- added data sets for inclusion in the archive. There is no formal submission agreement for data nodes.

The PDS discipline nodes each maintain a list of outstanding restorations. These are worked-off based on their priority within discipline. At some point this list will be completed and

only new project or data node data sets will be ingested into PDS. There is no formal agreement associated with discipline data restorations.

Each data set that is identified for ingest in PDS is assigned to a Central node data engineer. It is the responsibility of the data engineer to see that all archiving steps are completed. The archiving steps are called out in the PDS Data Preparation Workbook.

Typical Data Delivery Session. Typically a delivery session will consist of a single data set contained on one or more volumes of CD-ROM or CD-Recordable media. A data set is defined within PDS to be a group of homogenous data granules at the same data level (raw, decalibrated, reduced) which differ only in time of acquisition and major category of target body. For example, the images of Jupiter taken by both Voyager spacecraft comprise a single data set. The standard process includes up-front negotiations between PDS and the provider; the production of test products which are evaluated in the peer review; revised final test products which are validated by the data engineering staff at the central node; approval and production of CD-ROM volumes; distribution by the appropriated discipline node or the central node; entry of the data set into the PDS central catalog; and entry of the data set into the NSSDC ordering system.

Transformation Process. In most cases the original data formats are maintained when data is brought into PDS. This allows existing software tools to continue to be used with the data. Much of the data preparation involves carefully documenting the data format and preparing metadata (granule labels, index files and catalog templates).

Validation. Validation is generally performed as part of the peer review of a product or by using validation tools. In some cases (for example, Magellan), the project develops its own internal validation process. The main validation tool of the PDS is the Volume Verifier. This program is run by the Central Node data engineers on each product delivered from a project or a data restoration. It validates the format and content of all product labels, and validates data files using checksums.

Security. The only area where any special security issues exist involves the receipt of proprietary data. Some projects have one-year proprietary periods before data in released to the science community. The PDS policy is to avoid receipt of any proprietary data sets during the proprietary period.

III. INTERNAL FORMS

The PDS has developed standards for documenting data sets (templates) and individual data products (PDS labels) using a keyword=value label system called the Object Description Language (ODL). Recommendations are also provided for volume organization and data product formatting to optimize the utility of resulting data products.

The PDS standards are specified in the PDS Standards Document. Standard documentation requirements include templates describing the data set, instrument, mission, etc. These templates are included on data volumes and also entered in the PDS high-level catalog.

Standard terminology is maintained in the Planetary Science Data Dictionary, which is jointly maintained by the PDS and the multi-mission ground data system. The metadata values for new data products are carefully compared with the PSDD and existing values used wherever possible. Additions are made to the PSDD to add new standard values to accommodate new data sets and when justified new keywords are added to the PSDD. Data products can have specialized metadata values which are not cataloged in the PSDD.

The PDS product labeling system is flexible enough to allow nearly any data structure to be described. Labels can be attached to the beginning of the data file or detached in a stand-alone text file which points to the data file. In some cases a single label file is used to describe multiple data files. Detached labels can be used to describe data stored in other formats (FITS or HDF, for example). In cases where complicated raw telemetry formats are stored the Software Interface Specification (SIS) for the product is included in lieu of descriptive labels.

Archive Volume Components

An archive quality data set is required to contain the following components.

AAREADME.TXT	- Text summary of data contents.
VOLDESC.SFD	- Standard volume label.
VOLINFO.TXT	- Text description of data contents.
CATALOG	- DATASET.CAT, MISSION.CAT, INST.CAT
INDEX	- ASCII index for each granule on the volume.
SOFTWARE	- Software needed to interpret/display the data.
CALIB	- Calibration data sets.
BROWSE	- Browse products for this volume.

Peer Review. All restoration and data node produced data sets are required to undergo a peer review before acceptance as archive products. Products produced by flight projects do not go through a formal peer review process. In general there is ongoing negotiation between the data engineer or the discipline node staff and the data producer. The peer review team consists of a number of scientists familiar with the data set, the discipline node leader and one or more data engineers. All product documentation and sample products and software are supplied to the peer review group for evaluation. The peer review group determines the adequacy of documentation and quality of the data products and either approves the product or provides a set of liens which must be fixed prior to approval. The PDS nodes and data engineers have access to a Volume Verifier tool which aids in validating the quality of an archive volume. The volume verifier checks internal checksums, verifies that the index contains entries for all data products and validates the volume templates as well as the descriptive keywords supplied for each product.

Delivery Media. Discipline restorations and data node products are recorded on CD- ROM or CD-recordable media as a standard practice. Flight projects are urged to provide archive quality products on CD media but may not be able to due to funding constraints. Products

delivered to PDS on magnetic tape media are assigned to the PDS restoration queue. It is the goal of PDS to convert all data sets to CD-ROM or CD-recordable media which is replicated at a separate geographic facility. This separate facility is generally the National Space Science Data Center (NSSDC) at Goddard Space Flight Center.

IV. ACCESS

Nearly all access to PDS data sets is via the CD-ROM volumes which are distributed to the entire research community. Large discipline node data collections including a substantial volume of CD-ROM data are accessible via the Internet. Several of the discipline nodes have developed on-line retrieval systems customized to meet the needs of their discipline scientists.

Finding Aids. The Pilot PDS devoted substantial resources to designing a central catalog system and distributed query and processing capabilities at the discipline nodes. These efforts were largely dropped as the Planetary Data System focused on data restoration rather than data access. In general, most of the user community already had home grown tools for data analysis and were most concerned with getting access to the data sets. The growth of the user community due to Internet and increased usage of CD-ROM readers has spurred to prototype a more consistent finding aid. The PDS Navigator has been developed for selecting images from the Clementine mission. It includes three components, a forms-based traditional database retrieval capability, an image-based retrieval and a text-based retrieval.

Security. The high-level PDS catalog can be accessed via a group account. Most of the data access services at the discipline nodes require the user to obtain a valid account on the node computer.

Customer Service/Support

The order function of the PDS is distributed. Data inventories are kept at NSSDC, the PDS central node and at each discipline node. In general each site serves a special group of users:

- discipline node - members of the NASA funded discipline
- central node - other NASA scientists and engineers, other agencies
- NSSDC - other scientists, agencies, public and foreign users.

The PDS Operator at the central node handles requests for PDS documentation or standard data products. The discipline nodes handle data requests from within their discipline and also provide expert help in the utilization and interpretation of the data. Access to tools is also provided.

V. DISSEMINATION

The vast majority of data dissemination is done via CD-ROM disc. Several hundred copies of over 500 titles have been distributed to date.

Subscriptions. Nearly all PDS distribution is done via subscriptions or standing distribution lists. It is the responsibility of each discipline node to maintain a distribution list for its discipline scientists. This list determines the order amounts for most CD-ROM titles. The central node maintains a distribution list for engineering and management personnel and for other external recipients (reciprocal distribution, software developers).

Media/NetworkUse. Nearly all final products are delivered to the user community on CD-ROM. Archival products that need not be widely distributed are stored on CD-Recordable media, with a duplicate copy provided to the NSSDC. Most PDS data is available for downloading via anonymous ftp connection to a large CD-ROM jukeboxes at the central node and the imaging node.

Data Manipulation. Each discipline has a suite of government developed analysis tools which can be applied to the discipline data sets. These software packages are available for UNIX workstations or VAX VMS platforms. Several nodes provide the user a menu of processing functions that can be performed on selected data and will carry out requested processing and provide the results electronically or via media. The most widely used commercial tool is IDL.

Pricing Policy. The PDS distributes data to legitimate NASA researchers for no charge. There are no charges for on-line computer usage or data processing to NASA researchers. The NSSDC distributes CD-ROMs for $10 per volume.

Security. All PDS data sets are certified GTDA by the Department of Commerce and are distributable worldwide.

VI. SPECIAL CHARACTERISTICS

PDS has invested a substantial engineering effort in its common data dictionary, data standards and procedures for preparing archival quality data sets. By having these standards in place the PDS is able to demand better quality data sets of its data providers.

A.2 NATIONAL ARCHIVES AND RECORDS ADMINISTRATION'S CENTER FOR ELECTRONIC RECORDS

I. DOMAIN

Domain and Consumers

The Center for Electronic Records is the organization within the U. S. National Archives and Records Administration (NARA) that appraises, accessions, preserves, and provides access to federal records in a format designed for computer processing. NARA serves as the archives for the records of the United States federal government. Consumers for this data are as diverse as the electronic records they seek to access and range from individuals seeking to assert their rights to other government agencies to academic researchers, private consultants,

media personnel, and a wide variety of other users.

Data Producers

Originally this data is produced (created or received) by agencies of the U.S. federal government (producers). The data may concern virtually any area or subject in which the government is involved. They may come from a variety of computer application such as data processing, word processing, computer modeling, or geographic information systems. They can include records made directly by government employees or indirectly through government grants and contracts.

Special Features

The most noted special feature of NARA's Center for Electronic Records is the diversity of the collection of more than 29,000 data sets from more than 100 bureaus, departments, and other components of executive branch agencies and their contractors and from the Congress, the Courts, the Executive Office of the President, and numerous Presidential commissions. A small portion of the data originally were created as early as World War II. An even smaller portion contains information from the nineteenth century that has been converted to an electronic format. Most of the data, however, has been created since the 1960s. The major types of holdings and subject areas include agricultural data, attitudinal data, demographic data, economic and financial statistics, education data, environmental data, health and social services data, international data, and military data.

Scientific and technological data already transferred to the Center include the National Register of Scientific and Technical Personnel; the National Engineers Register; the 1971 Survey of Scientists and Engineers; major portions of the National Ocean Survey's Nautical Chart Data Base; numerous Environmental Protection Agency series relating to pesticide use, hazardous wastes, and pollution abatement; the Nuclear Regulatory Commission's Radiation Exposure Information Reporting System; biometric data sets and epiemiological studies (such as the National Collaborative Perinatal Project) from the National Institutes of Health, the Centers for Disease Control, and the National Center for Health Statistics; and text from presidential commissions on Three Mile Island, coal, and the Space Shuttle Challenger Accident. While the Center's scientific and medical holdings are rich and varied they do not fully reflect the extent and diversity of federal activity in this area.

II. INGEST

The ingest process begins with producers (records managers and records creators in federal agencies) inventorying all electronic records and determining how long to retain the records for current agency business. The next step in the process is for the producer and NARA to develop a *Request for Records Disposition Authority*, Standard Form 115 (SF 115). Here information on the content, retention and disposition, and the availability and extent of documentation and related reports is listed in the context of the producer's business needs for the information. Data with continuing value are listed as permanent and the timing and frequency of their transfer to NARA is established. The producer submits the SF 115 to

NARA for its review and appraisal. The Center for Electronic Records appraises electronic records items on all SF 115s. Identifying permanently valuable electronic records for retention by NARA's Center for Electronic Records involves cooperation between NARA and the producers. Through the process of scheduling and appraisal, the Center identifies and selects the electronic records it judges to have enduring value. The Center evaluates electronic records in terms of their evidential, legal, and informational value and their long-term research potential. Some of the factors in this appraisal evaluation include estimation of past, present, and probable future research value within the context of the data's origin and current use and its impact on federal programs and policy. Administrative and legal value, as well as the potential for linkage with other data, may bear on the decision. Unaggregated microlevel data sometimes has the greatest potential for future secondary analysis. Once the Center determines the records have enduring value, it then determines whether the records should be preserved in electronic format.

Submission Agreements

The actual Submission Information Package (SIP) between NARA and the agency that creates or receives the data is a *Request to Transfer, Approval, and Receipt of Records to the National Archives of the United States*, Standard Form 258 (SF 258) accompanied by the data object(s) and sufficient documentation and descriptive information to use the data. The SIP transfers physical and legal custody of the electronic records from the producer to NARA. This agreement is the end product of the ingest process described above. The SF 258 also contains any restrictions on access to the data which conform with exemptions listed in the U.S. Freedom of Information Act. The Center enforces all legitimate restrictions on access. The Center also works with the producer to determine if any "disclosure-free" version of the data can be produced for consumers.

Typical Delivery Session

This inventorying, scheduling, and appraisal process specifies the data object(s) and related metadata and documentation to be transferred and establishes the timing and frequency of submissions. Specific instructions for how the data are to be organized and when they should be submitted are established in the *Code of Federal Regulations* (36 CFR 1228.188). All data should be transferred on either open reel magnetic tape, tape cartridges, or CD-ROM. The CFR sets the specific technical requirements in terms of format, block size, and extraneous characters. While the current regulations also require that all SIPs should be transferred in a software-independent format, NARA staff recognize that the research potential and utility of some data would be significantly reduced if they were transferred in such a format. In such cases NARA works with the producers to determine the best mode of transfer.

What are the Information Objects that are Delivered? Producers typically will transfer a series consisting of one or more data sets with the related documentation which minimally should include the record layout and codes, methodology statements, technical information about the data including number of records and size. Ideally, the SIP also includes associated analyses and reports. Increasingly agency-created metadata also is included. The majority of

electronic records come as flat files of data; increasingly, however, text files and output from data base management systems, and geographic information systems also are transferred.

What are Collections? NARA organizes all Archival Information Collections (AIC) on the basis of Provenance and Original Order. Provenance maintains the identity of an Archival Information Package (AIP) or an AIC and preserves as much information as possible about its origins and custodial history. Within NARA this is accomplished through the use of Record Groups which reflect the structure of the federal government and subgroups and sub-subgroups which place the AIPs and AICs within the producer's place within its agency. Original order argues maintaining the contents of an AIP or AIC in the order developed and used by the producer. This helps reveal the producer's organization and how it used the data objects and can provide additional information to consumers. For electronic records, "original order" is expressed in the logical structure of files and databases and in the indexing which the producer used. Within NARA the basic unit for arrangement and description is the AIC which can include a number of related AIPs.

What Descriptive Information is Provided? The extent and quality of the descriptive information provided by the producer varies from quite sketchy to extremely detailed. NARA staff attempt to flesh out the producer-created descriptors with AIC level descriptions, title list entries, abstracts, and Dessimination Information Packages (DIP) and to provide the descriptive information in a variety of formats to reach different consumers.

What sorts of Validation Objects are Provided? Producers are required to transfer metadata and descriptors adequate to access, process, and interpret electronic records. For formatted data files the DIP must include a record layout with appropriate field definitions and codes. It frequently also includes methodology statements, input documents, data entry instructions, processing directions, sample outputs, reports and analyses of the information and system manuals.

What Transformation Processes are Performed Prior to Storage

What Metadata is Created? The most extensive metadata product created by NARA is the DIP. In the Introduction, Center staff discuss the origin, creation, and administrative uses of the data object(s), list related objects that are or will be available, and discuss characteristics of the data that could cause problems for consumers based on initial validation processes. The DIP also includes sample printouts of the data and tables and reports related to computer validation of the data. NARA also captures metadata on record layouts, domains, ranges, and links between files in a metadata database as a byproduct of the automated validation process. Other metadata created by Center staff include AIC descriptions, formatted abstracts, title line entries, and collective descriptions which place the data in a broader context. The Center anticipates that increasingly metadata created by the producer will be part of the SIP transferred to NARA.

What Validation is Performed? The Center's initial accessioning procedures include creating a new master and backup copy of each data object on new certified media to ensure the best physical media for long-term storage. At this time Center staff perform automated

comparisons of the data contents with the record layout and codes, and of the physical structure including the number of records, blocks, and bytes. Staff also perfect the DIP to facilitate secondary use of the data.

Security

All data are maintained off-line with consumer access only to copies of the data. The master and backup copies are maintained in separate secure stacks at two different physical locations. Data which require additional security measures, for example Census data subject to restrictions imposed under Title 13 of the *United States Code* and national security classified information restricted under Executive Order, are afforded the appropriate level of protection. The Center is moving to provide enhanced access to selected data onsite by providing reference copies on a wider variety of media and by providing a broader range of services and output products. This may include use of vendors who can provide enhanced access to the holdings utilizing "value-added" services.

III. INTERNAL FORMS

How do you Store your Data? All master and backup copies are stored on newly certified class 3480 magnetic tape cartridges. Some of the holdings have not yet been migrated from nine-track, 6250 bpi open-reel magnetic tape. Data are received and stored temporarily on other media including diskettes, 4mm, 8mm, CD-ROM, and various removable hard drives, although not all of these media conform with regulatory requirements.

Migration (Data). Based on recommendations from the media manufacturers, the National Technology Alliance, the National Institute of Standards and Technology, and various standards organizations, the Center has been migrating its data to new class 3480 magnetic tape cartridge when each media unit is ten-years old.

Migration (Metadata). Metadata has been stored in a variety of formats depending on the original format transferred with the data. Traditionally most metadata existed in textual format. The metadata captured in the validation process is maintained in a relational database. There are no current plans for migrating from this format, although the metadata can be exported in flat file format. The Center has been encouraging data producers to create and transfer metadata in electronic form. Within the next fiscal year the Center hopes to begin scanning and digitally converting metadata so it can be preserved and provided in an electronic format along with the data.

Migration (Format). The *Code of Federal Regulations* requires data producers to transfer all data in ASCII or EBCDIC with all extraneous characters removed from the data except record length indicators or tape marks and blocked at no higher than 32,760 bytes per block for open-reel and 37,871 bytes for class 3480 magnetic tape cartridge. When CD-ROM is used they must conform to ISO 9660 standard and the data must be in discrete files containing only the permanent data. Additional software files and temporary files may be included on the CD-ROM. The CFR also requires all electronic records to be transferred in a software-independent format. The Center works with data producers who cannot meet those

requirements to determine the most appropriate transfer and storage formats.

IV. ACCESS

What Finding Aids are Provided?

Information about the holdings are available in multiple levels of detail and by multiple sources as a way to provide various consumers with information about the Center's holdings. The least specific detail is available in the 1996 three volume *Guide to Federal Records in the National Archives of the United States* where electronic records series are described in the context of the larger holdings from a producer. Other collective descriptions include *Information About Electronic Records in the National Archives for Perspective Researchers*, General Information Leaflet 37, which also is available on the Center's homepage (http://www.nara.gov/nara/electronic), and a title list of data sets available on the Center's homepage and as a printout. Specific electronic records series descriptions were created as formatted metadata for a portion of the Center's holdings for inclusion in a proposed automated description data base which has not been implemented. The most detailed description for any data set is the DIP. Each DIP may contain a narrative describing the data file(s), the record layout and codes for the data, a methodology, sample input forms and questionnaires, annotations regarding the data validity, and a bibliography. The Center also has established an email site (cer@nara.gov) for queries regarding the Center's holdings and services.

Security.

All of the Center's holdings are maintained in environmentally controlled closed stacks which are accessible only by Center staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. The Center's national security classified data sets are in separate environmentally controlled stacks approved for the storage of classified information. All processing is performed in limited access processing rooms at NARA or at the National Institutes of Health computer center. Computer processing is done on closed systems which require both a registered logon and personal identification number or password to access the system. Researchers do not have direct access to any accessioned data. Presently they access copies of the data that they have purchased for their own use.

Customer Service/Support.

The Center has a staff dedicated to providing reference services to the public and to the staffs of other federal agencies. The staff responds to both general and specific inquiries by telephone, letter, email, or in-person visit and fills orders for copies of specific data and their DIPs. The staff also provides information from records to respond to researcher requests such as for casualty records from the Korean and Vietnam conflicts. The staff also functions as a filter between researchers and the data producers when problems develop in understanding or interpreting the data. The staff develop a variety of informational material about the Center's holdings and services, much of which is available online.

V. DISSEMINATION

Do You Support Subscriptions?

The Center will accept a standing order (subscription) for electronic records that it receives on a regular, periodic basis from producers of the Federal government. Under current NARA regulations all subscriptions must be prepaid prior to shipment of the data.

What Media/formats do you use?

Currently the Center provides copies of data files on either nine-track open-reel magnetic tape or class 3480 magnetic tape cartridges encoded in ASCII or EBCDIC, labeled or unlabeled and written to the maximum block size requested. The Center also can provide an exact copy of records in nonstandard formats, if the agency transferred them this way, but it cannot validate or verify the contents of these files. In the past these other formats have included packed decimal, zone-decimal, binary, National Information Processing System (NIPS), Statistical Analysis Software (SAS), Statistical Package for the Social Sciences (SPSS), or OSIRSIS. The Center recently expanded its media options to include diskettes for smaller data sets and CD-ROM. On-line transfer of data remains a more distant goal.

What Transformation (Value Added) is Provided?

The Center currently preserves data as received from the producers; it does not routinely provide extracts from the data or other value-added services beyond computer validation of the data contents and enhanced documentation. Planned enhancements provide for value-added services including extracts from the data.

Pricing Policies.

The Center uses a cost-recovery fee schedule developed by the National Archives Trust Fund. Currently, the charge for an exact copy of all data on a input cartridge or reel, regardless of the number of data sets on the media is $80.75 when copied to a class 3480 magnetic tape cartridge and $90.00 when copied to a nine-track open reel magnetic tape or a CD-ROM. The Center charges an additional $24.50 for each additional data set or file added to a media and $7.50 for each subsequent magnetic tape cartridge and $17.00 for each subsequent open-reel magnetic tape. Paper reproductions cost $0.25 per page.

Security.

The same security considerations developed in relation to Access apply to Dissemination. The Center's national security classified data is made available only to researchers who have both the appropriate security clearances and the appropriate need-to-know. Other restricted data are made available only with prior written approval of the creating agency or under the terms of the restrictions which must be supported as a legitimate exemption under the Freedom of Information Act.

VI. SPECIAL CHARACTERISTICS

NARA's Center for Electronic Records has a diverse collection which reflects the diverse activities of the federal government. The staff shape the holdings through the process of scheduling, appraisal and accessioning. Currently, the Center acquires less than one percent of all federal records created in an electronic format. The timing of the transfer of electronic records from the creating federal agency to NARA is negotiated with the creator to ensure that the records are available for agency use for as long as necessary for current business and that they are transferred to NARA as soon as practicable to ensure their long term preservation for secondary use. NARA is the only federal agency with an explicit archival mandate for Federal records and thus the only Federal agency that preserves and provides access to a wide range of historically valuable records for the indefinite future. As such it is an archives of last resort for the electronic records of some federal agencies which undertake an active data dissemination function while there is a researcher interest in the data but whose mandate ceases or may cease once the demand wanes or ceases.

A.3 LIFE SCIENCES DATA ARCHIVE

I. DOMAIN

What is the domain and who are the customers of the Archive and who are the producers of the data? What are the special features of this archive?

The Life Sciences Data Archive (LSDA) is responsible for collecting and disseminating data of NASA funded Life Sciences space flight investigations. There are two general goals for NASA space life science research; one, to find counter measures to problems encountered by human bodies as a result of space flight, and two, to broaden the understanding of the effect of gravity on living systems. The LSDA's primary customer is the life sciences research community, but it is also used by students, educators and the general public. The data archived in the LSDA is produced by both intramural and extramural investigators funded to perform flight experiments through NASA grants. It is anticipated that the archive may grow to include data from investigations which are completely ground based.

The LSDA is a distributed archive with responsibilities distributed to LSDA Nodes at various NASA Centers and Projects with life sciences activities. There are the LSDA Project Nodes (ARC, KSC, & JSC) which are responsible for the actual collection and cataloging of data, and there is a LSDA Data Distribution Node (Central Node) which is responsible for dissemination of the data to the public.

The LSDA contains animal, plant, and human space flight data. This archive is notable in that it contains a unique collection of data describing, in considerable detail, biology experiments carried out in space by NASA over the past thirty years. The nature of the data is highly varied and spans many life science disciplines.

The LSDA is also unique in that it provides both digital and non-digital information. The

non-digital data may be either reproducible or non-reproducible. Examples of reproducible, non-digital data are video and audio tape. An example of non-reproducible, non-digital data is a biomedical sample.

II. INGEST PROCESS

Submission Agreements

There are two major types of data producers; the NASA Flight Project offices that design hardware and manage the experiment, and the NASA funded Principal Investigator.

To acquire data from the NASA Flight Project offices, the LSDA Project Nodes work closely with them to acquire data during flight operations. The LSDA assists the NASA Flight Project Offices in distributing this data to the Principal Investigators and gathering it as an archival product. As the LSDA is relatively new (1993) there is also retrospective archiving of past missions being done on a funding available basis.

To acquire data fro the NASA funded Principal Investigator, there are a couple of methods of data collection currently being used depending on the "age" of the experiment. For previously flown experiments (prior to 1994) there is an informal submission agreement between the LSDA and the PI's that is based on cooperation, and is not binding. For experiments being selected for flight (after 1994) the funding agreements include a contractual stipulation that the PRINCIPAL INVESTIGATOR must supply the LSDA with raw data, analyzed data and a final science report.

These funding agreements are finalized when proposed investigations are selected for flight. At this time the PIs are sent a letter informing them, that upon acceptance of funding they will be responsible for delivering the data collected as part of their investigation in a form usable by the science~~s~~ community one year post flight.

After a one year proprietary period, submission of data to the LSDA begins. To assist in its submission, the LSDA Project nodes send the PRINCIPAL INVESTIGATOR a Data Inventory package. The PRINCIPAL INVESTIGATOR fills out the data inventory forms and returns them to the LSDA Project Node. The Project Node then contacts the PI to begin data submission. In order to clarify the "usable form" requirement throughout the entire LSDA project, the LSDA is in the process of developing a post flight data reporting handbook which explains exactly how the data should be provided to the archive.

Typical Delivery Session

A typical submission information package (SIP) consists of two parts; 1) the Data Inventory forms, and 2) actual data. The inventory forms are considered CI and contain information about the data types (i.e. physical samples, hardcopy/photographic/video, computer files, other formats) and Data Sets (i.e. title, description, treatments, parameters measured, research subjects and Ids, date/period of collection, collection location, analysis phase, comments). The actual data consists of physical biospecimens, spreadsheets, final science reports,

published articles, procedural documents, photographs, video tapes, analog tapes, digital or printed images, and other types of digital data files (i.e. HRM).

Upon receipt by the LSDA the CI will be cataloged and Archival Supporting Information added, including; experiment and mission ID, Principal Investigator and Co-Investigators name, and other linking information.

Collections

The Archival Information Unit's (AIU) are compiled per investigation, i.e., all AIUs for a single experiment make up an Archive Information Collection (AIC). During a typical ingest session each SIP is cataloged and Archive Supporting Information is added. This CI is entered into a database comprised of LSDA approved fields and uses valid values whenever possible. The Archive Supporting Information is developed by the LSDA personnel at the LSDA Project Node responsible for obtaining the data. The Archive Supporting Information provides layers of metadata for the data collection that describe the experiment, mission, hardware, personnel, sessions, biospecimen, and research subjects from which the data was collected.

It is anticipated that future uses of the archive will involve the creation of AICs based on discipline or measured parameters.

Transformation Processes

In most cases a set of data is kept in its original submitted form. Exceptions to this case include data submitted on outdated media requiring transfer to current media. As little transformation as possible is performed on the data at ingest in order to keep costs down and to insure the integrity of the data. There are some instances where the data has been collected in an application format that is not widely available and in this case the Project Node will transform the data into a more commonly accessible format. (e.g. spreadsheets created in Supernova are translated to MS Excel).

After the LSDA Project Node enters/creates this CI or metadata for the data collection and the individual data elements, the information goes through a validation process. This post-entry validation is accomplished by a second check of the data by the LSDA Project Node Manager. Content validation is further ensured by sending the completed catalog entries to the data originator (Principal Investigator, Flight Project Offices) for verification. The Principal Investigator reviews the information, makes corrections or additions and sends the information back to the Project Node. Edits are then made to the records and the information is once again printed and sent to the Principal Investigator. This process is repeated until the Principal Investigator is satisfied that his experiment data is accurately represented. At this point the Principal Investigator signs and returns a verification letter to the Project Node. The catalog or metadata is now ready for review by the LSDA Project Scientist before it is placed in the public record (via Web Site). The LSDA Project Scientist will review the data for overall form and cogency. Do you want to put anything in here about the LSDA review cycle or is review by the LSDA Project Scientist a sort of catch all term?

Security

The LSDA has strict security measures for data from human subjects which require sensitivity and secure handling due to the Human Data Privacy Act. Human data when received is coded to protect the identity of the crew members. Security procedures include keeping the data on magneto-optical disks stored in a locked file cabinet in a cipher locked room.

Overall security procedures stipulate that all digital data are backed up on a daily basis with off-site storage. Access to on-line servers is controlled through the use of password and/or address port filtering. Only data that is fully validated and approved for release is placed on publicly accessible servers.

III. INTERNAL FORMS

Storage

The LSDA back-up and storage procedures vary between LSDA Node types. Currently, the LSDA Data Distribution Node resides at Johnson Space Center. The LSDA Master Catalog and on-line data reside on a Microsoft SQL Server. These are backed up to tape daily. At the LSDA Project Nodes most of LSDA's data and metadata are stored on magnetic disks and backed up to tape. Long term storage is provided on CD-ROM.
Biospecimens are stored in -80 degree freezers.

AIUs are stored as a piece of CI (a spreadsheet, word processing document, strip chart or biospecimen) with the archival preservation information stored in a database record. The CI can only be, easily, accessed through the database record with it's descriptive information and the CI storage directory information. These AIUs are linked, through the database, into AICs via an Experiment Number. A space life sciences experiment is, in this sense, an AIC. It is a collection of tens or hundreds of AIUs.

Migration

The LSDA migration process is still in a developmental phase but there is some ongoing data migration. LSDA Project Nodes are in the process of converting information on outdated media (RA60's, RL02's) to CD-ROM format.

Migration of application formats (e.g. MS-Excel) and in particular, version changes, is an area of concern. The cost of continually updating all LSDA spreadsheets to the current version is prohibitive and storing and making available the application is also expensive and complicated. A universal read only format such as Adobe Acrobat might be the solution, but ~~it~~ is a proprietary format and it's life span is an unknown.

IV. ACCESS

Finding Aids

Access to LSDA information and data is handled through the LSDA Data Distribution Node at JSC. Users enter the LSDA through the World Wide Web (WWW) and search/retrieve information via the Master Catalog. The Master Catalog is a relational database with a WWW forms interface and allows users to search Archive Supporting Information across experiments and Missions to find data that meets their search criteria. Users can search within ten information groups; Experiments, Missions, Data Sets, Hardware, Documents, Personnel, Specimen or Subjects, Data Collection Sessions, Biospecimens, and Images.

There is currently no method available for searching data at a 'sub-AIU' level. The AIU record contains a considerable amount of detailed, searchable, data so that a collection of AIUs could be found for a particular manual sub-search.

*Security***.** The LSDA does not have any special security concerns for access to the public Master Catalog information and non-human digital data. It is freely available to anyone on the WWW. However, the human flight experiment data is subject to the Human Data Privacy Act, and therefore, security measures are required to control access to this data. The policies and procedures for access to the data are currently being developed.

*Customer Service/Support***.** The LSDA provides user support for questions and problems concerning the Master Catalog (on-line data request system) and for questions about the data being provided. The primary means of user feedback and support is through the LSDA Data Distribution Node. Questions are addressed to the LSDA through on-line "What do you think?" links located throughout the system. From these links a WWW forms interface allows users to submit questions. Specific questions about the data are currently addressed by the NASA Life Sciences Acquisition Scientist and the LSDA Program Scientist. Questions which can not be answered by these individuals are forwarded to the LSDA Project Node which collects the data. In some instances questions are forwarded to the Principal Investigator or NASA Flight Project Office who provided the data.

V. DISSEMINATION

Subscriptions. The LSDA does not support subscriptions since the publicly available Master Catalog is accessible to all users. The LSDA data is located using a catalog on the WWW. Most data is disseminated to the user through links in the catalog to an anonymous FTP site from which the data is downloaded. This means of data dissemination is, therefore, tightly linked to the data "finding" process. If data are in non-digital format, but are reproducible (i.e., hardcopy documents, or log books), users may request them through on-line ordering forms available in the Master Catalog. The requested information is reproduced via photocopying and shipped US Mail to the requester.

There are discussions about an update notification service to be offered in the future.

Media/Formats. The LSDA contains unique non-reproducible pieces of data such as microscope slides and space flight biospecimens. These unique resources are provided to a

requester after a scientific proposal has successfully undergone peer review. Biospecimens, once disseminated, are used to produce original data which is then ingested into the archive.

*Transformations***.** Currently, the LSDA does not provide many value added services. The data is stored and disseminated as provided by the data producer. Data are available in raw and summarized form. These summarized data are provided by the data producer. LSDA does convert data which are received in a non-standard format to a more usable form. Currently there are no data analysis tools available through the LSDA. However, the LSDA Project Nodes do ensure that all data sets have the minimum amount of information needed for understanding (e.g. explanation of all column headings are provided for the spreadsheets, etc.). These are the only "value added" processes that come from the raw data.

Security. Since the Master Catalog and non-human data in the LSDA are available to anyone on the WWW there is no special security in place for the dissemination of this data. However; the human flight experiment data is subject to the Human Data Privacy Act, and therefore is not openly available to users. Limited dissemination of human data will be allowed using policies and procedures that are currently being developed.

Pricing Policies. LSDA data that has been verified and cleared for release is available to the public, free of cost, through the Internet. If significant requests are generated for hardcopy documents, a processing fee for copying the document may be charged. As yet this has not been determined. In the future CD-ROMs with data may be generated. These CDs will be priced in order to recoup production and distribution costs.

A.4 NATIONAL COLLABORATIVE PERINATEL PROJECT (NCPP) 1959-1974

I. DOMAIN

Domain and Customers

The National Collaborative Perinatal Project was a multi-institutional, multi-year study of pregnant women and the children born from those pregnancies to provide baseline information useful for later determining the causes of neurological diseases which appeared in a portion of the studied population. The data came from medical histories, examinations, and observations. The records also contain socioeconomic, family history, and family health information. The data are used by a variety of medical and other researchers.

Data Producers

The predecessor to the U.S. National Institutes of Health's National Institute of Neurological Disorders and Strokes (NINDS) began the National Collaborative Perinatal Project in 1958. Fourteen university-affiliated medical centers across the United States participated in the study. Between 1959 and 1965 each cooperating medical center collected information on between 300 and 2000 pregnancies each year for a total of 55,908 pregnant women utilizing their clinic services. This represented between 14% and 100% of the women utilizing these services depending on the sampling rate employed at each clinic. The final population was

reduced to 39,215 due to miscarriages prior to twenty weeks, 445 multiple births, exclusion of subsequent or repeat pregnancies, and deletion of incomplete records due to women withdrawing from the study prior to its completion. The children were given neonatal examinations and follow-up examinations were through eight years of age. The last examinations were conducted in 1974. The computer data files resulting from the research that NINDS transferred to NARA consist of approximately 6,200,000 records organized into a Master File, a variable file, and eighteen work files, one of which consists of thirteen distinct data files.

Special Features

The Collaborative Perinatal Project was a longitudinal multi-disciplinary research effort which sought to relate the events, conditions, and abnormalities of pregnancy, labor, and delivery to the neurological and mental status of the children of these pregnancies and their siblings through eight years of age. The study sought to link any later appearance of cerebral palsy, mental retardation, learning disorders, congenital malfunctions, minimal brain dysfunction, convulsive disorders, visual abnormality, or communicative disorders to patterns during the perinatal period in order to develop strategies for prevention and intervention. The sample population is large enough so that statistically significant numbers of such disorders would appear in the children. Study of the records relating those children could result in the development of predictive factors and possible preventive care or intervention actions which could reduce future incidence rates.

The data are available in two formats: microfilm of the individual case files for the mother and child of approximately 270 pages per case file and the computer data files. Access to the microfilm and two of the computer data files is restricted because they contain personal identifiers. The National Archives has created a public use file for the Master File and Work File 16: Serum Specimen Inventory.

II. INGEST

The ingest process for transferring any federal agency records to NARA begins with the agency identifying the records and assessing their potential evidential, legal or research value. The next step is for the agency to develop a Standard Form 115, *Request for Records Disposition Authority,* and submit it to NARA. NARA staff then appraise the records in terms of their evidential, legal, and informational value and their long-term research potential. NARA and the creator then establish a transfer date, negotiate any restrictions on access, and initiate the ingest process.

Ingest for the NCPP computer data was a two phase process. In Phase One, from 1958 through 1974, NIH's NINDS funded the project and the cooperating institutions conducted the research. Contractors accumulated the original examination records, created the consolidated case files, microfilmed the records, normalized the data, and developed the Master File, an extract file of frequently used variables, and special files such as "refined diagnoses". The data were stored on 23 reels of magnetic computer tape recorded at 1600 bpi. Prior to 1980 the data were available only to NINDS, the cooperating hospitals, and

selected government researchers.

In Phase Two NINDS developed the documentation necessary for more generalized use of the data and negotiated a submission agreement, including access provisions, with NARA. Since the nonreleasable data could be made anonymous through creation of a Public Use File, the producer and NARA worked on transferring the data files first.

Submission Agreements

NARA and NIH executed the U.S. Government's standard transfer form, Standard Form 258, *Request to Transfer, Approval, and Receipt of Records to the National Archives of the United States*, in mid-1985. This transferred legal custody and preservation responsibility to NARA. A similar agreement for the microfilm examination records was executed in 1990 after NARA and NIH resolved the privacy and access concerns and NARA developed a statistical research form.

Delivery Session

The delivery session was a single transaction in which NIH provided NARA with copies of the 23 reels of magnetic tape containing the NCPP and the related documentation consisting of seven volumes containing the background of the study, the sample, data collection and data processing overviews, record layouts and coding for each variable, sample forms, and a bibliography of all published research through 1985. NINDS transferred the 8000 rolls of microfilm containing the examination records in 1990.

Transformation Process

NARA has maintained and preserved the original data format. The data are in a hardware and software independent EBCDIC format which facilitates wide researcher access. All data were copied to new nine-track open-reel magnetic tape when received in 1985 and are migrated to new media every ten years to ensure long-term preservation. The more than 7000 pages of documentation are available in both a page format and in a microfiche format on 75 fiche. The documentation has not been scanned or digitized.

Validation

Validation was performed as part of quality control throughout the life of the NCPP. Extensive use of the data during the life of the project (1958-1974) and its use by NIH approved researchers (1958-1985) provided a second de facto validation. NARA also validated sample portions of the data at the time of ingest. Continuing researcher use also validates the data contents.

Security

The computer data is maintained in environmentally-controlled closed stacks which are

accessible only to Center staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. NARA has created Public Use Files of the restricted data files to prevent unauthorized access to personal and medical data.

What Descriptive Information is Provided?

NARA has prepared multiple levels of descriptive information for the NCPP. These range from entries for each data file in the Center for Electronic Records' Title List, an abstract entry for the series, a series description, a full documentation package, to a series-level entry in the three-volume *Guide to Federal Records in the National Archives.*

What Validation Objects are Provided?

During its active life the NCPP established and used elaborate data collection, input and verification procedures. Extensive use also validates the information. NARA's routine transfer and storage procedures also validated the data. The extensive seven-volume documentation includes record layouts and codes, methodology statements, data analyses, and a bibliography of research use.

What Transformation Processes are Performed Prior to Storage?

What Metadata is Created? NARA staff supplemented the documentation with an abstract and introduction discussing the origin, creation, and uses of the data, including an explanation of restrictions on access and the characteristics of the Public Use File.
What Validation is Performed? NARA's accessioning and storage procedures included creating a new master and backup copy on new certified magnetic media and creating a Public Use File of the two restricted data files. Sample portions of each data set also were verified against the documentation.

III. INTERNAL FORMS

Storage. NARA maintains separate sets of the master and backup copies of the data and the Public Use Files on newly certified 3480 class magnetic tape cartridges.

Migration (Data). NARA migrated the NCPP from 23 nine-track, 1600 bpi open-reel magnetic tapes it received in 1985 and stored the data on seven nine-track, 6250 bpi open-reel magnetic tape. NARA migrated the data to four 3480 class magnetic cartridge when the media was ten years old.

Migration (Metadata). NCPP metadata is available in textual (7000+ pages) and microfiche (75 fiche) forms. There are no plans to scan or digitize the text.

Migration (Format). The data currently are encoded in EBCDIC with all extraneous characters removed. There are no plans to migrate the format at this time.

IV. ACCESS

What Finding Aids are Provided?

Information (of varying detail) about the NCPP is available in the 1996 three volume *Guide to Federal Records in the National Archives of the United States*, where the records are described in the context of the larger holdings of the National Institutes of Health; in *Information About the Electronic Records in the National Archives for Perspective Researchers*, General Information Leaflet 37; in *Title List: A Preliminary ad Partial Listing of Data Files in the National Archives and Records Administration*; and in the documentation package for NCPP. Much of this information is available on the Center's homepage (http://www.nara.gov/nara/electronic) or by posting an enquiry to the Center's e-mail site (cer@nara.gov).

Security

NCPP, like all of NARA's holdings, is maintained in environmentally-controlled closed stacks which are accessible only by authorized Center staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. Researchers do not have direct access to the data. Presently they acquire copies of the data on a cost-recovery basis permitting indefinite use of the data for their own purposes.

Customer Service/Support

The Center has a staff dedicated to providing reference services to the public and to the staffs of other federal agencies. The reference staff responds to inquiries by telephone, mail correspondence, e-mail, or in-person visits. They fill orders for copies of all or part NCPP and the relevant documentation. The staff also function as a filter between researchers and the NINDS when problems develop in understanding or interpreting the data.

V. DISSEMINATION

Do You Support Subscriptions?

NARA's Trust Fund is willing to establish accounts that allow researchers to acquire data that is transferred on a recurring basis. Since the NCPP stopped collection data in 1974 there is no need for a subscription for this data.

What Media/Format do you use?

Copies of the 32 data sets comprising the NCPP are available on seven nine-track open-reel magnetic tapes, six 3480 class magnetic tape cartridges, or two CD-ROM.

What Transformation (Value Added) is Provided?

The NCPP is provided as received from NINDS. NARA has created Public Use Files for the

two data files containing personal identifiers in conformance with the Freedom of Information Act and NARA restrictions on access to records whose release might result in unwarranted invasion of personal privacy.

Pricing Policies.

Electronic data sets are available on a cost-recovery fee schedule developed by the National Archives Trust Fund. Currently the charge for an exact copy of all NCPP data on a storage reel or 3480 class cartridge is $80.75 when copied to a 3480 class magnetic tape cartridge and $90.00 when copied to a nine-track open-reel magnetic tape. Copies on CD-ROM are $90.00 for the first file and $24.50 for each additional file written to the CD-ROM.. Paper reproductions cost $10.00 for the first 20 pages and $5.00 for each additional block of 20 pages. Microfiche reproductions cost $2.10 per fiche.

VI. SPECIAL CHARACTERISTICS

The National Collaborative Perinatal Project was a prospective study. NINDS expended more than $200 million over two decades to collect information on more than 58,000 pregnant women and their children at fourteen cooperating institutions. It is unlikely that a study of this duration and magnitude will be repeated. The data continue to constitute an important resource for biomedical and behavioral research in many areas of obstetrics, perinatalogy, pediatrics, developmental psychology and other fields.

A.5 ARCHIVE SCENARIO FOR THE *CENTRE DES DONNEES DE LA PHYSIQUE DES PLASMAS* (CDPP)

I. DOMAIN AND CUSTOMERS

The CDPP (*Centre des Données de la Physique des Plasmas* - Center for Data on Plasma Physics) is a new service currently being set up. It has been developed to ensure the Long-term conservation and availability of natural Plasma Physics data (magnetospheric plasma, planetary plasma etc.) for the international scientific community. More specifically, the data concerned is from either ground-based or space-flown experiments in which France has participated or wholly directed. The CDPP is designed around two principal components:

- A Technical Activity segment, located on the premises of the French space agency, CNES, mainly in charge of developing and maintaining the archive system. The latter has the following functions: addition of data and metadata to the system, preservation of data and metadata, organization of search and product ordering facilities, and dissemination.
- A Scientific Activity segment, located at the CESR (*Centre d'Etudes Spatiales des Rayonnements* - Center for the Study of Space Radiation), a science laboratory near CNES. The CESR is in charge of all aspects relating to scientific knowledge of the data: validating data with its producers, ensuring that the data is useable by the scientific community, setting up added-value services etc. This Center is also responsible for developing a WWW server to present CDPP services, supplying educational information

on Plasma Physics to the general public, and guiding users to access and dissemination functions.

The two complementary segments work closely together.

A number of associated laboratories will be able to join the two main components of the CDPP provided they offer a service (data dissemination or information) relating to natural plasma physics.

The archive system is currently being developed. The service is planned to be made available to the scientific community on 1/1/99.

Data Producers. Data producers are mainly either current or future experiments, or projects concerned with rehabilitating existing data. Ongoing experiments include, for example, the French experiments flown aboard Russian satellites (INTERBALL), aboard the US satellite (WIND), aboard the future European satellites (CLUSTER), or even some data from the EISCAT radars. The projects to rehabilitate existing data cover a many French experiments performed since 1975, mostly flown on European, US and Soviet satellites or probes.

II. INGEST

The CDPP has drawn up a specification for deliverable data products. The specification defines the characteristics (either mandatory or optional) that the data and metadata to be delivered to the CDPP must exhibit. It defines the rules systematically applied with respect to:

- file structure, data encoding and standardization of times and dates,
- orbit or trajectory data,
- the minimum content and format of catalogues,
- complementary information needed to use or interpret the data,
- etc.

The CDPP provides technical support in order to apply this specification to each data-producing project.

As far as future projects are concerned, the authorities empowered to make decisions on projects will make the drawing up of an obligatory data management plan. The plan must define exactly which data will be archived (physical values, raw data...), how the data will be organized, and when the data will be delivered to the CDPP.

One particular service within the CDPP is the SPID (*Service de Préparation des Informations et des Données* - Service for Preparing Information and Data), in charge of the interfaces with data-producing projects and the formatting of some metadata before its delivery to the archive system.

Submission Agreements. As far as future projects are concerned, the submission agreement shall be constituted by the project Data Management Plan, to be approved by both the project and the CDPP. As far as existing data to be rehabilitated is concerned, the framework is less formal: there is normally no project team left and no longer a budget specific to that project. Rehabilitation is thus the responsibility of a team of engineers from CNES and those of the Principal Investigator or members of his team. The CDPP suggests priorities for the work to be completed. It also influences the choices and compromises to be made with regard to the level of data to be archived.

Delivery Session

Data delivery. Data-producing projects must normally store the data produced before delivery. They do so using the facilities offered by the STAF, a multi-mission storage service at CNES. The main function of the STAF (*Service de Transfert et d'Archivage des Fichiers* - Service for Transferring and Archiving Files) is the Long-term physical storage of information. The interface is stable and therefore the technologies and storage media can thus be replaced or changed in-house without affecting the interface. The STAF also monitors and renews the media used.

From a user project viewpoint, the STAF appears as a virtual tree structure in which files may be stored. When all the data to be delivered has been produced, the delivery process merely amounts to a change of ownership of the STAF directories in which the data is stored. There is no actual physical movement of data.

Delivery of metadata. Metadata generally takes up less space than data. A delivery disk space is set up by the CDPP and the data-producing project has the right onto write to this space. When all the data and metadata has been delivered, the SPID can begin its checking and formatting (see below). This process is valid for a complete set of data, a partial delivery or an update of previously delivered metadata.

Transformation Process

The format of experiment data is not altered during the delivery process. On the other hand, metadata delivered will be subject to a kind of packing (without changing the contents) and new metadata will be created by the SPID. To give some examples:

The archive system manages the descriptions of both data collections and objects, browse data and documentary information in the form of graphs on collections and objects. The delivery of a new collection results in the creation of a new node in the data description graph and logical links with existing collections. The creation of this information, granting a global and consistent view of all the data and metadata available, is not within the domain of the data producer.

When the Principal Investigator delivers a Microsoft Word document describing an experiment, he places the corresponding file in the delivery disk space. The SPID will then use this file to create a documentary object descriptor giving the document title,

author, publishing body, language, associated keywords, stating the existence of an abstract etc.

The insertion of metadata in the archive system is mostly based on use of PVL (Parameter Value Language) and a DED (Data Entity Dictionary) which is configuration managed. One of the roles of the SPID will thus be to create this new metadata and construct the PVL structure describing it. Generally speaking, metadata appears as an extremely heterogeneous set of information objects. Using PVL means that these heterogeneous objects may be delivered in both a homogeneous and standard format.

Validation

The SPID is responsible for ensuring that the deliverable product specifications for each data set have been respected. It also performs a number of coherence checks, such as checking coherence between catalogue data and the files containing experiment data.

Once these checks have been completed, the results, together with all the metadata, are presented at a formal peer review whose purpose is to decide whether the CDPP can accept the data set and issue recommendations in this field. Once accepted, the CDPP becomes the guarantor of the data set. This review brings in scientists from outside both the CDPP and the Principal Investigator team.

Despite the various checks carried out, the scientific validity of the experiment data delivered to the CDPP remains the responsibility of the Principal Investigator or data-producing project.

Security. The delivery process for both data and metadata takes place within a dedicated environment accessible only by the data producer and the CDPP.

III. INTERNAL FORMS

Storage. The STAF multi-mission storage service (see above) takes charge of the data and metadata. This service currently uses StorageTek silos with 3490 cartridges and larger capacity Reedwood cartridges (10 Gigabytes uncompressed). The objects archived by this service are files. There are several different layers of service with regard to file retrieval time and file duplication. The STAF is in charge of all data migration involved when changing from old to new media or to a new technology medium. They do not affect the upper layers of the system.

Formats. The format of data stored must be independent of all operating systems. In practice, experiment data is usually in IEEE or ASCII code and divided up into sequential files. The application of CCSDS encoding for times and dates is compulsory for all record structure files. The syntax and semantics of each file must be described with EAST and a DED unless self-descriptive structures such as FITS or NCAR are used. As far as documentary information is concerned, no reference standard for the internal representation of documents has yet been applied.

Data Management

Data management revolves around use of a graph describing data collections and objects. For the purposes of simplification, this graph is usually known as a data graph. It is oriented and non-cyclical. The relations associating a node with its descending nodes are (from an object-oriented viewpoint) inheritance and composition relations. A data set, also known as a terminal collection, thus inherits the characteristics of all the collections above it.

Documentary information, browse data and event tables are also managed through graphs which are nonetheless distinct from the data graph. The graphs contain either explicit metadata or references to external files or documents.

IV. ACCESS

Access facilities are seen by the user through a WWW server. These facilities include aids to search for data collections and objects, means of retrieving certain metadata (such as documents and catalogues) immediately and ways of ordering data products which include special protective mechanisms for data not made public.

Finding Aids

The aids to search data of interest to the user are based on navigation within the different graphs: the experiment data collection and object graph, the browse data graph, the documentary object graph and the events table graph. These graphs are independent but a certain number of links are used to move from one to another. Navigation within the graphs is, depending on the case, through criteria such as a keyword (parameter measured, location of measurements etc.), time or other types of criteria.

The data object and collection graph grants several views of the data, and the final objects may be selected after several navigations within the graph.

The events table graph may be used to make indirect selections over time, such as selecting only data corresponding to a given instrument operating mode, or data corresponding to the periods during which a particular type of magnetospheric event was observed etc.

These aids may be used to select data which is stored either on the main archive site (at CNES) or at an associated laboratory.

Security

Without exception, metadata is visible and accessible to the general public without any prior authentication. On the other hand, data may only be ordered by a user previously authorized by the CDPP, as it normally implies the consumption of resources. The user makes his request for authorization by a form available on-line, indicating his name, e-mail address, the name of the laboratory he belongs to and the reasons for his request. Once the user has

received authorization to order products, he must authenticate his request (name and password) before ordering.

Data archived by the CDPP is usually public in nature, but in the case of recent data, data ordering may be temporarily restricted to one particular user group. The system must therefore be capable of handling access rights to the service (for ordering data) independently from access rights to the data itself.

Finally, the system is designed and has a number of protective measures such that any accidental or deliberate modifications to the data stored in the Center may be avoided.

Customer Service/Support

The system can handle profiles peculiar to each user, taking into account in particular the capability of the network linking him to Internet and the laboratory to which he belongs (laboratories directly supported by CNES, laboratories involved in cooperative projects with French laboratories, other laboratories etc.).

The CDPP has a customer support team able to reply to technical questions (how to use the system, read data etc.). This team can also direct the users to the Principal Investigator or data producers.

V. DISSEMINATION

Subscriptions. In its initial version, the system only accepts orders relating to data available in the system.

Media/Network Use

The data from Plasma Physics experiments is often bulky (a data set often contains between ten and several hundred Gigabytes). It is not planned to systematically create pre-defined, widely disseminated products as is often the case for planetary data, particularly as users are often interested in a specific period of time and not the whole data set.

Products may be delivered either over a network or on a variety of media (currently CD-ROM, DAT or Exabytes). The choice between these two types of delivery depends on the capacity of the network between the user and the CDPP at any given time.

As far as network deliveries are concerned, the system proposes the HTTP protocol at the user's initiative or the FTP protocol at the CDPP's initiative, but at a time specified by the user. The latter choice is subject to certain constraints. Deliveries of data via a network offer optional data compression and grouping facilities in the form of .tar files.

Data Transformation

The data objects distributed to scientific users are not necessarily identical to the data objects

stored in the system. Depending on the standards respected and tools available, a certain number of transformations of archived objects may be requested, in particular:

- Time-related retrieval which provides data corresponding to one (or more) time periods specified by the user. This kind of retrieval is only possible when times and dates have been encoded in compliance with CCSDS recommendations.
- Retrieval of fields, which permits the user to select fields of interest on the basis of an EAST data descriptor.

These transformations are known as "subsetting services". Other such transformations are planned for future versions, so as (for example) to be able to deliver data in the user's native machine format, or deliver data as physical values although it is stored as raw values.

Pricing Policy. The pricing policy has not yet been determined, but will probably include an invoice for dissemination of data on an external medium (CD-ROM, DAT, Exabytes).

Security. Whether data is public or not, data products ordered by a user are only visible and accessible by that user, whatever the mode of delivery.

ANNEX B. COMPATIBILITY WITH OTHER STANDARDS

This annex is not part of the standard.

(TBD)

ANNEX C. BRIEF GUIDE TO THE UML

This annex is not part of the standard.

A key to object relationships in the UML diagrams of this document is shown in Figure C1.

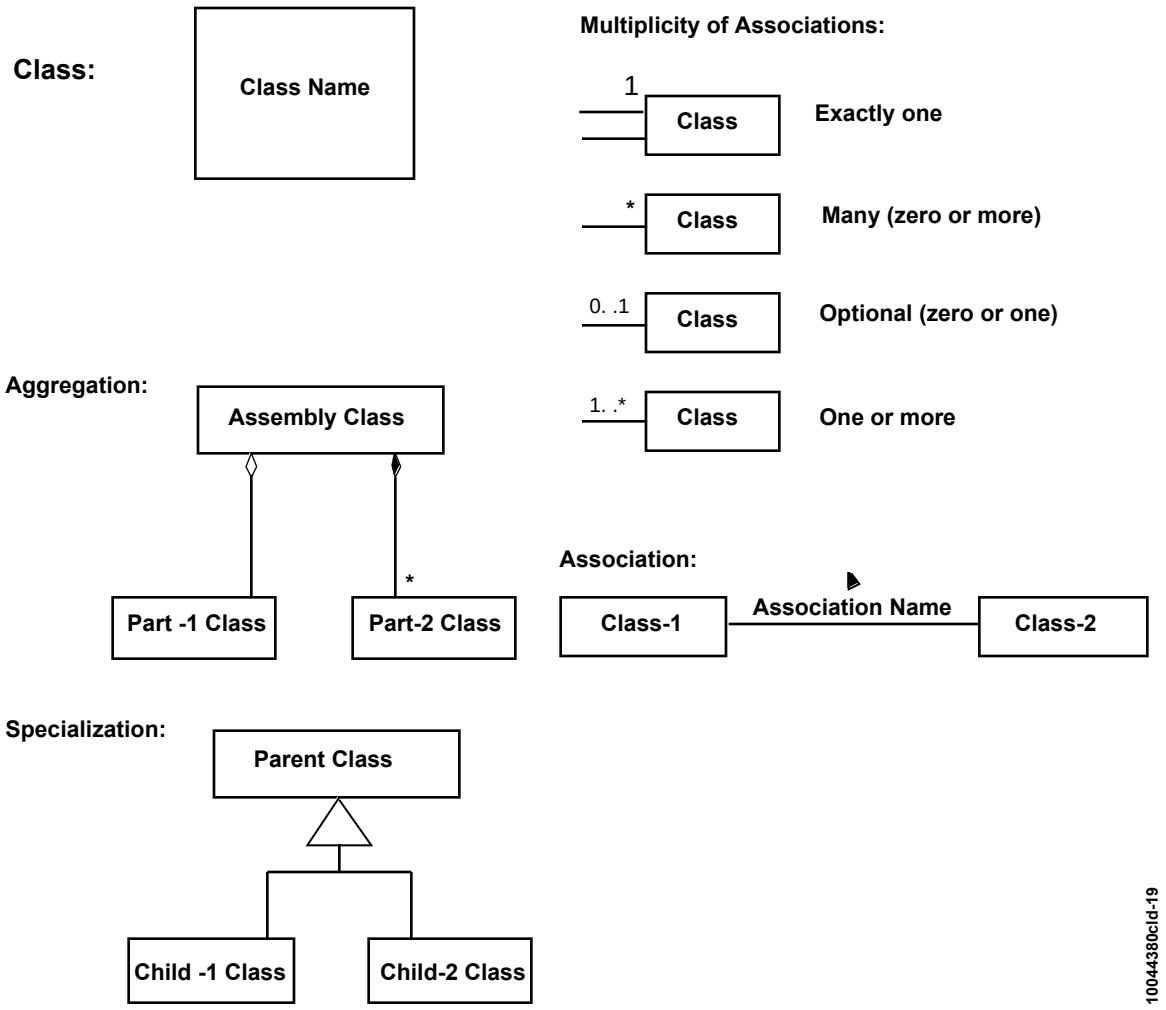


Figure C-1. Key to UML relationships

A Class is indicated by a rectangle containing the Class name. The UML representation of a class is a three-compartment rectangle with name in the top compartment attributes in the second compartment and methods in the lowest compartment. In this document the attributes and operations compartments are always empty and UML states empty compartments can be suppressed.

Classes of objects are related to one another through Associations and there are various multiplicities that may be attached to these associations as shown. The multiplicity refers to the number of instances, or objects, of that class that are involved in the relationship.

A solid line connecting two classes indicates the general association, among two classes. The line is labeled with an association name, indicating the nature of the association, and a solid arrowhead indicating the direction that the relationship should be read. The multiplicity of each class is shown next to the class near the association line. If the association forms a class that may have its own attributes or methods, that association class is shown as a rectangle connected to the solid line by a dashed line. The multiplicity may be omitted if the association is 1-to-1.

There are two particular associations that are commonly used - aggregation and specialization, and these have particular symbols to indicate them.

An Aggregation association is one where a class is considered to be a part of another class. In UML, a diamond connecting the aggregation association to the aggregated class shows association. There are two types of aggregation defined by UML.
Strong aggregation, where the part classes are physically stored as part of the aggregated class, is shown with a solid diamond. In a strong aggregation, if the aggregated class is destroyed, the child classes are also destroyed. aggregation, where the part classes are referred to by the aggregated class, is shown with an empty diamond. In a weak aggregation, if the aggregated class is destroyed, the part classes are not destroyed and may be aggregated into other new classes. Strong aggregation can be thought of as aggregation by value, while weak aggregation can be thought of as aggregation by reference.In the figure above, the aggregation association says that the Assembly class contains exactly one Part-1 class instance and zero or more Part-2 class instances. Also if an instance Assembly is destroyed the Part –1 instance will continue to exist but all the Part-2 instances will be destroyed.

A Specialization association is one where a child class inherits attributes and methods from the parent class. In UML, a broad triangle connecting the aggregation association to the parent class shows specialization. An instance of a child class contains all the attributes and methods contained by its parent class, so an instance of the child class can be used is any operation where an instance of the parent class would be valid. However, the child class may add any number of new attributes or methods so an instance of a parent class is not necessarily a valid replacement for the child class. In the figure above, the specialization association says that the Parent class attributes and methods are inherited by the Child-1 class and the Child-2 class.

ANNEX D. INFORMATIVE REFERENCES

This annex is not part of the standard.

[1] Procedures Manual for the Consultative Committee for Space Data Systems, CCSDS A00.0-Y-6. Yellow Book. Issue 6. Washington, D.C.: CCSDS, May 1994.

[2] "Preserving Digital Information", Report of the Task Force on Archiving of Digital Information, final report currently available from the URL <http://www.rlg.org/ArchTF/>, May 1, 1996

[3] "Unified Modeling Language Specification, Version 1.1", at http://www.rational.com/uml/resources, September 1, 1997.

[4] "Z39.50 Profile for Access to Digital Collections", currently available at the URL <http://vinca.cnidr.org/protocols/profiles/zdl.html>

[5] "Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation", by Rothenberg, J., Council on Library and Information Resources (CLIR) Publications, Report Pub77, currently available from URL http://www.clir.org/pubs/reports/reports.html , January, 1999.

[6] IEEE POSIX Open Systems Environment (OSE) Reference Model (IEEE 1003.0, 1995)

[7] NIST Application Portability Profile (APP, NIST Special Publication 500-xxx, April 1995)

ANNEX E: A MODEL FOR SOFTWARE USE IN REPRESENTATION INFORMATION

This annex is not part of the standard.

Sections 4.2 and 5 discussed that display software is often used to end the Representation Network. A way to view this information is as a layered Information Model as shown in figure E-1. In this model there are five layers of software. Each of these layers has well defined interfaces to the higher layers of the model. These interfaces are known as Application Program Interfaces or Service Access Points in other layered models. The following is an overview of the functionality of each layer and the data that is exchanged at each interface. This overview illustrates the process of getting bits from the media and adding Representation Information needed to make the information usable by the Consumer.

- The Media Layer simply models the fact that the bitstrings are stored on physical or communications media as magnetic domains or as voltages. The function of this layer is too convert that bit representation to the bit representation that can be used in higher level (i.e., 1 and 0). This layer has as single interface, which enable higher layers to specify the location and size of the bitstream of interest and receive the bits as a string of 1 and 0 bits. In modern computing systems device drivers and chips built into the physical storage interface provide much of this functionality.
- The Stream Layer hides the unique characteristics of the transport medium by stripping any artifacts of the storage or transmission process (such as packet formats, block sizes, inter-record gaps, and error-correction codes) and provides the higher levels with a consistent view of data that is independent of its medium. The interface between the Stream Layer and higher layers allows the higher layers to request Data Blocks by name and receive a bit/byte string representing those Data Blocks. The term *name* here means any unique key for locating the data stream of interest. Examples include path names for files or message identifiers for telecommunication messages. In modern computing systems, operating system file systems often provide this layer of functionality.

- The Structure Layer converts the bit/byte streams from the Stream Layer interface into addressable structures of primitive data types that can be recognized and operated by computer processors and operating systems. For any implementation, the structure layer defines the primitive data types and aggregations that are recognized. This usually means at least characters and integer and real numbers. The aggregation types typically supported, include a record (i.e. a structure that can hold more than one data type) and an array (where each element consists of the same data type). Issues relating to the representation of primitive data types are resolved in this layer. The interface from the Structure Layer to higher levels allows the higher levels to request labeled aggregations of primitive data types and receive them in a structured form that may be internally addressable. In modern computing systems programming language compilers and interpreters generally provides this layer of functionality.

- The Object Layer, which converts the labeled aggregates of primitive data types into

information, represented as objects that are recognizable and meaningful in the application domain. In the scientific domain, this includes objects such as images, spectra, and histograms. The object layer adds Semantic meaning to the data treated by the lower layers of the model. Some specific functions of this layer include the following:

- Defines data types based on information content rather than on the representation of those data at the structure layer. For example, many different kinds of objects—images, maps, and tables—can be implemented at the structure level using arrays or records. Within the object layer, images, maps, and tables are recognized and treated as distinct types of information.

- Presents applications with a consistent interface to similar kinds of information objects, regardless of their underlying representations. The interface defines the operations that can be performed on the object, the inputs required for each operation and the output data types from each

- Provides a mechanism to identify the characteristics of objects that are visible to users, operations that may be applied to an object, and the relationships between objects.

The Interface between the Object Layer and the Application Layer allows the higher levels to specify the operation that is to be applied to an object, the parameters needed for that operation and the form in which results of the operations will be returned in. One special interface allows the user to discover the semantics of the objects such as operations available, and relationships to other objects. In modern computing systems subroutine libraries or object repositories and interfaces supply this functionality.

- The Application Layer contains customized programs to analyze the Data Objects and present the analysis or the data object in a form that a Data Consumer can understand. In modern computing systems application programs supply this functionality.

The problem with using software to end Representation Networks is that the programs that are saved do not include the information needed to enable the lower levels of the layered model to extract the information from the bits on the media. These services are usually provided by the vendor-supplied operating systems, device drivers, and file systems. When data is moved to other media or different software platforms, the interfaces to these levels may be changed. This migration process is further discussed in Section 5 of this document.

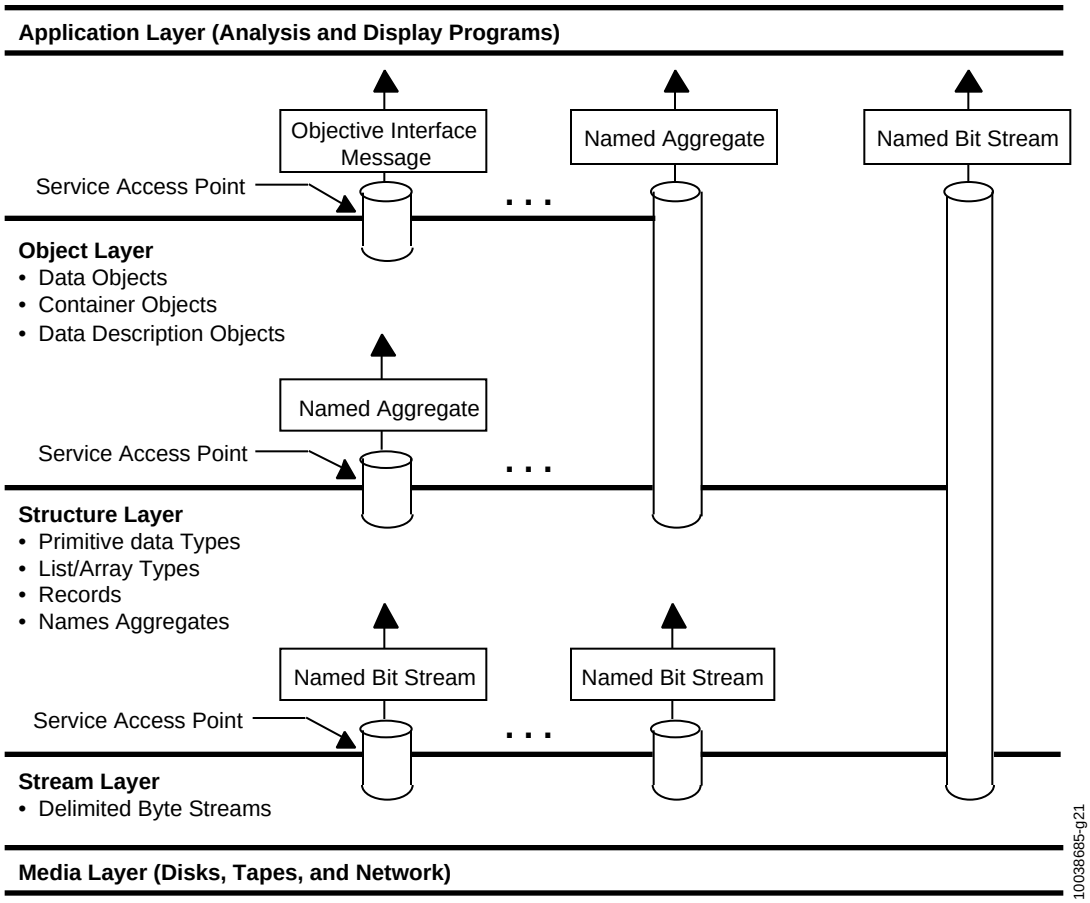


Figure E-1: Layered Information Model

*** *[End of document]* ***