Consultative Committee for Space Data Systems

RECOMMENDATION FOR SPACE DATA SYSTEM PRACTICES

Producer-Archive Interface Methodology Abstract Standard

CCSDS 651.0-M-1

MAGENTA BOOK

May 2004

AUTHORITY

Issue:	Magenta Book, Issue 1
Date:	May 2004
Location:	St. Hubert, Canada

This document has been approved for publication by the Management Council of the Consultative Committee for Space Data Systems (CCSDS) and reflects the consensus of technical panel experts from CCSDS Member Agencies. The procedure for review and authorization of CCSDS Reports is detailed in the *Procedures Manual for the Consultative Committee for Space Data Systems*.

This document is published and maintained by:

CCSDS Secretariat
Space Communications and Navigation Office, 7L70
Space Operations Mission Directorate
NASA Headquarters
Washington, DC 20546-0001, USA

STATEMENT OF INTENT

The Consultative Committee for Space Data Systems (CCSDS) is an organization officially established by the management of member space Agencies. The Committee meets periodically to address data systems problems that are common to all participants, and to formulate sound technical solutions to these problems. Inasmuch as participation in the CCSDS is completely voluntary, the results of Committee actions are termed **Recommendations** and are not considered binding on any Agency.

This **Recommendation** is issued by, and represents the consensus of, the CCSDS Plenary body. Agency endorsement of this **Recommendation** is entirely voluntary. Endorsement, however, indicates the following understandings:

– Whenever an Agency establishes a CCSDS-related **standard**, this **standard** will be in accord with the relevant **Recommendation**. Establishing such a **standard** does not preclude other provisions which an Agency may develop.

– Whenever an Agency establishes a CCSDS-related standard, the Agency will provide other CCSDS member Agencies with the following information:

  • The **standard** itself.

  • The anticipated date of initial operational capability.

  • The anticipated duration of operational service.

– Specific service arrangements are made via memoranda of agreement. Neither this Recommendation nor any ensuing standard is a substitute for a memorandum of agreement.

No later than five years from its date of issuance, this **Recommendation** will be reviewed by the CCSDS to determine whether it should: (1) remain in effect without change; (2) be changed to reflect the impact of new technologies, new requirements, or new directions; or, (3) be retired or canceled.

In those instances when a new version of a **Recommendation** is issued, existing CCSDS-related Agency standards and implementations are not negated or deemed to be non-CCSDS compatible. It is the responsibility of each Agency to determine when such standards or implementations are to be modified. Each Agency is, however, strongly encouraged to direct planning for its new standards and implementations towards the later version of the Recommendation.

FOREWORD

The purpose of this Recommendation is to identify, define and provide structure to the relationships and interactions between an information Producer and an Archive.

Through the process of normal evolution, it is expected that expansion, deletion, or modification to this Report may occur. This Report is therefore subject to CCSDS document management and change control procedures which are defined in the *Procedures Manual for the Consultative Committee for Space Data Systems*. Current versions of CCSDS documents are maintained at the CCSDS Web site:

http://www.ccsds.org/

Questions relating to the contents or status of this report should be addressed to the CCSDS Secretariat at the address on page i.

At time of publication, the active Member and Observer Agencies of the CCSDS were:

Member Agencies

– Agenzia Spaziale Italiana (ASI)/Italy.
– British National Space Centre (BNSC)/United Kingdom.
– Canadian Space Agency (CSA)/Canada.
– Centre National d'Etudes Spatiales (CNES)/France.
– Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)/Germany.
– European Space Agency (ESA)/Europe.
– Instituto Nacional de Pesquisas Espaciais (INPE)/Brazil.
– Japan Aerospace Exploration Agency (JAXA)/Japan.
– National Aeronautics and Space Administration (NASA)/USA.
– Russian Federal Space Agency (FSA)/Russian Federation.

Observer Agencies

– Austrian Space Agency (ASA)/Austria.
– Central Research Institute of Machine Building (TsNIIMash)/Russian Federation.
– Centro Tecnico Aeroespacial (CTA)/Brazil.
– Chinese Academy of Space Technology (CAST)/China.
– Commonwealth Scientific and Industrial Research Organization (CSIRO)/Australia.
– Communications Research Laboratory (CRL)/Japan.
– Danish Space Research Institute (DSRI)/Denmark.
– European Organization for the Exploitation of Meteorological Satellites (EUMETSAT)/Europe.
– European Telecommunications Satellite Organization (EUTELSAT)/Europe.
– Federal Service of Scientific, Technical & Cultural Affairs (FSST&CA)/Belgium.
– Hellenic National Space Committee (HNSC)/Greece.
– Indian Space Research Organization (ISRO)/India.
– Institute of Space and Astronautical Science (ISAS)/Japan.
– Institute of Space Research (IKI)/Russian Federation.
– KFKI Research Institute for Particle & Nuclear Physics (KFKI)/Hungary.
– Korea Aerospace Research Institute (KARI)/Korea.
– MIKOMTEK: CSIR (CSIR)/Republic of South Africa.
– Ministry of Communications (MOC)/Israel.
– National Oceanic & Atmospheric Administration (NOAA)/USA.
– National Space Program Office (NSPO)/Taipei.
– Space & Upper Atmosphere Research Commission/Pakistan.
– Swedish Space Corporation (SSC)/Sweden.
– United States Geological Survey (USGS)/USA.

DOCUMENT CONTROL

Document	Title	Date	Status
CCSDS 651.0-B-1	Producer-Archive Interface Methodology Abstract Standard	May 2004	Original issue
CCSDS 651.0-M-1	Producer-Archive Interface Methodology Abstract Standard	May 2004	Original issue, reconfirmed in 2010 with new document number

CONTENTS

CONTENTS (continued)

Figure Page

Table

CONTENTS (continued)

1 INTRODUCTION

1.1 PURPOSE AND SCOPE

The purpose of this Recommendation is to identify, define and provide structure to the relationships and interactions between an information Producer and an Archive. This Recommendation defines the methodology for the structure of actions that are required from the initial time of contact between the Producer and the Archive until the objects of information are received and validated by the Archive. These actions cover the first stage of the Ingest Process as defined in the Open Archival Information System (OAIS) Reference Model (reference [1]). This Recommendation describes parts of the functional entities Administration ('Negotiate Submission Agreement') and Ingest ('Receive Submission' and 'Quality Assurance').

This Recommendation accomplishes the following:

- identifies the different phases in the process of transferring information between a Producer and an Archive;
- defines the objective of each of these phases, the actions that must be carried out during these phases, and the expected results (e.g., administrative, technical, contractual) at the end of a phase;
- forms a general methodological framework, which should be able to be applied and reused in those processes that relate to the Producer-OAIS Archive interface (this general framework should also provide sufficient flexibility for each particular case);
- forms a basis for the identification and/or development of standards and implementation guides in the community in question;
- forms a basis for identification and/or development of a set of software tools that will assist the development, operation and checking of the different stages in the process of information transfer between the Producer and the Archive.

NOTE – The term 'Archive' refers to an Archive that is in compliance with the OAIS Reference Model. This Recommendation uses terminology as defined in the OAIS Reference Model (reference [1]).

1.2 APPLICABILITY

The methodology defined in this Recommendation applies both to the information Producer and to the Archives to which this information must be transmitted, where such Archives are conformant to the Reference Model.

This methodology could also be of interest and be fully or partially applied to Archives that are not conformant to the Reference Model.

This methodology is of relevance to Archives containing physical, as well as digital, objects.

1.3 RATIONALE

Relationships between Archives and the Producers are rarely simple and easy. There are serious difficulties with the management of the Producer-Archive Interface in all the contexts which have been analyzed in preparation of this Recommendation (e.g., traditional Archives, libraries, Scientific Data Centers, business Archives).

These difficulties generally lead to an increased workload and may have negative consequences on the quality of the archived information. They can also have a negative effect on the relationship between the Archive and the Producer.

Potential problems include the following:

– the digital objects received do not conform to what the Archive expects;

– what the Producer delivers has not been clearly defined;

– the ingest schedule is not fulfilled by the Producer;

– errors in transfers are detected late by the Archive, or are not detected prior to use.

Within this context, the development of a standard methodology in this domain should aid in reducing these problems.

1.4 CONFORMANCE

This standard provides an abstract general methodology framework applicable to the interface between the Archive and the Producer. The aim is to first create a ‘Submission Agreement’, then transfer Submission Information Packages (SIPs) to the Archive, and finally to check these SIPs. In order for this standard to be easily applicable to a particular community, a specific standard or ‘community standard’ could be created in order to take into account all of the specific features of the community in question.

This community standard will be considered to conform to this abstract standard if:

– all of the actions have been considered and tailored as appropriate within the context of that community;

– the methodology for creating the community standard has addressed the various work phases defined in section 4, ‘Creating a Producer-Archive Interface Methodology Community Standard from the Producer-Archive Interface Methodology Abstract Standard’.

In the case that this abstract standard is directly used by a Producer and an Archive within the framework of a certain Producer-Archive Project, the methodology applied will be deemed as conforming to the abstract standard if all of the actions have been considered and implemented as appropriate within the context of that project.

1.5 DOCUMENT STRUCTURE

1.5.1 HOW TO READ THIS DOCUMENT

All readers should study subsections 1.1 (Purpose and Scope), 1.2 (Applicability) and 1.4 (Conformance) in order to understand the objectives and applicability of this Recommendation.

Readers seeking an overview of the methodology should also read section 2, 'An Overview of the Producer-Archive Interface Methodology'.

Those who will use the methodology should read the entire document.

NOTE – Working knowledge of the concepts and vocabulary defined in the OAIS Reference Model (reference [1]) is required in order to understand this Recommendation. Annex A contains a targeted overview of the OAIS Reference Model dedicated to the Methodology Abstract Standard.

1.5.2 ORGANIZATION BY SECTION

Section 1 defines the purpose, scope, applicability, rationale and definitions for terminology used in this Recommendation. It also specifies what is required for conformance to this standard.

Section 2 contains a general overview of the methodology, the players involved, their relationships and the activity phases that should be organized to manage the submission of information to an Archive for preservation and access.

Section 3 analyzes in detail each of the four phases defined in the methodology for all submissions. The phases are as follows:

- preliminary;
- formal definition;
- transfer;
- validation.

Section 4 describes the work stages that enable a methodological community standard to be created in conformance with this abstract standard.

The annexes listed here are not part of the this Recommendation and are provided for the convenience of the reader:

- Annex A contains a targeted overview of the OAIS Reference Model (reference [1]) dedicated to the Methodology Abstract Standard.
- Annex B contains the informative references.

– Annex C provides a table showing the correspondence between the preliminary phase and the formal definition phase.

1.6 DEFINITIONS

1.6.1 ACRONYMS AND ABBREVIATIONS

AIP	Archival Information Package
ASCII	American Standard Code for Information Interchange
CCSDS	Consultative Committee for Space Data Systems
DED	Data Entity Dictionary
DEDSL	Data Entity Dictionary Specification Language (reference [2])
DIF	Directory Interchange Format
DTD	Document Type Definition
EAD	Encoded Archival Description
EAST	Enhanced Ada SubseT
FGDC	Federal Geographic Data Committee
IEEE	Institute of Electrical and Electronic Engineers
ICA	International Council on Archives
ISO	International Organization for Standardization
MARC	MAchine-Readable Cataloging
OAIS	Open Archival Information System
PDI	Preservation Description Information
PDF	Portable Document Format
PVL	Parameter Value Language
RM	Reference Model
SGML	Standard Generalized Markup Language

SIP	Submission Information Package
TEI	Text Encoding Initiative
UML	Unified Modeling Language
XML	eXtensible Markup Language

1.6.2 TERMINOLOGY

Following is a short glossary of the OAIS terminology indispensable for this document. The terminology used is fully defined in reference [1], except the definitions printed in italics. Only brief definitions are provided here. This terminology does not seek to replace existing terminology in the various domains related to archiving. Each domain should be able to apply this methodology while retaining their specific terminology.

When first used in the text, the terms defined in the terminology are shown in bold.

Access: The OAIS entity that contains the services and functions which make the archival information holdings and related services visible to Consumers.

Archival Information Package (AIP): An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

Archive: An organization that intends to preserve information for access and use by a Designated Community.

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. This can include other OAISs, as well as internal OAIS persons or systems.

Content Data Object: The Data Object, that together with associated Representation Information, is the original target of preservation.

Content Information: The set of information that is the primary target for preservation. It is an Information Object comprised of its Content Data Object and its Representation Information. An example of Content Information could be a single table of numbers representing, and understandable as, temperatures, but excluding the documentation that would explain its history and origin, how it relates to other observations, etc.

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

Data Entity Dictionary (DED)*: A collection of semantic definitions of various data entities, together with a few mandatory and optional attributes about the collection as a whole. Data Entity Dictionaries may pertain to a single product, i.e., all the data entities within a single product are described in a corresponding single dictionary, or the Data Entity Dictionary*

may be a discipline-oriented dictionary that holds a number of previously defined data entity definitions which may be used by data designers and users as references.

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

Data Submission Session: A delivered set of media or a single telecommunications session that provides data to an OAIS. The Data Submission 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 these are represented on each media delivery or in the telecommunication session.

EAST*: The EAST language is a CCSDS and ISO norm. EAST offers means to describe the syntax of a data file, including:*

- *the fields in which it can be decomposed;*
- *structure (simple or composite);*
- *type (integer, real, enumerated, array, record, list);*
- *range (min value, max value);*
- *coding (ASCII, binary);*
- *location (rank, length);*
- *optionality (mandatory or not and, if not, presence condition);*
- *eventually, variable dimension (for arrays).*

Fixity Information: The information which documents the authentication mechanisms and provides authentication keys to ensure that the Content Information Object has not been altered in an undocumented manner.

Information: Any type of knowledge that can be exchanged. In an exchange, it is represented by 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 its Representation Information.

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

Meta-data*: Data about the content, the quality, condition and other characteristics of the data* (from the FGDC Standards Reference Model, reference [3]).

Packaging Information: The 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.

Preservation Description Information (PDI): The information which is necessary for adequate preservation of 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. This can include other OAISs or internal OAIS persons or systems.

Producer-Archive Project*: A Producer-Archive Project is a set of activities and the means used by the information Producer as well as the Archive to ingest a given set of information into the Archive.*

Representation Information: The information that maps a Data Object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of bits (i.e., a Data Object) is mapped into a symbol.

Submission Agreement: The 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 a telecommunication session.

In the framework of this abstract methodology, the Submission Agreement will also deal with other aspects such as validation, change management and schedule.

Submission Information Package (SIP): An Information Package that is delivered by the Producer to the OAIS for use in the construction of one or more AIPs.

Transfer*: The act involved in a change of physical custody of SIPs.* This definition is derived from the International Council on Archives [ICA] Dictionary on Archival Terminology (reference [4]).

The terms '**class**', '**association**', **and** '**aggregation**' refer to UML terminology.

1.7 APPLICABLE REFERENCES

[1] *Reference Model for an Open Archival Information System (OAIS*). Recommendation for Space Data System Standards, CCSDS 650.0-B-1. Blue Book. Issue 1. Washington, D.C.: CCSDS, January 2002. [Equivalent to ISO 14721:2003.]

[2] *Data Entity Dictionary Specification Language (DEDSL)—Abstract Syntax (CCSD0011).* Recommendation for Space Data System Standards, CCSDS 647.1-B-1. Blue Book. Issue 1. Washington, D.C.: CCSDS, June 2001. [Equivalent to ISO 21961:2002.]

[3] *FGDC Standards Reference Model*. Washington, DC: Federal Geographic Data Committee, March 1996. http://www.fgdc.gov/standards/refmod97.pdf

[4] *Dictionary of Archival Terminology: English and French with Equivalents in Dutch, German, Italian, Russian, and Spanish.* International Council on Archives, Handbook No. 7. 2nd ed., 1988.

[5] *Space Data and Information Transfer Systems—Standard Formatted Data Units—Structure and Construction Rules*, Recommendation for Space Data System Standards, CCSDS 620.0-B-2. Blue Book. Issue 2. Washington, D.C.: CCSDS, May 1992. [Equivalent to ISO 12175:1994.] (Note: You should ensure that this standard has the Technical Corrigendum - CCSDS 620.0-B-2 Cor. 1, November 1996 - applied.)

2 AN OVERVIEW OF THE PRODUCER-ARCHIVE INTERFACE METHODOLOGY

2.1 THOSE INVOLVED AND THEIR RELATIONSHIPS

2.1.1 THE PRODUCER

In conformance with the definition given in the Reference Model, the term '**Producer**' designates the persons and systems which supply the **Archive** with **Information** to be preserved.

The term 'Producer' thus covers a wide variety of situations: the Producer can be an editor, a scientific team, a laboratory, a company department, a Ministry, an administrative body, a private individual, etc.

The Producer's activities can be multiple and varied and they may require the involvement of a whole group of people with different skills and professions.

For the purpose of this methodology, it is assumed that the Producer is represented by a single person who has the responsibility for all the activities related to a phase, and for each of the phases identified in this methodology.

The Producer has his own management. This management defines the objectives and responsibilities of the Producer's activity, and provides him with the necessary resources. This management may be different from or the same as the Producer. In this Recommendation, the Producer and the Producer's management are differentiated and considered to be two different functions even if they are assumed by the same person.

2.1.2 THE ARCHIVE

The Archive is an OAIS Archive. The main responsibility of an Archive is to preserve a set of information and to make this available in an intelligible and useable form to a defined Designated Community.

In that context, the term 'Information' is used as defined in subsection 1.6.2 of this document, as well as in subsection 2.2.1 of the OAIS Reference Model (reference [1]). The OAIS framework is summarized in annex A.

The responsibilities of the Archive (e.g., which information to archive and which Designated Community) are defined by the OAIS Management.

2.1.3 GENERAL FRAMEWORK OF PRODUCER-ARCHIVE RELATIONSHIPS

There are a wide range of relationships and context situations that can exist between a Producer and an OAIS Archive, and they include the following:

- They can have the same management. This is the situation in a company, in which one department is entrusted to archive the information produced by the other departments.
- They can have different management, the **transfer** of data to be archived is, nonetheless, of an obligatory nature. This is the case for government Archives and Legal Deposit Libraries, whose tasks are defined by regulations or law.
- They can have a voluntary relationship when there is no obligation for the Producer to co-operate with the Archive. These Archives are called collecting Archives. Collecting Archives often specialize in one type of record such as labor union records, business records, commercial broadcasting records, or immigration records.
- They can have a contractual relationship. This is the case for 'commercial Archives', i.e., companies specializing in archiving and who ensure the preservation of information for other companies.

In some cases there is no relationship established between the Archive and the Producer, e.g., this is the case when an institutional library is entrusted to archive all electronic publications (CD-ROM) and, due to the great number of editors or to their non co-operation, there is no relationship—and thus no negotiation—between the Producer and the Archive. In this case, the library could decide to create a department, within its own structure, to collect electronic publications to be archived and prepare the SIPs. This department plays the role of a Producer with respect to the Archive department.

2.1.4 NEGOTIATION FOR AN AGREEMENT

The conditions under which negotiation takes place between the Producer and the Archive depend on the nature of the relationship between the Producer and the Archive and whether the archiving is mandatory or not.

This negotiation can be of an iterative nature. Negotiations should result in a '**Submission Agreement**'. This agreement precisely and thoroughly defines the different **Data Objects** which are to be transmitted to the Archive, the means used to transfer this data, the transfer schedule, etc.

In the absence of a relationship between the Producer and the Archive, as discussed previously, there is no negotiation with the actual Producer. For example, the Archive may collect information from various Web sites. In essence the Archive establishes a virtual Submission Agreement with the actual Producer without any negotiation beyond that involved in conformance to Web protocols. Virtual Submission Agreement is understood in the sense defined in subsection 2.3.2 reference [1].

Whatever the Archive/Producer relationships may be, experience shows that negotiations are easier when they are initiated very early on in the information creation process. It is always easier to agree on a data format before, rather than after, data are produced.

2.2 THE PRODUCER-ARCHIVE PROJECT

A **Producer-Archive Project** is a set of activities and the means used by the information Producer and the Archive to ingest a given set of information into the Archive.

Under the agreement between the Producer and the Archive, the Producer agrees to provide a set of information defined in the framework of a Producer-Archive Project. The following are contained within this set of information:

– the primary information that must be preserved;

– the complementary information, which is necessary for the **Archival Information Packages (AIPs)** to be made up, could include the following:

  • Information delivered by the Producer within the context of the Producer-Archive Project in question;

  • Information delivered by the same Producer within the context of the previous Producer-Archive Project;

  • Information delivered by another institution (for standards, for instance);

  • Information delivered by the Archive itself (Reference, **Fixity Information** of AIPs).

Periodic updates of the agreement may be required because additional data is collected, or the scope of data provided has been expanded to include additional areas of information. Technological changes or new standards may also imply agreement updates (see subsection 3.2.2.6).

2.3 THE PHASES

2.3.1 GENERAL DESCRIPTION

The Producer-Archive Interactions consist of four different phases:

– The Preliminary Phase, also known as a pre-ingest or pre-accessioning phase, includes the initial contacts between the Producer and the Archive and any resulting feasibility studies, preliminary definition of the scope of the project, a draft of the SIP definition and finally a draft Submission Agreement.

– The Formal Definition Phase includes completing the SIP design with precise definitions of the digital objects to be delivered, completing the Submission Agreement with precise contractual **transfer** conditions such as restrictions on access and establishing the delivery schedule.

– The Transfer Phase performs the actual transfer of the SIP from the Producer to the Archive and the preliminary processing of the SIP by the Archive, as it is defined in the agreement.

– The Validation Phase includes the actual validation processing of the SIP by the Archive and any required follow-up action with the Producer. Different systematic or in-depth levels of validation may be defined. Validations may be performed after each delivery, or later, depending on the validation constraints.

Each phase is carried out in chronological order. However, the transfer phase may overlap the validation phase.

Each phase is divided into a number of sub-phases (e.g., the sub-phases identified in summary-table 3-1) that also must be carried out in chronological order.

Each of these sub-phases is made up of one or more action tables. The action tables and the actions can be carried out in any order.

2.3.2 RELATIONSHIPS BETWEEN THE PHASES

Figure 2-1 provides a view of the relationships between these phases. In each text box on the top of the diagram, there is a brief indication of the goals of each phase. On the bottom, the outputs between each phase are depicted, as follows:

– The preliminary phase leads to a summary document on the feasibility of the Producer-Archive Project and approves proceeding to the formal definition phase (or stopping the project).

– This document is the basis on which the formal definition phase is developed. The formal definition phase leads to the Submission Agreement, which summarizes all the aspects of the formal definition phase, being drawn up. This agreement refers to a **Data Dictionary** and a formal model. Both of these elements are needed in order to proceed with the transfer phase.

– The outputs of the transfer phase are **Information Objects** that are input to the validation phase. As previously mentioned, validation may be able to be started before all the Information Objects have been delivered. The transfer and validation phases are often carried out partially in parallel, as there is iteration when all of the information to be submitted is not submitted at once.

– The Archive sends the Producer its validation report for the objects received, or forms reporting the anomalies found (the Archive may also acknowledge receipt of SIPs after ingest, and only notify the Producer if there were anomaly forms or invalid data).

There can be a significant lapse in time between the formal definition phase and the actual transfer phase. Within the Archives the transfer phase and the validation phase can take place concurrently if the actual transfer phase occurs over an extended length of time.

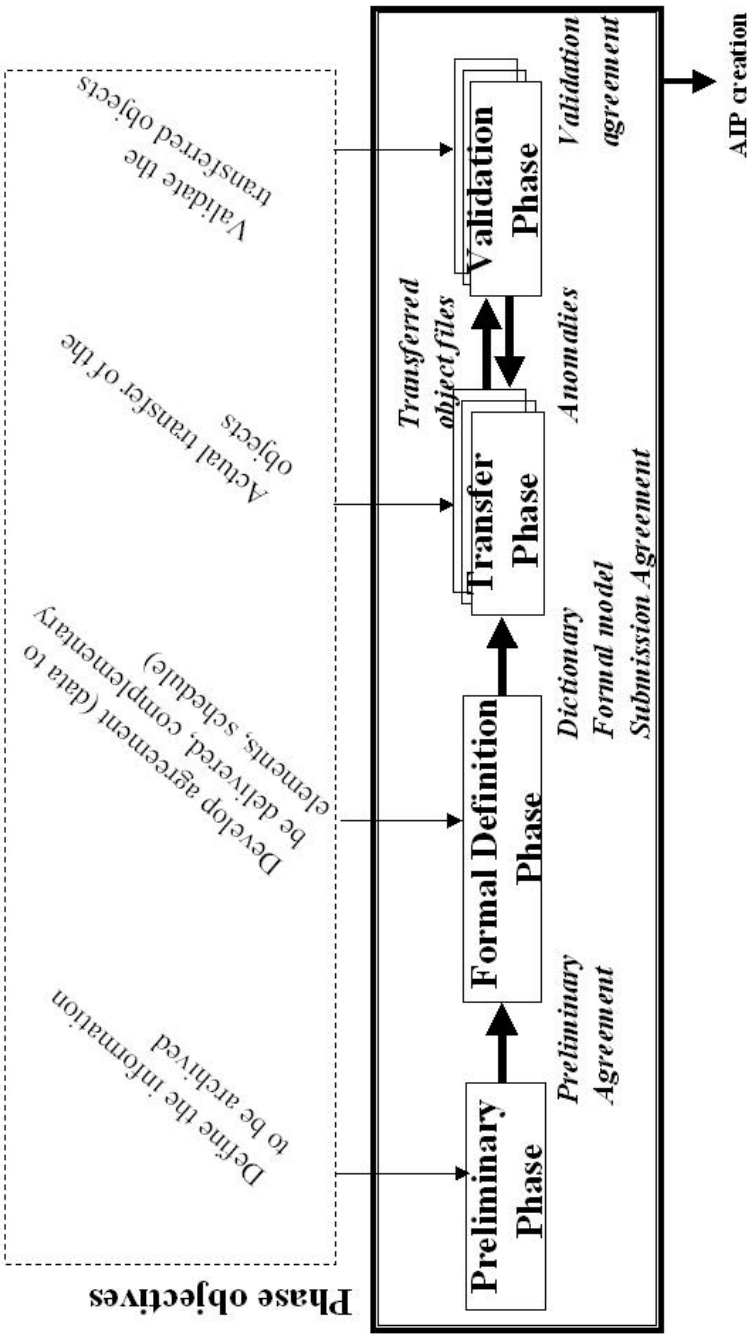

Figure 2-1: Main Phase Objectives and Outputs

3 DETAILED DESCRIPTION OF PHASES

NOTE – Each of the four phases detailed in this section may be divided into sub-phases (explained at the beginning of each phase). In this case, these sub-phases are identified in the left column in a summary table at the head of each related section and are to be addressed in the order given. The sub-phases are characterized by action tables identified in the right column. The action tables and the actions can be carried out in any order.

Each action is referenced with an identifier of the form C-n:

C = 'P, F, T, or V'. This character references the phase and means respectively 'Preliminary, Formal Definition, Transfer, Validation'.

n = 1 to total number of actions in the phase.

3.1 PRELIMINARY PHASE

The aims of the phase are as follows:

- to identify the primary information which the Archive must preserve;
- to establish a preliminary definition of the different Data Objects that will be transmitted to the Archive by the Producer;
- to analyze all aspects of feasibility;
- to decide on the feasibility of the Producer-Archive Project, from the Producer's as well as from the Archive's point of view;
- to make an estimate of the required resources;
- to draw up a Summary Document and, if appropriate, a preliminary agreement;

This phase is fundamental. It establishes the foundation on which a Producer-Archive Project can be built in the best way. Whenever possible, this preliminary phase **should** be carried out very early on, even before the information to be archived has been produced. This observation is based on practical experience. The preliminary phase is itself made up of three sub-phases:

- First contact.
- Preliminary definition, feasibility study and assessment of the Producer-Archive Project. The different subjects discussed in this phase are listed in table 3-1 and are detailed throughout subsection 3.1-2.
- Drawing up of the preliminary agreement.

These sub-phases are accomplished within the context of the standards, guides and tools available for this phase.

Table 3-1: Summary Table for Preliminary Phase

Summary of the Preliminary Phase	
Sub-phase	**Action Table**
First contact	First contact
Preliminary definition, feasibility and assessment	Information to be archived
	Digital objects and standards applied to these objects
	Object references
	Quantification
	Security conditions
	Legal and contractual aspects
	Transfer operations
	Validation
	Schedule
	Permanent impact on the Archive
	Summary of costs, risks
	Critical points
Establishment of a preliminary agreement	Establishment of a preliminary agreement

3.1.1 FIRST CONTACT

Table 3-2: Action Table for Preliminary Phase: First Contact

Id	Preliminary Phase: First Contact	Involves
P-1	Identify the contact persons and work organization	Producer and Archive
P-2	Exchange of general information	Producer and Archive

The first contact between the OAIS Archive and the Producer can be made on the initiative of the Archive, the Producer, the Archive management, or even by an external entity.

P-1 Identify the Contact Persons and Work Organization: This is the stage to agree in principle on how to proceed with the preliminary phase **in conformity with this Abstract Standard methodology** and to identify the main contact persons, both on the Producer's side and on the Archive's side. Complementary contact persons for specific questions (e.g., technical, administrative) can also be identified and their roles should be defined. These persons may also ask for help from experts depending on the point examined (e.g., standards, legal questions). The list of potential contacts includes appropriate subject matter specialists from the Archives.

The organization and division of work between the Producer and the Archive for this phase should also be defined at this point.

P-2 Exchange of general information: The Producer and the Archive have to exchange information:

- The Producer provides the Archive with a set of general information that concerns the type of the information to be preserved, its context, its schedule and its constraints. The Producer may also provide expectations regarding requirements of the Designated Community.

- The Archive provides the Producer with a description of its role, its general mode of operation, the standards that it generally applies, the tools that may be used in the Producer-Archive Interface, etc., and an assessment as to whether or not this information is appropriate for this Archive.

- The Archive submits the document describing the Producer-Archive Interface Methodology to the Producer, including the following items:

  – the main phases of the methodology, the basic requirements and the ways to apply them (e.g., Data Dictionary, models of the data to be delivered), including advantages and constraints;

  – the available service aids and tools for the methodology's application (e.g., existing data dictionaries, tools for creating dictionaries or formal models, service aids for creating descriptors).

At this point, each of the two partners can supply all information that may be useful to the project, i.e., general documents, reference documents, documentary references, and Internet site references.

3.1.2 PRELIMINARY DEFINITION, FEASIBILITY AND ASSESSMENT

This is the focus point of the preliminary phase. It should result in the following:

– Identification of the information to be preserved by the Archive and descriptions of the main features or properties of that information, as well as any dependencies on other information stored in this or some other Archive.

– Establishment of a preliminary definition of the Data Objects, Data sets and subsets, description of the main features or properties of those objects, sets and sub-sets, that the Producer is expected to transmit to the Archive.

– Both Archive and Producer create an assessment of the project cost.

– Each party will conduct a feasibility study of the project from their own point of view. Feasibility covers all aspects (including technical, financial, and legal) that could put the project in jeopardy.

– All the elements required to draw up a Submission Agreement have been collected.

The remainder of this subsection deals with a whole group of topics that must be analyzed as part of the preliminary phase. The depth of the analysis needed to reach the goal is not, a priori, defined. This depends on the context, the information to be archived and those involved. Definition of the required depth of analysis point by point is thus the responsibility of the Producer and the Archive.

The topics discussed in the remainder of this subsection are approached in the form of actions to be carried out by the Archive, the Producer, or both parties depending on the context. There is often interdependence between these topics.

Most of these topics can be approached and treated at the same time, e.g., information and standards, while respecting the dependencies (e.g., categories of digital objects must be identified before considering the likely number and size of the digital objects).

The Producer and Archive should ask the following questions for each of the topics examined:

– Does the topic concern the Producer-Archive Project?

– What level of definition should be reached in the preliminary phase?

– Is the topic critical for the Producer-Archive Project?

Some topics can be completely covered in this phase, whereas other topics should be further developed in the formal definition phase (these should be specified and noted in the summary document).

3.1.2.1 Information to Be Archived

Table 3-3: Action Table for Preliminary Phase: Information to Be Archived

Id	Preliminary Phase: Information to Be Archived	Involves
P-3	Identify the Content Information to be preserved	Producer and Archive
P-4	Identify the complementary information	Producer and Archive
P-5	Identify the Designated Community	Producer and Archive
P-6	Define Consumer access to the information	Producer and/or Archive
P-7	Assess the planned duration of the preservation of this information by this Archive	Producer and/or Archive
P-8	Assess the feasibility and costs induced by the previous actions	Producer and/or Archive

The Producer and the Archive shall develop the interdependent actions described in P-3 through P-8.

P-3 Identify the Content Information to be preserved: This is the primary starting point and it is important at this stage to clearly define and delimit the information which constitutes the primary object of the Producer-Archive Project. If there are still some open options, this is the time to make these explicit. The preliminary phase cannot be completed until this has been accomplished.

P-4 Identify the complementary information: The **Representation Information** and **Preservation Description Information (PDI)**. Draw up an inventory of the available data and information and those which must be created or gathered, and if necessary identify those that are mandatory for the preservation and those that are only useful.

Example of Complementary Information in a Space Mission

A Space Science mission is composed of experiments, an experiment produces data sets (main, auxiliary data, images), and a data set is a set of homogeneous files.

- Descriptive files for the mission and the experiments are the PDI, to include context, source file names from the laboratory, and references. The data sets (and their data files) are the Data Objects. A data set is described by a Directory Interchange Format (DIF) file.

- The data files are described by the following Representation Information:

  * An **EAST** (ISO language for data description) structure file, giving the exact structure bit per bit of the data files (syntactic representation).

  * A **Data Entity Dictionary (DED)** file describing the semantics of the data files.

P-5 Identify the Designated Community: Specifically identify how and by whom the data will be used, e.g., whether for the general public or for researchers. This action affects the required level of information (high or low) and the previous action, 'Identify the complementary information'. It also affects access (e.g., search by keyword, by author, by time-related or geographic criteria) and the next action, 'Define **Consumer** access to the information'. Obtain a preliminary identification of the Descriptive Information required. However, it should be noted that for some institutional and/or governmental Archives neither the Producer nor the Archive has a precise idea of how the information to be preserved will be used. Even with scientific observation Archives, 10 years after data production, scientific data is used in ways that the Producers could not even imagine.

P-6 Define Consumer access to the information: Define to the extent known (also see the subsection 3.1.2.5, 'Security Conditions'), including the following:

- unrestricted or limited access;
- free or paid access;
- availability and access authorization over time (defining when the records are available to specific classes of Consumers);
- required service level, i.e. speed, performance, type of access (e.g., interactive server, data transfer by network or on a digital media), typical selection criteria and requested volumes of data dissemination expected, and research aids.

P-7 Assess the planned duration of the preservation of this information by this Archive: Assess duration and attempt to identify a successor Archive if appropriate.

P-8 Assess the feasibility and costs: The Producer and the Archive assess the costs induced by the actions listed within the definition of Consumer access. If this cost reveals clear non-

feasibility, stop the work at this stage and possibly restart on a new basis. This remark is valid for additional actions listed in table 3-4.

3.1.2.2 Digital Objects and Standards Applied to These Objects

Table 3-4: Action Table for Preliminary Phase: Digital Objects and Standards Applied to These Objects

Id	Preliminary Phase: Digital Objects and Standards Applied to These Objects	Involves
P-9	Make a preliminary identification of the Data Objects related to the different categories of information to be archived	Producer and Archive
P-10	Define the rules and standards related to these objects that are accepted by the Archive	Archive
P-11	Describe the tools available for the application of the rules and standards known by the Archive	Archive
P-12	Provide the rules and standards applied to Data Objects by the Producer	Producer
P-13	Describe the tools available for application of the rules and standards known by the Producer	Producer
P-14	Assess the compatibility and study solutions	Producer and Archive
P-15	Assess the efforts to be made and the associated costs	Producer and Archive

P-9 Make a preliminary identification of the Data Objects: This enables a first list of object categories to be drawn up. These include the **Content Data Objects**, which contain the primary information to be preserved, the Data Objects containing Representation Information on the primary Data Objects, and the Data Objects describing the context and source of the primary information.

The Producer and the Archive must ensure that both ingest and future preservation actions preserve the significant properties (such as accuracy and precision in number representation) of the Information Objects. (See Section 3.2.2)

For each of these object categories, priority is given to the Content Data Objects and their associated Representation Information. The Archive and Producer should attempt to reach an agreement on what the Producer will create and what the Archive will receive.

P-10, P-11, P-12, P-13 Define the Rules, Standards and Tools: The following paragraphs cover actions and provide some examples concerning discussion of rules, standards and tools:

- Standards applicable to Content Data Objects: data files in ASCII or binary, the form of which is defined by a specific application, particular standards applicable to the geographic representation of information or the representation of time and dates, standards related to a profession, sound, image, video files, SGML or XML files conforming to a DTD or a predefined schema, PDF files, etc.
- Standards applicable to Data Objects containing the Representation Information of **Content Information**: simple reference to a standard that should also be archived or use of a syntactic data description language (e.g., EAST), semantic description language (DEDSL, SGML, PVL, XML), etc.
- Standards applicable to **Meta-data** levels: ISO/TC211 standards for the description of geographic data, MARC for libraries, DIF for scientific data, DTD EAD for the archivists, etc.
- If the standards accepted by the Archive do not correspond to those used by the Producer, it is possible that the availability of tools aiding the use of these standards could enable the partners to find common ground. Possible solutions should be analyzed in terms of technical feasibility and cost. If the objects already exist, what are the necessary migration efforts? Otherwise, what would be the effort required to create the objects to satisfy the requirements?

P-14 Assess the compatibility and study solutions: Assess the compability between the rules, standards and tools already in place and those that should be used. Carry out a study of the possible solutions.

P-15 Assess the efforts and associated costs: Deduce from the previous study what resources must be deployed and the relevant costs.

3.1.2.3 Object References

Table 3-5: Action Table for Preliminary Phase: Object References

Id	Preliminary Phase: Object References	Involves
P-16	Draw up an inventory of the information on the existing identification rules or nomenclature within the domain, legal provisions, and standards	Producer and Archive
P-17	Define the rules that could or should be applied within the context of the Producer-Archive Project	Producer and Archive
P-18	Assess the associated costs	Producer

P-16 Draw up an inventory of the information: The Archive provides the Producer with information on:

- the existing identification rules or nomenclature (e.g. bibliographic description, namespaces);
- any possible legal provisions imposed by applicable local, provincial, state or national policy, guidelines or legislation;
- the standards used.

P-17 Define the Rules: The Producer and Archive negotiate the pertinent rules to be applied to the Producer-Archive Project.

P-18 Assess the associated costs: The Producer evaluates the cost of these constraints.

3.1.2.4 Quantification

Table 3-6: Action Table for Preliminary Phase: Quantification

Id	Preliminary Phase: Quantification	Involves
P-19	Estimate the data volume to be transmitted to the Archive	Producer
P-20	Assess the permanent data volume to store	Archive
P-21	Assess the storage capability needed for the ingest process	Archive
P-22	Assess the associated costs	Archive

P-19 Estimate the data volume: The Producer must estimate the volumes to be transmitted in the short, medium and long term (global volume, minimum, average, and maximum planned size of files, number of files), as well as the frequency of the transfer sessions. These elements have an influence on the technique used for the transfer.

P-20 Assess the permanent data volume: The Archive must estimate the permanent global data volume to store with the elements (listed in P-19) provided by the Producer. This estimate implies an associated cost for the Archive. This cost is evaluated in subsection 3.1.2.10.

P-21 Assess the storage capability: The Archive must assess the storage needed for the ingest process (data storage before transformation to AIP and transfer to OAIS storage function).

NOTE – The preceding action is dependent on the choices made for the standards applicable to transmitted Data Objects. For the Data Objects containing scientific observations it has frequently been noted that the volume of data coded in ASCII can be twice as large as the same data coded as IEEE floating numbers. In much the same way, the size of a file structured in XML can be much larger than the same file in simple text. Although it may be possible to use the same format, the transfer format and the storage format do not need to be the same.

P-22 Assess the associated costs: The Archive must assess the cost associated with the storage needs.

3.1.2.5 Security Conditions

Table 3-7: Action Table for Preliminary Phase: Security Conditions

Id	Preliminary Phase: Security Conditions	Involves
P-23	Identify the requirements for confidentiality of the information and for authentication of the source of the information in the transfer between the Producer and the Archive	Producer and Archive
P-24	Identify the requirements for security of the holdings at the Archives	Archive
P-25	Identify the requirements for confidentiality of the information and for authentication of the source of the information in the transfer between the Archive and the Consumer	Producer and Archive
P-26	Identify the standards and tools that could be used	Producer and Archive
P-27	Assess the associated costs	Producer and Archive

P-23 Identify requirements for confidentiality and authentication between Producer and Archive: The Producer and the Archive need to consider:

- Confidentiality of the information in the transfer: This means that the Archive and Producer must implement required measures such as encryption of the information held by the Producer and the use of secure transfer techniques.
- Authenticity of the information in the transfer: This may imply the establishment of encoding and signature mechanisms—at a digital object transmission level—in order to guarantee this authenticity. In-depth validation and particular attention to documentation are also important aspects.

P-24 Identify requirements for security: Implementation of specific measures for security of the holdings, including storage vaults, limiting physical access, separation of master and copy, etc, may be required by the Archive and may include the following:

- Specific storage measures for the Archive.
- This also implies the subsequent establishment of techniques to guarantee the integrity of preserved objects (including the definition of backup procedures). Documenting the preservation process and maintaining an untouched set of the data in archival storage are also important aspects.
- The Archive must take into account the change of technology in the long term.

P-25 Identify requirements for confidentiality and authentication between Archive and Consumer: Confidentiality of the information in the transfer between the Archive and the Consumer and authenticity of the information in the transfer between the Archive and the Consumer is subject to the same considerations discussed in P-23. Furthermore, numerous Consumers on different sites may access the same Archive. This could impact on the techniques used.

P-26 Identify standards and tools: For each action examined in this subsection, the following should be made explicit: identification of the applicable regulations and specification of standards and tools that could be used.

P-27 Assess the associated costs: Assess costs to cover these aspects.

3.1.2.6 Legal and Contractual Aspects

Table 3-8: Action Table for Preliminary Phase: Legal and Contractual Aspects

Id	Preliminary Phase: Legal and Contractual Aspects	Involves
P-28	Define the nature of the relationships between the Archive and the Producer	Producer and Archive
P-29	Assess the problem of intellectual property	Producer and Archive
P-30	Define the conditions for access to data	Producer and Archive
P-31	Address Archive certification	Archive
P-32	Provide the standards and tools used	Producer and Archive
P-33	Assess the associated costs	Producer and Archive

This subsection examines all of the aspects that involve legal consideration. These aspects depend to a large extent on the nature of the relationships between the Archive and the Producer that should thus be made explicit.

P-28 Define the nature of the relationships between the Archive and the Producer: The Archive and the Producer should examine and answer the following questions:

- Does the Producer-Archive Project enter into the context of statutory government archiving? What are the consequences of this aspect of the project?
- If the relationship between the Archive and the Producer are of a contractual type, what is the aim of the contract and how are the responsibilities for the Archive defined within this contract?
- What are the specific responsibilities implied by their relationships?

P-29 Assess the problem of intellectual property: Is the data to be archived subject to intellectual property rights? What are the consequences for the Archive? The Archive must, of course, already be familiar, or become familiar, with the national or international legislation on copyrights. Does the transfer of data between the Producer and the Archive imply a transfer of these rights?

- If so, what documents should be provided in order to legalize this transfer?
- If not, what obligation does the Archive have with respect to this data?

When negotiating intellectual property rights the Archive should distinguish between preservation and access. It may be necessary to secure an agreement to preserve, although no one will be granted access. This may be the only way to prevent loss of historically important material, as the original medium and technology are unlikely to survive long enough for copyright expiry.

P-30 Define the conditions for access to data: What obligation does the Archive have with respect to information protection and access to this information? Define the rules which govern these conditions (e.g., authorized persons, immediate access, or authorized after a legal lapse of time).

P-31 Address Archive certification: The different issues brought up here may also imply that the Archive should be certified with respect to an Archive certification baseline, if this in fact exists.

P-32 Provide the standards and tools used: For each topic examined, the following should be made explicit: identification of the applicable regulations, and specification of the standards and tools that could be used.

P-33 Assess the associated costs: Assess costs to cover these aspects. These aspects should be included in the Submission Agreement.

3.1.2.7 Transfer Operations

Table 3-9: Action Table for Preliminary Phase: Transfer Operations

Id	Preliminary Phase: Transfer Operations	Involves
P-34	Make a preliminary definition of the SIPs	Producer and Archive

P-35	Exchange the requirements and constraints with respect to the transfer of Data Objects and identify possible solutions	Producer and Archive
P-36	Assess associated costs	Producer and Archive

P-34 Make a preliminary definition of the SIPs: The Producer and the Archive should together study the possible solutions as regards the SIP. More precisely, it is important to study the packaging of the different Data Objects for their transmission to the Archive.

P-35 Exchange the requirements and constraints with respect to the transfer: The Producer and Archive exchange their transfer constraints and requirements for network or media support (e.g., compact disc). They identify communication protocols and the tools which could be used (e.g., ftp, http) and adapted (depending on the frequency and volumes). It may be necessary to envisage an automated transfer, a secure transfer for which the required level of security should be defined (also see subsection3.1.2.5). Producer and Archive identify the possible solution(s), taking into account the identified requirements and constraints.

P-36 Assess associated costs: Assess the associated costs related to these operations.

3.1.2.8 Validation

Table 3-10: Action Table for Preliminary Phase: Validation

Id	Preliminary Phase: Validation	Involves
P-37	Supply the Producer with information on the SIP validation procedures, the reject procedures, and the tools that are applied by the Archive	Archive
P-38	Study the development or modification of the validation tools required	Archive
P-39	Study the implementation of quality methods (and tools) to answer needs	Producer
P-40	Assess associated costs	Producer and Archive

P-37 Supply information on SIP validation: The Archive provides the Producer with general information dealing with:

– Validation procedures for the SIPs that it uses. It is important to distinguish, on the one hand, the validation methods for the reception of a SIP with conformity to the model and, on the other hand, the validation methods that concern the content of SIP objects.

- Reject procedures in the event of an anomaly.
- Validation tools. Some of these tools may be supplied to the Producer for validation at this end before transfer.

P-38 Study validation tools: The Archive may need to modify existing tools, or develop new tools, in order to adapt to the context of the Producer-Archive Project.

P-39 Study quality methods: The Producer makes an independent study of the actions to be considered in order to fulfill the quality and validation requirements of the Archive.

P-40 Assess associated costs: The Producer and Archive each assesses their costs associated with these actions.

3.1.2.9 Schedule

Table 3-11: Action Table for Preliminary Phase: Schedule

Id	Preliminary Phase: Schedule	Involves
P-41	Define a preliminary schedule	Producer and Archive

P-41 Define a preliminary schedule: The Archive and the Producer must negotiate a preliminary schedule for data production, transfer, validation, data archiving, and data availability for the Designated Community.

3.1.2.10 Permanent Impact On the Archive

Table 3-12: Action Table for Preliminary Phase: Permanent Impact On the Archive

Id	Preliminary Phase: Permanent Impact On the Archive	Involves
P-42	Assess the permanent impact and the associated costs on the Archive	Archive

P-42 Assess permanent impact and associated costs: These actions are the Archive's responsibility. They concern an assessment of any possible future impact on archiving the data in question, beyond the ingest operation time. This impact and the associated costs take into account:

- The permanent data volume to store, which is estimated in subsection 3.1.2.4. This volume may imply an increase in the number of storage archive volumes, or changes in the media type and an associated cost.
- The necessary long-term preservation actions (for example, media renewal, duplication, re-packaging, and transformation of information). Long-term migration should also include plans for transfer of information to another Archive in the case of closure of the current Archive.

- Establishment of specific precautionary measures to avoid the loss of data (e.g., destruction, alteration), such as copying data to another Archive. In the event of loss or alteration of data, the Archive will inform the Archive Management, the Producer (if it is still available) and the Designated Community of any measures taken.
- Security requirements (also see subsection 3.1.2.5).

It is important that the Archive defines and maintains a cost model to be able to estimate the cost of maintaining the Archive when the speed and direction of technological changes is not known in advance.

3.1.2.11 Summary of Costs, Risks

Table 3-13: Action Table for Preliminary Phase: Summary of Costs, Risks

Id	Preliminary Phase: Summary of Costs, Risks	Involves
P-43	Carry out a cost summary and estimate risks	Producer and Archive

P-43 Carry out a cost summary, estimate risks: Producer and Archive should make a summary of the different costs, based on the activities outlined so far in subsection 3.1.2, on a short, medium and long-term basis. Each side should assess the costs that may be implied for them. The following aspects should be taken into account:

- possible changes, either on the side of the Producer or Archive, which would require new investment in the end (e.g., new data collection, technical changes, etc.);
- available resources and means (human and material);
- risks on either the side of the Archive or Producer;
- available budgets (possibly readjust them).

This summary could lead to numerous negotiations that in turn could lead to an agreement on both sides.

3.1.2.12 Critical Points

Table 3-14: Action Table for Preliminary Phase: Critical Points

Id	Preliminary Phase: Critical Points	Involves
P-44	Assess the critical points	Producer and Archive

P-44 Assess the critical points: The Producer and Archive must assess, from among all the points that have already been raised, which ones may cause serious problems and could imply a risk of complete or partial failure for the Producer-Archive Project.

3.1.3 ESTABLISHMENT OF A PRELIMINARY AGREEMENT

Table 3-15: Action Table for Preliminary Phase: Establishment of a Preliminary Agreement

Id	Preliminary Phase: Establishment of a Preliminary Agreement	Involves
P-45	Draw up a document that summarizes the preliminary phase, with a feasibility assessment and a recommendation on proceeding with the formal definition phase (or stopping it)	Producer and/or Archive
P-46	Make a preliminary agreement to proceed to the next phase	Producer and Archive

P-45 Draw up a summary document: This is the last sub-phase and concluding step of the preliminary study examined above (the first two sub-phases of this Preliminary Phase). The Producer and/or the Archive have to draw up an understandable document - how the drafting of the document is divided up must be decided between the two parties - that is a summary of he previous analyses. Particularly, this summary document provides a basis on which the feasibility of the project can be decided and also contains the critical points of the project. The conclusion is a recommendation on proceeding with the formal phase, or stopping the project. In this last case, alternative solutions should be considered (e.g., financing).

P-46 Make a preliminary agreement to proceed to the next phase: At this stage, the Producer and the Archive should make a preliminary agreement. This is not yet the final Submission Agreement (which is finalized at the end of the formal definition phase), but a preliminary agreement to proceed with the next phase, which is the formal definition phase. This agreement could be part of the previous summary document.

The preliminary agreement contains the first elements:

- the SIP content (Content Information, PDI, **Packaging Information**, Descriptive Information), and the data model;
- a first submission timetable;
- data access restrictions;
- validation procedures;
- revision and re-negotiation clauses.

3.2 FORMAL DEFINITION PHASE

The aim of this phase is the negotiation of the 'Submission Agreement', which includes a complete and precise definition of:

- the data to be delivered to the Archive by the Producer;
- the contractual and legal aspects;
- the complementary elements required to define the transfer and validation process;
- the schedule.

The formal definition phase is itself made up of three sub-phases:

- organization of the formal definition phase;
- formal definition (see table 3-16 for a summary of the components of this phase).
- drawing-up of the Submission Agreement for approval by the Producer and the Archive.

This is accomplished in the context of the standards, guides and tools available for this phase.

The topics discussed so far in 3.2 are dealt with in a more precise way in the following paragraphs in the form of check-lists of actions to be carried out. They may require negotiation between the Archive and the Producer. Most of these topics can be examined and dealt with at the same time, while respecting inter-dependencies (e.g., the information must be identified before creating the Data Dictionary).

Table 3-16: Summary Table for Formal Definition Phase

Summary of the Formal Definition Phase		
Sub-phase	**Action Table**	
Organization of the Formal Definition Phase	Organization of the Formal Definition Phase	
Formal definition	Information to be preserved and Model of Data Objects to be delivered	General project context and definition of Information Objects
		Creation of a Data Dictionary
		Construction of a formal model
	Formalization of contractual and legal aspects	
	Definition of transfer conditions	
	Validation definition	
	Delivery schedule	
	Change management after completion of Submission Agreement.	
	Feasibility, costs and risks assessment	
Submission Agreement	Submission Agreement	

Figure C-1 in Annex C shows the relation between the stages of the preliminary phase and those of the formal definition phase.

The actions identified in the preliminary phase are treated in a formal way in this phase. Certain sections of the formal definition phase are new. The following should be taken into account:

– Subsection 3.1.2.4 (‘Quantification’) of the preliminary phase broaches numerous aspects which are partly discussed in subsections 3.2.2.1.2 (‘General Project Context and Definition of Information Objects’), 3.2.2.3 (‘Definition of Transfer Conditions’), and 3.2.2.7 (‘Feasibility, Costs and Risks Assessment’).

- The actions discussed in subsection 3.1.2.10 ('Permanent Impact on the Archive') should be reassessed regarding their costs in subsection 3.2.2.7 ('Feasibility, Costs and Risks Assessment').
- The critical points identified by the partners in the Summary of the Preliminary Phase must be dealt with separately in the section related to the formal definition phase.

3.2.1 ORGANIZATION OF THE FORMAL DEFINITION PHASE

Table 3-17: Action Table for Formal Definition Phase: Organization of the Formal Definition Phase

Id	Formal Definition Phase: Organization of the Formal Definition Phase	Involves
F-1	Setup the management of the formal definition phase	Producer and/or Archive
F-2	Specify the points previously raised which are to be made explicit in the formal definition phase	Producer and Archive

F-1 Setup the management of the formal definition phase: The Archive and the Producer must negotiate the organization of the formal definition phase, as well as the definition of their individual roles and responsibilities, as follows:

- plan the different archiving stages (production, transfer, ingest), identify the key points and specify how technical approval is obtained (plan the validation phase);
- define the documents to be produced and identify who is producing and maintaining these documents.

F-2 Specify points to be made explicit: The Archive and the Producer must specify the points in the preliminary phase that need to be examined in greater depth.

3.2.2 FORMAL DEFINITION

3.2.2.1 Information to Be Preserved and Model of Data Objects to Be Delivered

This subsection discusses the precise definition of the information to be transferred from the Producer to the Archive. This definition is a formal model of objects to be delivered. This Model contains a definition of the objects to be delivered that is as precise and non-ambiguous as possible.

Three main work stages are required to create this model:

- Description of the general objectives and project context, definition of all the information Objects, definition of the coding, format, and Information Object identifiers, put into the form of an understandable document. Create an unambiguous record of the decisions and agreements. The record should be available to both the

Producer and the Archive. All of these points have already been studied in the preliminary phase.

– Definition of the object classes associated with the aforementioned Information Objects, and creation of an associated Data Dictionary to list these definitions.

– Construction of the formal model of the Producer-Archive Project.

3.2.2.1.1 General Project Context and Definition of Information Objects

Table 3-18: Action Table for Formal Definition Phase: General Project Context and Definition of Information Objects

Id	Formal Definition Phase: General Project Context and Definition of Information Objects	Involves
F-3	Define the general project context as well as the list and contents of the information elements to be delivered	Producer and Archive
F-4	Define the formats, coding rules, and standards to be applied for the objects to be delivered	Producer and Archive
F-5	Define the volume indicators	Producer
F-6	Define the references for the objects to be delivered	Producer and Archive
F-7	Choose the tools on the Producer's side	Producer and Archive
F-8	Write a description of the Information Objects referring to a Data Dictionary and a model (part of the final agreement)	Producer and/or Archive

F-3 Define the general project context: At this stage the Producer and Archive must agree on all the information elements to be preserved and on the following content to be delivered:

– Content Information: Data Object and Representation Information (syntactic and semantic).

– Preservation Description Information (provenance, context, reference, fixity).

– Descriptive Information.

It is presumed that the Designated Community and the access conditions have already been identified during the preliminary phase. This has an impact on the level of complementary information to be archived with the Data Objects, as well as on the Descriptive Information.

The Producer and the Archive must agree on the contents of the documents describing the information elements. Several levels can be established, e.g., a standard document model

(with a table of contents model), or specifications that define the required elements, the recommended elements, and the optional elements.

F-4 Define the formats, coding rules, and standards: The Producer and the Archive must then choose the format, the coding rules, and the standards to be applied, for each of the objects defined in F-3, drawing on the elements already provided during the preliminary phase. Some objects already exist, while others do not. If the format of existing objects does not correspond to the specified format, the Producer and the Archive must reach an agreement (e.g., migrations).

F-5 Define volume indicators: The Producer provides the Archive with information on the volume measurements (e.g., estimated total volume to be archived and also granular information on the volume of Content Data, mean and maximum size of a file).

F-6 Define the references: Producer and Archive define the references of the information elements, drawing on the results of the preliminary phase.

F-7 Choose the tools: Producer and Archive define the tools to be installed by the Producer or acquired by the Archive (to aid with data production, production of descriptors, document production, etc.).

F-8 Write a description of the Information Objects: A description of the elements previously negotiated by the Archive or the Producer must create an unambiguous record of the decisions and agreements. The record should be available to both the Producer and the Archive. This description will be part of the final Submission Agreement. This description refers to the Data Dictionary and the formal model (defined in subsections 3.2.2.1.3 and 3.2.2.1.4).

NOTE – The Packaging Information is defined in the transfer stage (see subsection 3.2.2.3).

3.2.2.1.2 Creation of a Data Dictionary

Table 3-19: Action Table for Formal Definition Phase: Creation of a Data Dictionary

Id	Formal Definition Phase: Creation of a Data Dictionary	Involves
F-9	Define the object classes and their attributes, set up the associated Data Dictionary	Producer and Archive
F-10	Code the Data Dictionary	Producer or Archive

F-9 Define the object classes and their attributes: From the information already provided, the Producer and the Archive define the classes of the objects associated with all the defined

information and their attributes. These classes could be subject to change (see subsection 3.2.2.6, 'Change Management after completion of the Submission Agreement').

F-10 Code the Data Dictionary: The complete, formal and precise definition of the different classes of Data Objects to be delivered constitute the project Data Dictionary. This Data Dictionary could conform to the Data Entity Dictionary Abstract Standard (reference [2]). Its implementation could conform to the references [B2] or [B3], or be subject to specific implementation.

It is recommended that the Data Dictionary be modeled on existing dictionaries.

3.2.2.1.3 Construction of a Formal Model

Table 3-20: Action Table for Formal Definition Phase: Construction of a Formal Model

Id	Formal Definition Phase: Construction of a Formal Model	Involves
F-11	Define the model of the data to be delivered	Producer and Archive
F-12	Draw up a model representation	Producer or Archive

F-11 Define the model of the data to be delivered: The formal model identifies the different instances of Data Objects that will be delivered. This model defines the nature of the relationships between these different instances. It also provides a logical and coherent overall view of the whole set of objects. How the model is created depends on the transfer possibilities (i.e., whether objects will be delivered in a separate manner or not). The granularity of the model will enable the definition of the Data Objects, or set of Data Objects, that may be delivered independently. This data or set of Data Objects is the basis for the definition of the SIPs. There is no single, unique model; moreover, this model may be subject to change (see subsection 3.2.2.6, 'Change Management after completion of the Submission Agreement').

F-12 Draw up a model representation: It is recommended that the model be defined using a formal language. A text document may accompany the model, if this is useful, particularly for complex models.

3.2.2.2 Formalization of Contractual and Legal Aspects

Table 3-21: Action Table for Formal Definition Phase: Formalization of Contractual and Legal Aspects

Id	Formal Definition Phase: Formalization of Contractual and Legal Aspects	Involves

F-13	Draw up legal and contractual agreements between the Archive and the Producer concerning the data (part of the final agreement)	Producer and Archive

F-13 Draw up legal and contractual agreements: This step concerns formalizing all the points already stated in the preliminary phase and reaching an agreement on this matter by the Archive and the Producer. In particular, if a transfer of intellectual property must take place, the conditions and the date of this transfer must be defined at this level.

3.2.2.3 Definition of Transfer Conditions

Table 3-22: Action Table for Formal Definition Phase: Definition of Transfer Conditions

Id	Formal Definition Phase: Definition of Transfer Conditions	Involves
F-14	Define the communication procedures (digital network, protocols, media, etc.)	Producer and Archive
F-15	Define the Packaging Information of delivered objects (in what form the data is delivered)	Producer and Archive
F-16	Define a transfer session (functional and time-related structure of the transfer of digital objects)	Producer and Archive
F-17	Define the initial transfer test	Producer and Archive
F-18	Identify the tools that may be used during the transfer phase	Producer and Archive
F-19	Write a description of the transfer procedures (based on F-14 through F-18).	Producer and/or Archive

F-14 Define the communication procedures: The Archive and the Producer must precisely define the communication procedure—type of transfer and type of media used for the transfer of objects—drawing on the elements in the preliminary phase, and taking account of elements which have an impact on the scale of transfer and reception operations, such as data volume and frequency, maximum number of objects delivered by session, and maximum and mean object size. The volume of the data delivered by session has been estimated in subsection 3.1.2.4.

Various scenarios may occur for the transfer of data from the Producer to the Archive. Potential scenarios include transfer via a physical media and transfer via a network where, for example, the Archive fetches data from a predefined site. The communication procedures may involve the particular means used in order to ensure the security conditions identified in

the preliminary phase (see subsection 3.1.2.5, 'Security Conditions') to include authenticity, integrity and/or confidentiality of the data.

F-15 Define the Packaging Information: The Archive and the Producer must agree on the technical choices concerning Packaging Information and those already looked at in the preliminary phase.

Producer and Archive must define how the objects or set of Data Objects of the formal model will be packaged. For example, a set of attributes about a data file might be expressed using XML and be combined with the data file bytes using a standard packaging approach such as ISO 12175 (reference [5]).

F-16 Define a transfer session: The actual transfer of Data Objects is divided into successive sessions. The notion of time-sequence also structures the data transfer into successive stages. This is a logical concept regardless of the physical resources used.

A Submission Session is a term defined in the OAIS Reference Model (reference [1]). It is an operation that enables data transfer from the Producer to the Archive to be carried out. A transfer session thus corresponds to the set of objects that are delivered by:

- transmission on a private or public (Internet) network, by ftp, E-mail, http, etc.;
- delivering a package of one or more physical media.

The Archive and the Producer must define:

- on the one hand, the functional structure of a session. A session may be a homogeneous package of objects (e.g., set of documentation, file packet of scientific data), or a retransfer of data following non-conformities, or an update.
- on the other hand, the structure with respect to time. In fact, very often all the instances of the Data Objects of a model may not be delivered simultaneously, but in several sessions (depending on the data production, the means of transfer, etc.). This process can be spread over several months, or several years or be ongoing.

The characteristics of the session (e.g., identifier, date, version, start and end date in the case of an ongoing process) must take into account the previous items concerning the functional structure of the session, and its structure with respect to time. This information could be in a file provided simultaneously.

Lastly, the Archive and the Producer must establish a procedure for sending/receiving messages (e.g., forms, e-mails, acknowledgement of receipt), depending on needs. The Archive must have precise information on the contents of a session and, in turn, inform the Producer of the correct reception of the objects (e.g., in order to acknowledge session reception, the Archive may send an e-mail, provide a receipt or a letter of acknowledgement to the Producer indicating the date and contents of the reception).

A systematic validation is described in subsection 3.2.2.4, 'Validation definition'.

F-17 Define the initial transfer test: The Producer and Archive must:

- Define the test SIPs.
- Identify the various kinds of tests, the aim of which is to check the following:
  - On the one hand the nominal functioning of the transfer: tests at the utmost limit (maximum volume of a file, maximum number of files), and then test performance. Test of the integrity of the objects received.
  - On the other hand, the procedures in the event of breakdown (for example, in the case of the transfer being interrupted).

F-18 Identify tools used for the transfer phase: The Producer and Archive identify the software to be used by each other to manage the transfer. The choice of software can have an impact on the description of the transfer procedures.

F-19 Write the transfer procedures: This step entails the writing of a description of the transfer procedures defined between the Archive and the Producer. This description will be part of the Submission Agreement.

3.2.2.4 Validation Definition

Table 3-23: Action Table for Formal Definition Phase: Validation Definition

Id	Formal Definition Phase: Validation Definition	Involves
F-20	Define immediate validation plan	Archive
F-21	Define an in-depth validation plan	Archive
F-22	Define the procedures for rejection, re-transfer, object acceptance (forms, anomaly forms, technical approvals, reviews, etc.)	Producer and Archive
F-23	Define the initial validation test	Producer and Archive
F-24	Identify the validation tools	Archive
F-25	Write a description of the validation procedures	Archive

F-20 Define immediate validation plan: Systematic validations are carried out in a systematic way at the time of object reception. In this case, errors lead to immediate rejection.

The Archive informs the Producer about the systematic validation carried out after reception. Important points to consider are as follows:

- Completeness (all the objects in the session have been correctly received).
- Integrity (the objects have not undergone any deterioration: checking with indicators such as volume).
- Conformity to the formal model. The objects delivered must correspond to the objects already identified in the model, and they must conform to the Data Dictionary (attributes).

F-21 Define an in-depth validation plan: A more in-depth level of validation, which depends on the quality required by the Archive, may be carried out later. In this case, a classification of non-conformities must be established.

In addition to systematic validation, this is a more in-depth validation of the data included in the SIPs, such as checking the coherence of the syntactic description of a file with respect to a described file, or checking the contents of text documents.

The Archive informs the Producer of the desired validation level, the necessary validation time (and the conditions for this validation to take place, in particular, the elements which must be present). These checks can concern objects delivered in different transfer sessions. The Archive can establish a validation classification.

The checks conducted automatically should be distinguished from those that are conducted manually. These checks can be carried out in a complete manner or random sampling:

- Automatic checks, such as:
  - Checking the structure of a document (e.g., table of contents, conformity to a DTD). This structure was defined during the Information Object definition phase.
  - Checking the structure of a data file with its syntactic description (e.g., EAST descriptor for a scientific data file).
- Manual checks:
  - Checking the intelligibility of document contents by partially or fully rereading (under no circumstances can the relevance and clarity of the semantic description of a file containing scientific observations be checked automatically).
  - Lastly, validation by experts representing the Designated Community should be considered. However, the feedback can reveal inadequacies in the data model and thus lead to changes. It is essential to ensure that all the information delivered, possibly supplemented by other information already held by the Archive, enable the AIP to be created containing all the required qualities from a Consumer point of view. The comprehensiveness and relevance of the information can only be determined by a peer review composed of experts and representatives of the Designated Community. The archivists may, if they consider it appropriate, invite the data Producer to this peer review.

F-22 Define the procedures for object acceptance: In each of these two previous cases (systematic and in-depth validation plan), the agreement or rejection procedures must be defined and approved by the Archive and the Producer.

The Archive and the Producer agree on the (total or partial) acceptance or (total or partial) rejection procedures of the session in the event of non-conformity with previous elements (e.g., anomaly forms, other forms). They also decide on the re-transfer procedures (and the deadlines). A technical report can close this phase. After these validations, the Archive can, for example, ask for modification of certain objects or complementary information.

F-23 Define the initial validation test: These tests both validate the data and ensure that the data is what should be transferred.

The Archive and Producer must:

- Define the test SIPs.
- Identify the kinds of tests:
  - Test the validation means (tools, procedures).
  - Test conformity to the test SIPs received. In the case of anomalies of the objects, the Archive alerts the Producer of the non-conformities. The Producer must correct the anomalies before the actual start-up of the deliveries.

F-24 Identify the validation tools: The Archive identifies and informs the Producer about the tools to be used for the validation. The Archive and Producer then discuss the possibility for the Producer to re-use these tools.

The Archive and the Producer identify the tools to be installed on both sides (some tools can be installed on the premises of the Producer so that validation can be carried out at that end. For example, a tool enabling a check of the compliance of an XML document with its DTD).

F-25 Write a description of the validation procedures: The procedure should cover all actions by both the Archive and the Producer. This description will be part of the Submission Agreement.

3.2.2.5 Delivery Schedule

Table 3-24: Action Table for Formal Definition Phase: Delivery Schedule

Id	Formal Definition Phase: Delivery Schedule	Involves
F-26	Define a reference delivery schedule (part of the final agreement)	Producer and Archive
F-27	Define the procedures to implement in the event of the schedule not being followed	Producer and Archive

F-26 Define a reference delivery schedule: Define a schedule with respect to the different objects or sets of objects that will be transferred. This schedule is an updated and completed version of the preliminary phase schedule. The type of elements delivered includes data files, descriptive files, timetables, and key dates.

F-27 Define the procedures to implement in the event of the schedule not being followed: The schedule must be regularly revised and the reasons for any divergence must be analyzed. The Producer and the Archive must specify the procedure to follow in the event of divergence.

3.2.2.6 Change Management After Completion of the Submission Agreement

It must be understood that the methodology presented in this Recommendation must not be interpreted statically. **This section takes into account changes that could occur after completion of the Submission Agreement.** This subsection takes into account changes that may be an upgrade request from the Archive or the Producer (for example, improvements in performance), as well as imposed changes (for example, data cannot continue to be produced because of a critical production failure).

The Producer and the Archive must agree to follow the change management process and to take into account the following actions in the future.

Table 3-25: Action Table for Formal Definition Phase: Change Management After Completion of the Submission Agreement

Id	Formal Definition Phase: Change Management After Completion of the Submission Agreement	Involves
F-28	Identify the origin (who) and the causes for the change	Producer and Archive
F-29	Identify the scenarios for managing the change	Producer and Archive
F-30	Assess the work to perform, the cost and the feasibility per	Producer and Archive

	scenario	
F-31	Make relevant decisions after discussion	Producer and Archive
F-32	Define and execute action plan	Producer and Archive

F-28 Identify the origin and the causes for the change: The Origin and the Causes of the Change Can be Numerous. The change may be requested by the Producer or by the Archive. Also, the origin of the change may be an evolution to an environment fully independent from the Producer and the Archive (for example, gradual obsolescence of a network technology or a media used for transfer).

The change may be temporary or definitive (for example, failure of a measurement device in a scientific experiment resulting in temporary or permanent stoppage of data production).

Several categories of causes of change may be defined:

– Infrastructure: the Producer and the Archive rely on a set of hardware, software and communication facilities. These facilities constantly evolve and may force the SIP delivery conditions to be modified.

– Information: the Producer may have to alter the schedule or extent of information to be delivered to the Archive. The information production may need to be stopped earlier than originally planned, or it may need to be extended. The nature of the information to be delivered may need to be altered. Information already delivered and validated may need to be revised and re-delivered due to any number of circumstances.

– Resources: the resources scheduled to perform the tasks defined in the formal agreement are no longer available.

– Legal: legal and formal aspects may need to be altered (for example, concerning copyright, change in ownership of rights to the information may restrict distribution by the Archive).

F-29 Identify the scenarios for managing the change: The Archive and the Producer must identify the possible scenarios for managing the change. Each scenario in the study should consider the entire ingest process and should include at least the following aspects:

– Impact on Data Objects:

  • impacts on the definition of objects to be delivered;

  • impacts on the formal model and the DED;

  • impacts on the volume of data to deliver;

  • impacts on objects already delivered.

– Impacts on the transfer procedure.

– Impacts on the validation procedure.

F-30 Assess the work to perform, the ensuing cost and the feasibility per scenario: The Producer and the Archive must assess the work to perform according to the previously identified scenarios. It should also include the impact on:

– The delivery schedule (and the frequency).

– Consumers (according to the schedule or the contents of the delivered Data Objects).

– The tooling.

– The human resources.

– The Archive in the long term.

This assessment results in a cost and feasibility study.

F-31 Make relevant decisions after discussion: The Archive and the Producer have in their possession the scenarios and their impacts for managing the change. The decision on how to proceed and the consequences on the Submission Agreement shall depend on the degree of severity of the change:

– A minor change will be taken into account without any modifications to the Submission Agreement.

– A more extensive change must be approved formally. The agreement may be the subject of a document that will be appended to the Submission Agreement, without this Agreement being fully renegotiated.

– A major change implies renegotiation of the Submission Agreement. There may be two outcomes to this renegotiation:

  • An agreement which may require that certain actions carried forward during the preliminary phase be produced again, and necessitate a modification to the Submission Agreement.

  • A disagreement which momentarily or definitively entails shutdown to the process.

F-32 Define and execute action plan: If the change is to be effectively taken into account, the Producer and the Archive must define the action plan to incorporate the change and must execute that plan.

3.2.2.7 Feasibility, Costs and Risks Assessment

Table 3-26: Action Table for Formal Definition Phase: Feasibility, Costs and Risks Assessment

Id	Formal Definition Phase: Feasibility, Costs and Risks	Involves

	Assessment	
F-33	Validate the project's feasibility	Producer and Archive
F-34	Assess the costs for the Archive and the Producer	Producer and Archive
F-35	Estimate the risks	Producer and Archive

F-33 Validate the project's feasibility: This step concerns the validation of the feasibility of the project, assessed in the preliminary phase.

F-34 Assess the costs: The Archive and the Producer must re-assess their costs separately (producing internal documents).

At this stage, the Archive must reexamine the points that only concern the Archive (see subsection 3.1.2.10, 'Permanent Impact on the Archive', as well as all tasks related to data ingest; see also subsection 3.1.2.4, 'Quantification').

F-35 Estimate the risks: The Archive and the Producer have to reexamine the risks estimated in the preliminary phase (see subsection 3.1.2.11, 'Summary of Costs, Risks'). Technical, financial, schedule, human and organizational aspects should be taken into account. The Archive and the Producer have to identify the actions necessary to minimize these risks.

3.2.3 SUBMISSION AGREEMENT

All of the elements resulting from this formal definition (Data Dictionary, model, etc.) must be approved by the Producer and the Archive.

Table 3-27: Action Table for Formal Definition Phase: Submission Agreement

Id	**Formal Definition Phase: Submission Agreement**	**Involves**
F-36	Draw up the Submission Agreement	Producer and/or Archive

F-36 Draw up the Submission Agreement: The formal definition phase is concluded by drawing up the Submission Agreement. This document is the result of all the preceding negotiations. It regroups all the textual descriptions for each of the paragraphs that make up the formal definition phase:

- information to be transferred (e.g., SIP contents, SIP packaging, data models, Designated Community, legal and contractual aspects);
- transfer definition (e.g. specification of the **Data Submission Sessions**);
- validation definition;

- change management (e.g. conditions for modification of the agreement, for breaking the agreement);
- schedule (submission timetable).

In some cases, there can be several 'Submission Agreements' between a Producer and an Archive, and these different agreements cover different and independent sets of information and result in several Producer-Archive Projects. When applying this methodology to subsequent Producer-Archive Projects, Submission Agreements associated with any previous Producer-Archive Project can be used to guide the completion of the new Submission Agreement.

Note that the Producer may not be able to agree on all planned data sets, but on sets or sub-sets of information, due to constraints linked to long term Data production (for example, the lack of resources may imply changes in data production).

3.3 TRANSFER PHASE

The aim of this phase is the actual transfer of the Data Objects between the Producer and the Archive.

Physical Objects may also need to be transferred, but the conditions of transfer and validation of the Physical Objects is outside the scope of this abstract standard.

During a Data Submission Session, one or more SIPs are delivered. The SIP is, in turn, composed of one or more digital Data Objects, the characteristics of which are described in the Data Dictionary.

Each object delivered is in reference to an object that has been previously identified with respect to a data model.

There is no sub-phase associated with the transfer phase. The subjects of the transfer phase are dealt with in a more precise way in the following subsections in the form of lists of actions to be carried out.

Table 3-28: Summary Table for Transfer Phase

Summary of the Transfer Phase
Action Table
Carry out the transfer test
Manage the transfer

3.3.1 CARRY OUT THE TRANSFER TEST

Table 3-29: Action Table for Transfer Phase: Carry Out the Transfer Test

Id	Transfer Phase: Carry Out the Transfer Test	Involves
T-1	Initial transfer test	Producer and Archive

T-1 Initial transfer test: To ensure full agreement on both sides, some initial submissions should be performed on the 'test data' before the beginning of the data delivery. These tests must be carried out as defined in action F17. After these tests have been carried out, the anomalies arising must be corrected and the operating parameters of the transfer must be adjusted. It can then be determined whether the differences between the performance shown and the expected performance require a review of the agreement or the schedule.

A test transfer may not be necessary for each new Submission Agreement. The Archives may not require a test transfer from a Producer with which the Archive has a good working relationship and has had no prior transfer or data validation problems.

All of these tests must be carried out before the start-up of the actual transfer operations.

3.3.2 MANAGE THE TRANSFER

Table 3-30: Action Table for Transfer Phase: Manage the Transfer

Id	Transfer Phase: Manage the Transfer	Involves
T-2	Ensure the proper execution of the data transfer operation from both the Producer and Archive sides	Producer and Archive

T-2 Ensure the proper execution of the data transfer operation: This phase consists of ensuring that the data transfer takes place correctly, both on the side of the Producer and the Archive:

- Adhering to the schedule for the Data Submission Sessions (transfer within planned time periods). This implies handling a timetable for transmissions from the Producer and for receptions by the Archive (e.g., progress indicators).
- The establishment and respect of procedures defined in the formal definition phase (e.g., session contents, packaging, media supports).
- Making sure that the operation runs well technically, including good network transmission (e.g., no cut-off, no transfer problems). This implies establishing a maintenance service to ensure the correct operation of the communication networks and to carry out appropriate actions in the event of failure.

- In the case of media transfers, making sure that the media sent by the Producer has been received by the Archive, that it has not been damaged, and that it is readable.
- Management of transmission anomalies, re-transfers.
- Sending acknowledgements of receipt per session by the Archive

In this phase the Archive and Producer should use the tools identified in the formal definition phase for the transfer.

3.4 VALIDATION PHASE

The aim of this phase is to carry out the validation of delivered objects, manage the anomalies detected, and accept all the objects transferred.

There is no sub-phase associated with the validation phase. The subjects of the validation phase are handled in a more precise manner in the following subsections in the form of lists of actions to be carried out.

Table 3-31: Summary Table for Validation Phase

Summary of the Validation Phase
Action Table
Carry out the validation test
Manage the validation

3.4.1 CARRY OUT THE VALIDATION TEST

Table 3-32: Action Table for Validation Phase: Carry Out the Validation Test

Id	Validation Phase: Carry Out the Validation Test	Involves
V-1	Initial validation test	Producer and Archive

V-1 Initial validation test: The tests must be carried out as defined in the formal definition phase:

- The initial test ensures full agreement on both sides. The systematic validation plan should be performed on ‘test data’ before the beginning of the data delivery.
- It should be taken into account that the validation tests are related to the types of information on which they are applied. These must be performed prior to the first deliveries of this information, and thus may be spread out in time, according to the

arranged schedule. In addition, the test phases may reappear in the course of time if new information categories are defined.

3.4.2 MANAGE THE VALIDATION

Table 3-33: Action Table for Validation Phase: Manage the Validation

Id	Validation Phase: Manage the Validation	Involves
V-2	Apply the validations	Archive
V-3	Manage the results of the validation	Producer and Archive

In this phase, the Archive should use the validation tools and processes identified in the formal definition phase.

V-2 Apply the validations: Check the conformity of the delivered objects with respect to the model of objects to be delivered and validate their contents. Two validation plans were identified in the formal definition phase:

– Systematic validation:

  • These validations are carried out after each transfer session.

  • At this stage, the Archive implements the systematic validation plan defined in the formal definition phase. In order to do this, the Archive must have installed the required tools.

  • All non-conformity, at this stage, implies rejection of the delivered objects during the session, and an anomaly form is sent to the Producer. The non-conformity is dealt with by both the Archive and Producer.

– In-depth validation:

  • These validations are not necessarily carried out in every session. They may be carried out when there is a coherent package of information, or at the end of the Producer-Archive Project when all the Data Objects are present. Some checks may require the presence of several files that are not necessarily delivered at the same time.

  • At this stage, the Archive carries out the checks defined in the in-depth validation plan in the formal definition phase.

  • The Archive must have already installed the required tools for the automatic checks.

V-3 Manage the results of the validation: This means that:

- the Archive identifies and sends out diagnostic and/or irregularity forms in accordance with the procedure defined in the formal definition phase;
- the Archive and the Producer manage the anomaly forms.

The Archive sends the Producer an acknowledgement that the Data Objects it has received have been validated and accepted (there may be a first level and then a second level agreement).

4 CREATING A PRODUCER-ARCHIVE INTERFACE METHODOLOGY COMMUNITY STANDARD FROM THE PRODUCER-ARCHIVE INTERFACE METHODOLOGY ABSTRACT STANDARD

4.1 PURPOSE

The purpose of this section is to define the rationale and expand on the approach for creating a Producer-Archive Interface Methodology Community Standard from the Abstract Standard, discussed in sections 1 through 3. As defined in subsection 1.4, this Community Standard will be conformant with the Abstract Standard if the following conditions are met:

- all of the actions have been considered and tailored as appropriate within the context of that community;
- the methodology for creating the Producer-Archive Interface Methodology Community Standard has addressed the various work phases defined in this section;

We recommend that this Abstract Standard be referenced from the Producer-Archive Interface Methodology Community Standard as providing the framework for the Community Standard.

NOTE – The term community is used here in a very broad and open sense. It could be a huge set such as the Archives, Producers and Consumers of scientific data files or document files for libraries. On the other hand, it could be limited to just one Archive and to the community of the information Producers related to this Archive.

Taking into account the specific features of the Producer-Archive community may give rise to a new standard. From this standard, when a large community is addressed, further tailoring could be used to create specific standards for sub-communities.

4.2 EXAMPLES OF CREATORS OF PRODUCER-ARCHIVE INTERFACE METHODOLOGY COMMUNITY STANDARDS

Defining the breadth of the community enables one to know who might undertake the task of creating a Producer-Archive Interface Methodology Community Standard.

According to the breadth of the community, this could for example include any of the following:

- National and international standardization bodies, which are usually organized and structured by grouping the players addressing a certain problem (e.g., ISO).
- National and international organizations of the community itself. This could be a regulatory organization with the role of coordinating activities of the community itself (e.g., the International Council on Archives [ICA]).

– An Archive that creates the implementation standard to be proposed to its information Producers.

The list shown above is merely an example and the purpose of this list is to show the different contexts in which a Producer-Archive Interface Methodology Community Standard may be created.

4.3 PHASES FOR DEVELOPING A COMMUNITY PRODUCER-ARCHIVE INTERFACE STANDARD

4.3.1 DEFINING TERMINOLOGY

This Abstract Standard has been drawn up with a neutral vocabulary defined for basic purposes in the OAIS Reference Model OAIS (reference [1]).

In order for the Producer-Archive Interface Methodology Community Standard to be used and easily understood by the players in the community, the vocabulary of the community itself should be used by the developers of the Community Standard whenever possible. Where there are multiple or ambiguous definitions and uses of a term, the Community Standards developers should adopt the terminology from the Abstract Standard.

It is advisable, but not mandatory, for the Community Standards developers to provide an equivalence table between the vocabulary of the Abstract Standard and the vocabulary of the community, as an annex.

4.3.2 INFORMATION MODEL OF THE COMMUNITY

The terminology must enable the Community Standards developers to define the main Information Objects of the community and the general attributes of the relevant Data Objects.

In addition to this terminology, the Community Standards developers must define the relationships among the objects, attributes and their behavior.

The development of the community model should lead to the creation of the Data Dictionary and the formal model needed for a Producer-Archive Project.

4.3.3 OTHER COMMUNITY STANDARDS

In addition to referencing this Recommendation, the Community Standards developers should identify and reference any standards related to or relevant to the Producer-Archive Interface Methodology Community Standard. The developers should also identify missing standards for purposes of targeting further development efforts.

4.3.4 COMMUNITY TOOLS

The Community Standards developers should identify community tools that may or must be used with regard to each of the phases in the process. These tools might include procedures, work instructions, metrification tools, standard value lists, and authoritative references.

4.3.5 ADDRESS THE ACTIONS DEFINED IN THE ABSTRACT STANDARD

The creator of the Community Standard must analyze each action defined in the abstract standard within the context of the community, and determine for each action whether it:

- can be applied as is to the Community's context;
- does not apply in the Community's context;
- applies but needs to be modified.

The developers may add actions specific to the community.

4.4 BEST PRACTICES FOR CREATING A COMMUNITY STANDARD

Begin by defining the community as broadly as is practical.

Include a diverse and representative membership to the committee writing the standard.

Publicize the work in progress, as appropriate, (e.g. on an existing or new community web site) in order to solicit diverse viewpoints and build community acceptance of the resulting standard.

Submit the draft Producer-Archive Interface Methodology Community Standard to a standardization body if appropriate.

ANNEX A

TARGETED OVERVIEW OF THE OPEN ARCHIVAL INFORMATION SYSTEMS (OAIS) REFERENCE MODEL DEDICATED TO THE METHODOLOGY ABSTRACT STANDARD

(This annex is **not** part of the Recommendation.)

A1 PURPOSE

The purpose of this annex is to provide a brief overview of the important terms and concepts, as defined in the OAIS Reference Model (reference [1]), necessary to understand this Producer-Archive Interface Methodology Abstract Standard. Readers are urged to read the full OAIS Reference Model Recommendation to fully understand the concepts.

The OAIS Reference Model is a framework for understanding and applying concepts necessary for long-term digital information preservation (where long-term is long enough to be concerned about changing technologies). It is also a starting point for a model addressing non-digital information. It does not specify any implementation.

A2 OPEN ARCHIVAL INFORMATION SYSTEM

A2.1 DEFINITION

What is meant by an 'Open Archival Information System?'

'Open' simply refers to the fact that this standard was developed in an open forum and is freely available.

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

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

There are many terms that need to be used in well-defined ways in order to construct a Reference Model. The OAIS Reference Model contains a glossary of these terms, and a few of the more important of these are defined below as needed.

A2.2 ENVIRONMENT MODEL

Figure A-1 depicts the OAIS as a box with three primary interfaces.

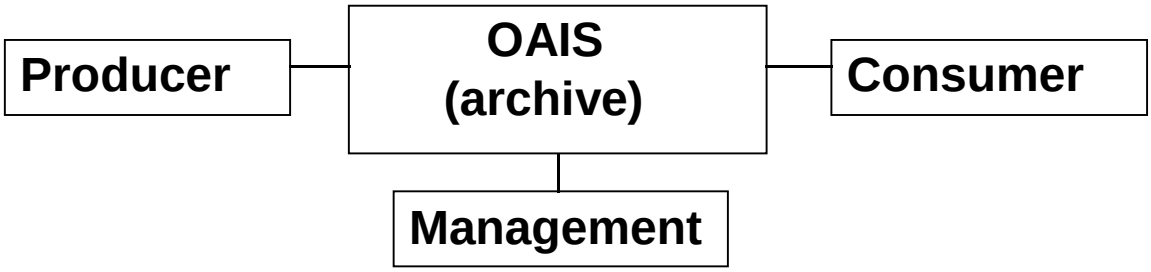

Figure A-1: OAIS Environment Model

In figure A-1, producers play the role of those who provide the information to be preserved. Management plays the role of those who set overall OAIS policy, where the OAIS is only one of its concerns. Day-to-day administration of the OAIS is handled by an Administration function within the OAIS box. Consumers play the role of those who interact with the OAIS services to find information of interest and to access this information.

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

A2.3 INFORMATION MODELING

As mentioned above, information is expressed by some type of data. It is the interpretation of the data, using additional Representation Information, that yields the information desired. This is depicted schematically in figure A-2. Consider a simple example to clarify the relationships.

Figure A-2: An Information Object

Consider a Data Object to be a particular string of 128 bits in a file. Given the information that these bits are to be interpreted by applying the ASCII standard, an understanding of the data (bit string) as a sequence of ASCII characters is obtained. This process has converted the Data Object (bit string), using the ASCII standard (Representation Information), into an Information Object that is more meaningful than the original bit string. Note that in order to preserve the Information Object, it is necessary to preserve not only the bit string, but also

the ASCII standard, which is the Representation Information, and the association between the two.

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

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

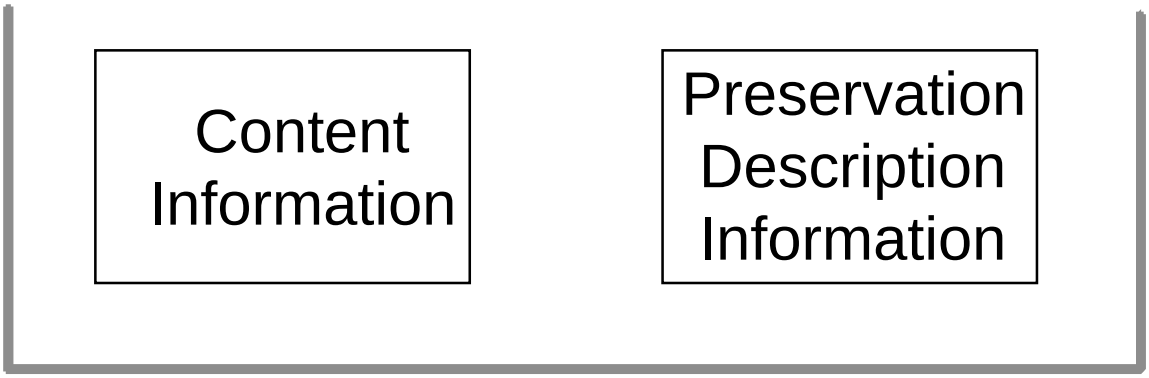

Figure A-3: Information Package Definition

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

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

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

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

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

- First, the Reference Information is used to provide one or more systems of identifiers by which to identify the Content Information. For example, this might include bibliographic attributes and/or a Digital Object Identifier.
- Second, the Provenance Information describes the history of the Content Information, including the chain of custody, so that Consumers can better judge how much to trust the information.
- Third, the Context Information relates the Content Information to other information outside the Information Package. This provides Consumers with an understanding of how the information being preserved relates to a wider environment.
- Fourth, the Fixity Information is used to help ensure that the Content Information is not altered in an undocumented manner. For example, this might include checksums and digital signatures.

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

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

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

The use of the three variants of an Information Package is shown in figure A-4.

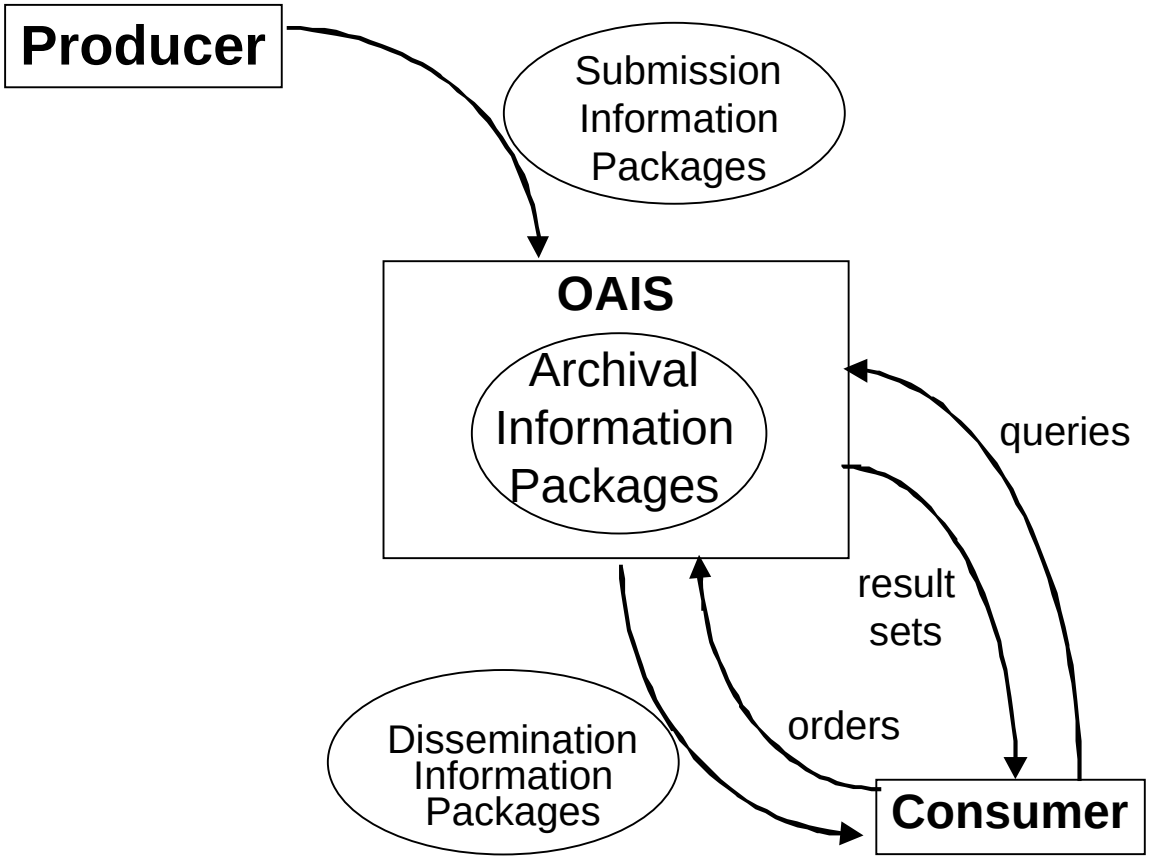

Figure A-4: External Data Flow View

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

The OAIS Reference Model goes into additional detail regarding the modeling of an AIP. It would not be appropriate to present all of this detail here, but some additional modeling is needed and is shown in figure A-5.

Figure A-5 is an example of the more formal modeling, using the Unified Modeling Language, of information in the OAIS as applied to the AIP.

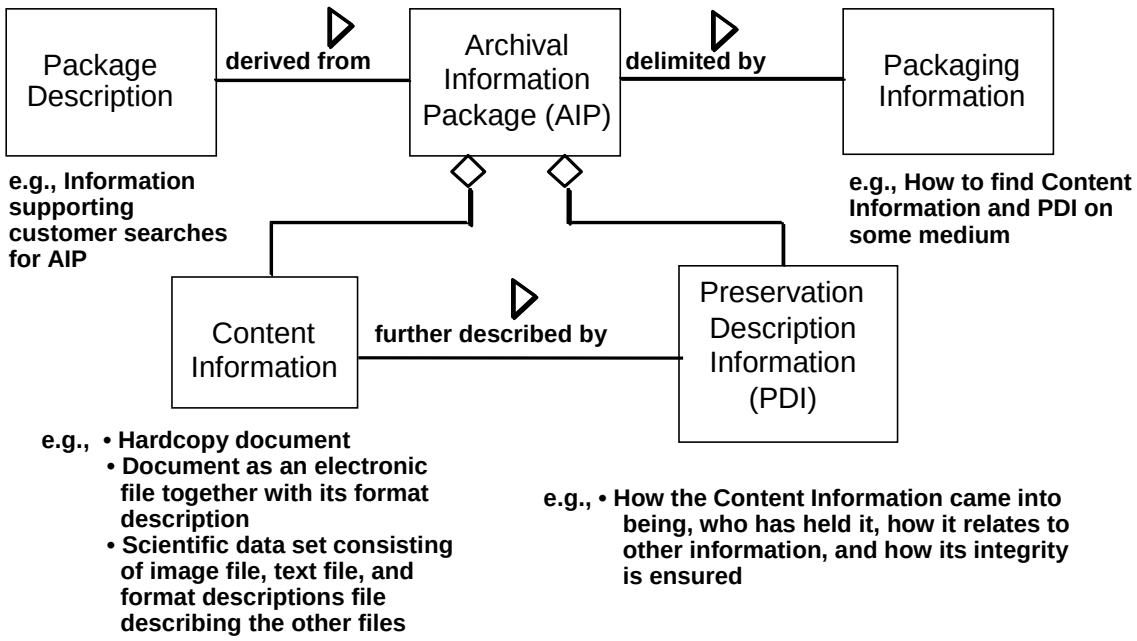

Figure A-5: Archival Information Package

The diamonds under the AIP box indicate that the AIP is a container holding two types of information: the Content Information and the Preservation Description Information. Examples of these types of information are given in the text below each of the boxes.

For example, the Content Information may be a hardcopy document, an electronic document with its Representation Information, or a set of files corresponding to a scientific data set with its Representation Information. Note that the Representation Information will include a format description, and may include additional semantic information such as that provided by a Data Dictionary. It is important for the OAIS to ensure that the Content Information and Preservation Description Information are understandable to the expected Consumer community. Such a community is referred to as the Designated Community for the given AIP.

What is new in this expanded view of an AIP are two additional types of associated information. The one on the right is called Packaging Information and it is used to bind the Content and PDI. The one on the left is called Package Description and it is used to support searching for the Content Information.

Packaging Information is the information that is used to logically, or actually, bind the Content Information and Preservation Description Information into a recognizable package with its constituent parts. It allows one to actually find the constituent parts on some media. It might be implemented using file systems, directory structures, pointers, and generic languages like XML.

The Package Description is used to hold the type of information needed by access aids, to support a Consumer's search for and retrieval of desired Content Information. It is most likely to be implemented in databases, and it is viewed as information that is most likely to

be updated over time. A card catalogue is an example. It is not critical for preservation because it can be regenerated, in principle, if needed.

Having looked at the information modeling aspects of the OAIS Reference Model, it is time to take a brief look at the modeling of archive functions.

A2.4 FUNCTIONAL MODELING

Six primary functions have been identified, as previously noted:

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

The conceptual relationships of the six functional areas, along with the three variations of Information Packages, are shown in figure A-6.

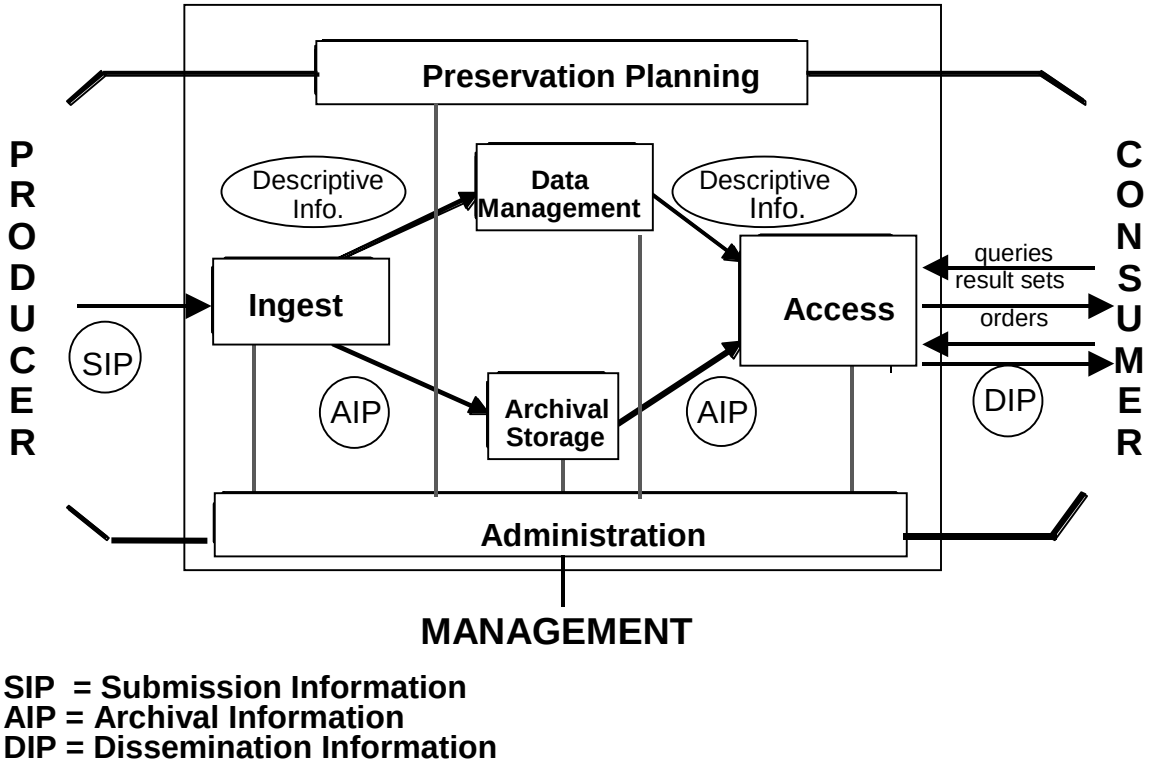

Figure A-6: OAIS Functional Entities

Figure A-6 may be understood as follows. Conceptually, a SIP is provided by a Producer to the Ingest entity. An AIP is created and delivered to Archival Storage. Related Descriptive Information is provided to Data Management. A Consumer searches for, and requests, information using appropriate Descriptive Information and access aids. The appropriate AIP is retrieved from Archival Storage and transformed by the Access entity into the appropriate Dissemination Information Package for delivery to the Consumer. This is all under the guidance of the Administration entity. Preservation strategies and techniques are recommended by Preservation Planning and put in place by the Administration entity.

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

To summarize, the OAIS Reference Model is applicable to all digital Archives, their Producers and Consumers.

It establishes common terms and concepts for comparing archival concepts and implementations, but it does not specify a particular implementation.

It identifies a minimum set of responsibilities that must be discharged for an Archive to call itself an OAIS Archive.

It provides detailed models for archival function and for the information associated with Archives.

Although not discussed in this annex, the OAIS Reference Model also provides perspectives on migration, emulation and interoperability among OAISs.

ANNEX B

INFORMATIVE REFERENCES

(This annex is not part of the Recommendation.)

[B1] *Unified Modeling Language*. Version 1.1. Cupertino, CA: Rational Software Corporation, September 1, 1997. <http://www.rational.com/uml/resources>.

[B2] *Data Entity Dictionary Specification Language (DEDSL)—XML/DTD Syntax (CCSD0013).* Recommendation for Space Data System Standards, CCSDS 647.3-B-1. Blue Book. Issue 1. Washington, D.C.: CCSDS, January 2002. [Equivalent to ISO 22643:2002]

[B3] *Data Entity Dictionary Specification Language (DEDSL)—PVL Syntax (CCSD0012).* Recommendation for Space Data System Standards, CCSDS 647.2-B-1. Blue Book. Issue 1. Washington, D.C.: CCSDS, June 2001. [Equivalent to ISO 21962:2002]

[B4] *The Data Description Language EAST Specification (CCSD0010).* Recommendation for Space Data System Standards, CCSDS 644.0-B-2. Blue Book. Issue 2. Washington, D.C.: CCSDS, November 2000. [Equivalent to ISO 15889:2003]

ANNEX C

LINKS BETWEEN PRELIMINARY AND FORMAL DEFINITION PHASE

(This annex is **not** part of the Recommendation.)

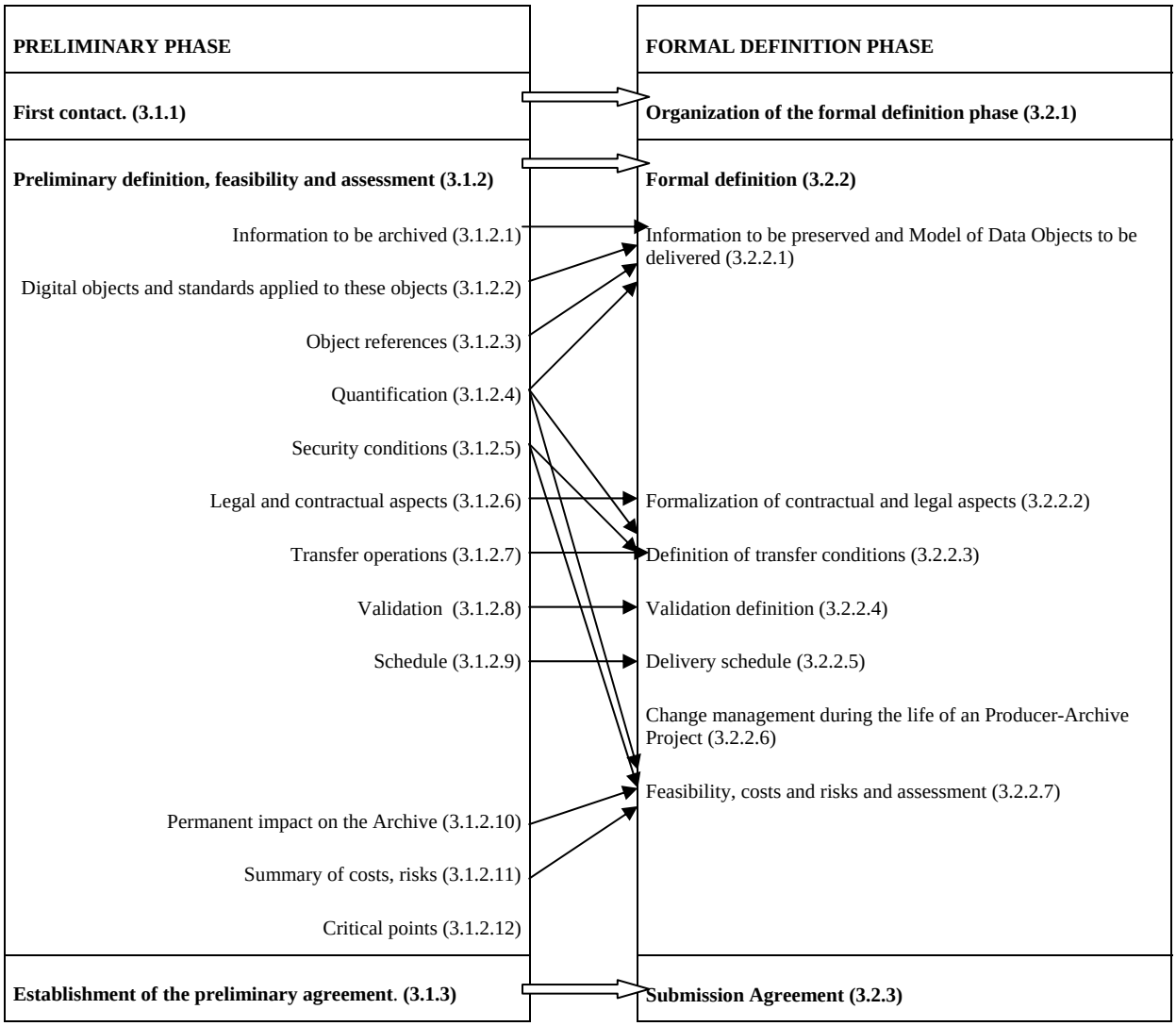

Figure C-1: Correspondence Between Preliminary and Formal Definition Phases

NOTE – In this table, the large open arrows describe the links between sub-phases levels. The fine arrows describe the links between groups of actions in a sub-phase.