

Draft Recommendation for  
Space Data System Practices

|  |
| --- |
| Information Preparation to Enable Long Term Use |

PROPOSED Draft Recommended Practice

CCSDS 653.0-W-0.12

WHITE Book

* April 2016

AUTHORITY

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
|  | Issue: | White Book, Issue 0.11 |  |
|  | Date: | April 2016 |  |
|  | Location: | Darmstadt, Germany |  |
|  | | | |

**(WHEN THIS RECOMMENDED PRACTICE IS FINALIZED, IT WILL CONTAIN THE FOLLOWING STATEMENT OF AUTHORITY:)**

This document has been approved for publication by the Management Council of the Consultative Committee for Space Data Systems (CCSDS) and represents the consensus technical agreement of the participating CCSDS Member Agencies. The procedure for review and authorization of CCSDS documents is detailed in *Organization and Processes for the Consultative Committee for Space Data Systems* (CCSDS A02.1-Y-4), and the record of Agency participation in the authorization of this document can be obtained from the CCSDS Secretariat at the address below.

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

FOREWORD

[Foreword text specific to this document goes here. The text below is boilerplate.]

Through the process of normal evolution, it is expected that expansion, deletion, or modification of this document may occur. This Recommended Practice is therefore subject to CCSDS document management and change control procedures, which are defined in *Organization and Processes for the Consultative Committee for Space Data Systems* (CCSDS A02.1-Y-4). 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 document should be addressed to the CCSDS Secretariat at the address indicated on page i.

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

Member Agencies

* Agenzia Spaziale Italiana (ASI)/Italy.
* Canadian Space Agency (CSA)/Canada.
* Centre National d’Etudes Spatiales (CNES)/France.
* China National Space Administration (CNSA)/People’s Republic of China.
* Deutsches Zentrum für Luft- und Raumfahrt (DLR)/Germany.
* European Space Agency (ESA)/Europe.
* Federal Space Agency (FSA)/Russian Federation.
* Instituto Nacional de Pesquisas Espaciais (INPE)/Brazil.
* Japan Aerospace Exploration Agency (JAXA)/Japan.
* National Aeronautics and Space Administration (NASA)/USA.
* UK Space Agency/United Kingdom.

Observer Agencies

* Austrian Space Agency (ASA)/Austria.
* Belgian Federal Science Policy Office (BFSPO)/Belgium.
* Central Research Institute of Machine Building (TsNIIMash)/Russian Federation.
* China Satellite Launch and Tracking Control General, Beijing Institute of Tracking and Telecommunications Technology (CLTC/BITTT)/China.
* Chinese Academy of Sciences (CAS)/China.
* Chinese Academy of Space Technology (CAST)/China.
* Commonwealth Scientific and Industrial Research Organization (CSIRO)/Australia.
* Danish National Space Center (DNSC)/Denmark.
* Departamento de Ciência e Tecnologia Aeroespacial (DCTA)/Brazil.
* European Organization for the Exploitation of Meteorological Satellites (EUMETSAT)/Europe.
* European Telecommunications Satellite Organization (EUTELSAT)/Europe.
* Geo-Informatics and Space Technology Development Agency (GISTDA)/Thailand.
* Hellenic National Space Committee (HNSC)/Greece.
* Indian Space Research Organization (ISRO)/India.
* Institute of Space Research (IKI)/Russian Federation.
* KFKI Research Institute for Particle & Nuclear Physics (KFKI)/Hungary.
* Korea Aerospace Research Institute (KARI)/Korea.
* Ministry of Communications (MOC)/Israel.
* National Institute of Information and Communications Technology (NICT)/Japan.
* National Oceanic and Atmospheric Administration (NOAA)/USA.
* National Space Agency of the Republic of Kazakhstan (NSARK)/Kazakhstan.
* National Space Organization (NSPO)/Chinese Taipei.
* Naval Center for Space Technology (NCST)/USA.
* Scientific and Technological Research Council of Turkey (TUBITAK)/Turkey.
* South African National Space Agency (SANSA)/Republic of South Africa.
* Space and Upper Atmosphere Research Commission (SUPARCO)/Pakistan.
* Swedish Space Corporation (SSC)/Sweden.
* Swiss Space Office (SSO)/Switzerland.
* United States Geological Survey (USGS)/USA.

PREFACE

This document is a draft CCSDS Recommended Practice. Its ‘White Book’ status indicates that its contents are not stable, and several iterations resulting in substantial technical changes are likely to occur before it is considered to be sufficiently mature to be released for review by the CCSDS Agencies.

Implementers are cautioned **not** to fabricate any final equipment in accordance with this document’s technical content.

DOCUMENT CONTROL

|  |  |  |  |
| --- | --- | --- | --- |
| **Document** | **Title and Issue** | **Date** | **Status** |
| CCSDS 000.0-W-0.1 | Information Curation Process, Proposed Draft Recommended Practice, Issue 0.1 | April 2014 | Original proposed draft |
| CCSDS 000.0-W-0.1a | Information Curation Process, Proposed Draft Recommended Practice, Issue 0.1a | June 2014 | Added Scope and Purpose Text, Import Abbreviations and Terminology |
| CCSDS 000.0-W-0.2 | Information Curation Process, Proposed Draft Recommended Practice, Issue 0.2 | September 2014 | Reworked Abbreviations from other standards. Still need to incorporate Purpose and Scope and Terminology from other documents. |
| CCSDS 653.0-W-0.3 | Information Lifecycle Framework, Proposed Draft Recommended Practice, Issue 0.3 | June 2015 | Renamed document, Entire document reworked to include only material from the agreed project description document. |
| CCSDS 653.0-W-0.4 | Information Lifecycle Framework, Proposed Draft Recommended Practice, Issue 0.4 | June 2015 | Expanded descriptions of Lifecycle stages. Lifecycle activities section added. |
| CCSDS 653.0-W-0.4 | Information Lifecycle Framework, Proposed Draft Recommended Practice, Issue 0.4-DG | August 2015 | Current draft. Entire document reworked. Activities removed, Topics added. |
| CCSDS 653.0-W-0.5 | Information Lifecycle and Long Term Usage, Proposed Draft Recommended Practice, Issue 0.5 | August 2015 | Renamed document. Merged 2 V0.4 version as agreed at telecon. |
| CCSDS 653.0-W-0.6 | Information Lifecycle and Long Term Usage, Proposed Draft Recommended Practice, Issue 0.6 | October 2015 | Updated Activities, Updated list of topics |
|  |  |  |  |
| CCSDS 653.0-W-0.7 | Information Lifecycle and Long Term Usage, Proposed Draft Recommended Practice, Issue 0.7 | November 2015 | Updates at CCSDS meeting. Remove detail section. |
| CCSDS 653.0-W-0.8 | Information Lifecycle and Long Term Usage, Proposed Draft Recommended Practice, Issue 0.8 | January 2016 | Updates following discussions at telecom |
| CCSDS 653.0-W-09 | Information Lifecycle and Long Term Usage, Proposed Draft Recommended Practice, Issue 0.8 | January 2016 | Updates following telecom |
| CCSDS 653.0-W-10 | Information Lifecycle and Long Term Usage, Proposed Draft Recommended Practice, Issue 0.10 | February 2016 | Updates based on telecons |
| CCSDS 653.0-W-11 | Information Preparation to Enable Long Term Use | March 2016 | Title changed to align with the scope and purpose of the document.  Updates based on telecons in March 2016 |
| CCSDS 653.0-W-12 | Information Preparation to Enable Long Term Use | April 2016 | Add details to placeholder sections |

CONTENTS

Section Page

[DOCUMENT CONTROL v](#_Toc449294437)

[CONTENTS vii](#_Toc449294438)

[1 Introduction 1](#_Toc449294439)

[1.1 purpose and scope 1](#_Toc449294440)

[1.1.1 Context 1](#_Toc449294441)

[1.2 applicability 3](#_Toc449294442)

[1.3 rationale 3](#_Toc449294443)

[1.4 conformance 4](#_Toc449294444)

[1.5 document structure 4](#_Toc449294445)

[1.6 definitions 4](#_Toc449294446)

[1.6.1 acronyms and abbreviations 4](#_Toc449294447)

[1.6.2 terminology 5](#_Toc449294448)

[1.7 NOMENCLATURE 6](#_Toc449294449)

[1.7.1 NORMATIVE TEXT 6](#_Toc449294450)

[1.7.2 INFORMATIVE TEXT 6](#_Toc449294451)

[1.8 References 6](#_Toc449294452)

[2 Overview 8](#_Toc449294454)

[2.1 Stages 8](#_Toc449294455)

[2.2 Topics 8](#_Toc449294456)

[2.3 ICP Participants 9](#_Toc449294457)

[3 INFORMATION TOPICS of Interest for Long-Term Perservation 10](#_Toc449294458)

[3.1 OAIS DEFINED Information Objects 11](#_Toc449294459)

[3.1.1 Content Information 11](#_Toc449294460)

[3.1.2 Preservation Description Information (PDI) 12](#_Toc449294461)

[3.1.3 Package Description Information. 13](#_Toc449294462)

[3.1.4 Packaging Information. 13](#_Toc449294463)

[3.1.5 Issues Outside the Information Model 13](#_Toc449294464)

[4 THE MAIN STAGES of the ICP and Information to be gathered 15](#_Toc449294465)

[4.1 Overview of Information ICP stages 15](#_Toc449294466)

[4.2 Details of the Stages 15](#_Toc449294467)

[4.2.1 FormulatION Stage 15](#_Toc449294468)

[4.2.2 ImplementATION Stage 15](#_Toc449294469)

[4.2.3 OperatION Stage 16](#_Toc449294470)

[4.2.4 INITIAL ExploitATION Stage 16](#_Toc449294471)

[5 ICP Framework - Activities detail 17](#_Toc449294472)

[6 ANNEX USE CASE: LTDP Workflow (Informative) 23](#_Toc449294475)

[7 ANNEX USE CASE: Large Project (Informative) 32](#_Toc449294483)

[7.1 Mapping to Stages 32](#_Toc449294484)

[7.1.1 Formulate Stage Example 32](#_Toc449294485)

[7.1.2 Implement Stage Example 32](#_Toc449294486)

[7.1.3 Operate Stage Example 33](#_Toc449294487)

[7.1.4 Exploit Stage Example 33](#_Toc449294488)

[7.2 Examples of Topics 33](#_Toc449294489)

[7.2.1 Data Objects Example: 33](#_Toc449294491)

[7.2.2 Representation Information Example: 34](#_Toc449294493)

[7.2.3 PDI Example: 35](#_Toc449294494)

[7.2.4 Reference Information Example: 35](#_Toc449294495)

[7.2.5 Provenance Example: 35](#_Toc449294496)

[7.2.6 Context Example: 35](#_Toc449294497)

[7.2.7 Fixity Example: 36](#_Toc449294498)

[7.2.8 Access Rights Information Example: 36](#_Toc449294499)

[7.2.9 Package Description Information Example: 36](#_Toc449294500)

[7.2.10 Packaging Information Example: 37](#_Toc449294501)

[8 ANNEX USE Case: Mapping to Generally Accepted RecordKeeping Practices (GARP) (Informative) 38](#_Toc449294516)

[9 ANNEX POTENTIAL USE Case (Informative) 40](#_Toc449294517)

[9.1 Small research project 40](#_Toc449294518)

[9.2 Large research project 40](#_Toc449294519)

[9.3 Aircraft Manufacture 40](#_Toc449294520)

[10 Informative References (Informative) 42](#_Toc449294521)

[11 Security Considerations (Informative) 43](#_Toc449294522)

[11.1 Introduction 43](#_Toc449294523)

[11.2 security concerns with respect to the CCSDS document 43](#_Toc449294524)

[11.3 Data privacy 43](#_Toc449294525)

[11.4 Data integrity 43](#_Toc449294526)

[11.5 Authentication of communicating entities 43](#_Toc449294527)

[11.6 Control of access to resources 43](#_Toc449294528)

[11.7 Availability of resources 43](#_Toc449294529)

[11.8 Auditing of resource usage 43](#_Toc449294530)

[11.9 Potential threats and attack scenarios 43](#_Toc449294531)

[11.10 Consequences of not applying security to the technology 43](#_Toc449294532)

# Introduction

## purpose and scope

There is a well-recognized need to capture digital information associated with a great variety of activities in virtually all recognized disciplines. The general term **Information Creation Project** (ICP) is used here for such an activity

Associated with such an Information Creation Project is a set of **Primary Data,** some of which may need long term preservation. This may be the main objective of the Project, such as with a satellite research effort to gain information about the Sun, or it may be a secondary objective such as the need to document the development of a new airplane to meet safety and maintainability issues. An ICP may create a collection of pre-existing information.

It is also well recognized that much of this information has long term value and needs to be preserved. Unfortunately it is often much less clear, for a given activity, what **Additional Information** is needed to ensure long term preservation and reuse/exploitation of the Primary Data. The purpose of this Recommended Practice is to provide guidance on this topic, building on the concepts and terminology provided by OAIS. It is assumed that the ICP will submit an appropriate subset of the Primary Data and Additional Information to an archive for long term preservation.

The Additional Information is unlikely to be known at the start of an ICP, and it is unwise to leave its collection to the end of the ICP. Therefore the ICP is conceived of as a set of stages, and within each stage this Recommended Practice identifies the Additional Information which should be collected, created or improved in order to be able to preserve and utilize/exploit (some part of) the Primary Data for the long-term.

This document should be of use to funders, researchers, information creators, archive managers and end-users by helping to increase the effectiveness of preservation activities and the exploitation of information. It should reduce the effort required by the archive if preservation related information has been gathered at the earlier stages.

This guidance can form the basis on which plans, including Data Management Plans, can be constructed, updated and monitored, to achieve the objectives noted above. It should be applicable to ICPs where the data already exists as well as where data is to be created in the future.

This Recommended Practice does not cover all aspects of an ICP. Aspects of the activities it does specify do not have to be carried out strictly sequentially, and indeed some may be revisited and improved at several of the stages.

The activities of the archive are described by OAIS. ICP aspects such as costing, risk management, policies and workflow, and service architectures are covered by the specific project plans. These are therefore not addressed here except at a high-level. Each aspect could be addressed in a variety of ways. It is expected that full treatment of these issues will require additional, more focused, standards.

While this recommendation originates in the space community, it is being designed in a generic way and should be applicable to any science domain and to the wider records management and archival communities, to information created in an individual project or, perhaps, by an organisation as a whole.

### Context

This Recommended Practice accomplishes the following:

* divides an ICP into 4 stages and identifies the Additional Information to be collected or improved at each stage;
* forms a basis for the specifications of Data Management Plans
* forms a basis for the identification and/or development of additional standards and implementation guides including those that address particular concerns in more detail;
* 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 of the ICP.

This Recommended Practice fits into the overall context defined by a number of other standards. Somesare

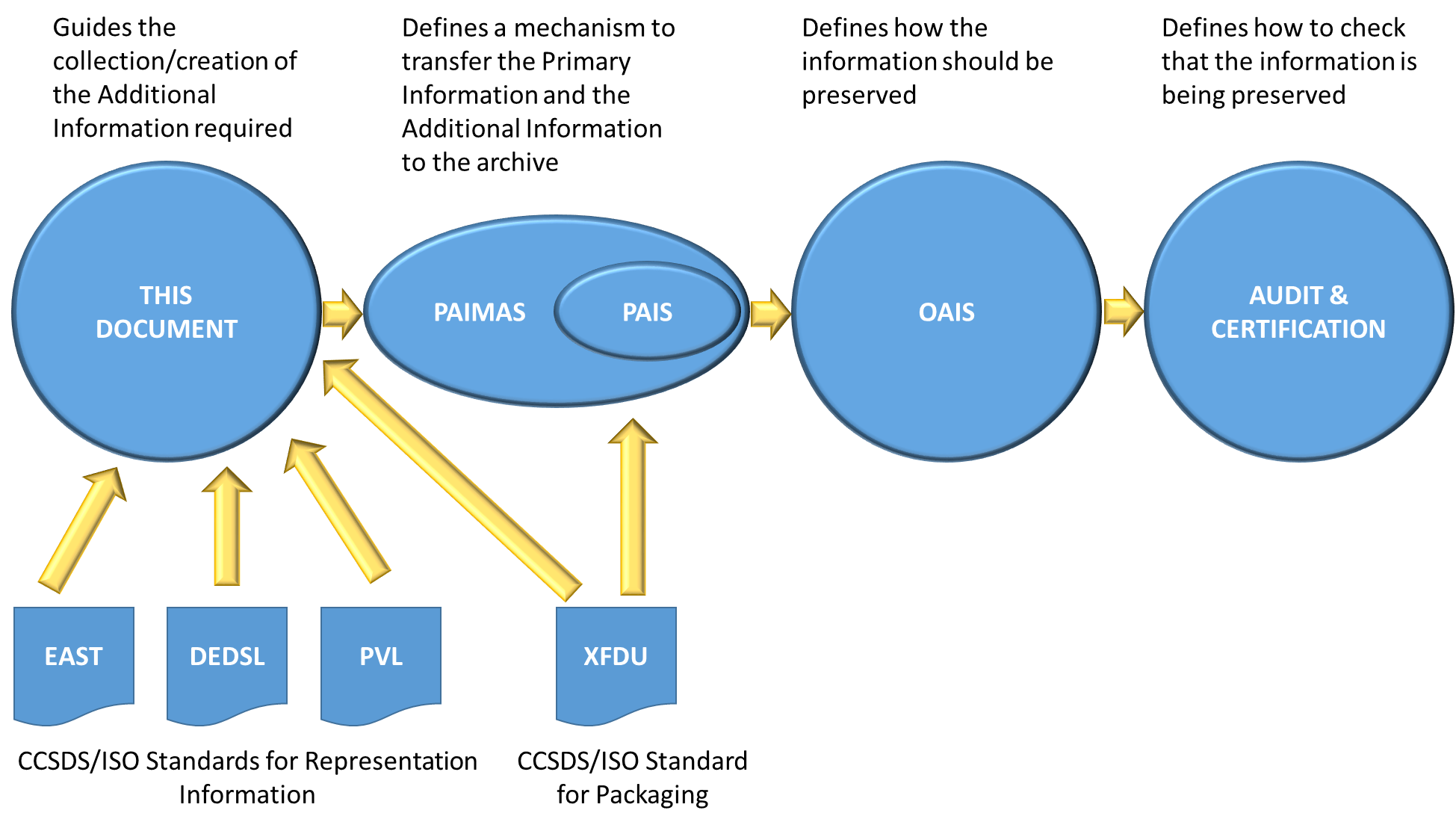


Figure 1‑1 Relationship between standards

OAIS [1] is one of the most widely recognized and applied archival standards available today. An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community.

PAIMAS [2] defines a methodology for transferring data from an Information Producer to an Archives based on the four following phases: Preliminary, Formal Definition, Transfer, Validation. Required activities during each phase are identified.

PAIS [3] provides the abstract syntax and an XML implementation of descriptions of data to be sent to an archive. These descriptions are negotiated agreements between the data Producer and the Archive and facilitate production of agreed data by the Producer and validation of received data by the Archive. The Recommended Standard includes an abstract syntax and one possible concrete implementation for the packages.

The Audit and Certification of Trustworthy Digital Repositories Recommended Practice [4] provides metrics for use in assessing the trustworthiness of digital repositories or archives.

In addition there are other CCSDS/ISO standards may be used to create Representation Information ([5],[6]and [7]) and also to package information [8]. There are many other techniques for creating Additional Information but these are outside the scope of this document.

In addition, the archival community has an existing, well established, set of concepts and terminology. The relationship with these and the OAIS concepts which underpin this document is described in Annex C

## applicability

The considerations/processes defined in this document apply to any activities producing information which is (or may be) re-used and preserved for significant periods. It is applicable, for example, to individuals who create information that may need long-term preservation and to organizations which have a mandate to make such information available for the long term.

## rationale

Data that is collected or created needs to have additional information associated with it if it is to be independently understandable, usable and trusted as being authentic. That additional information changes over time, as hardware, software, the general environment and users’ tacit knowledge changes. OAIS uses the terms Representation Information and Preservation Description Information for this associated information. It must be accumulated over the life of the ICP. For example Provenance Information will accumulate over time, recording the things which have happened to the data.

In the case of information created by individual projects, funders are increasingly asking that Data Management Plans accompany any request for project funding, however these tend not to evolve with the project and are difficult to monitor.

Many project models have been proposed. However they do not focus on the activities needed at each stage which will help to ensure that the data can be optimally exploited over the long term.

There are a small number of generally applicable stages in the ICP where, typically, the responsibility is handed on from one individual or team to another. Each of those individuals or teams has specific knowledge about the information which subsequent individuals or teams may not possess. Therefore there is a need to specify the information to be captured at each of those stages. Improvements may be needed to, for example, the Representation Information, which was recorded in an earlier stage; this may arise if the information is better understood or reformatted or re-processed in later stages.

Therefore there is a need for guidance as to what additional information should be captured or improved through the various stages of the ICP.

This document should enable:

* the Producer (including for example scientists who create the data) to capture and record the relevant information in a timely manner;
* the Archive to be assured that it will receive adequate information to enable it to perform preservation activities and support exploitation (e.g. re-use or secondary use) of the information
* the user to re-use information more easily
* the funder/sponsor to be assured that the resources that they contribute to the creation of the information will have suitable pay-back

## conformance

Conformance to this recommended practice requires that Additional Information defined here, appropriately mapped to the applicable terminology, be collected at the various stages identified.

Annexes A-??? provide examples of such mapping.

## document structure

Section 3 defines the topics about which information should be collected. Section 4 outlines the stages of the ICP and identifies the major pieces of information related to eventual re-use and exploitation which need to be collected while section 5 shows the way in which that information may evolve through the ICP.

## definitions

### acronyms and abbreviations

|  |  |
| --- | --- |
| **CCSDS** | Consultative Committee for Space Data Systems |
| **DMP** | Data Management Plan |
| **OAIS** | Open Archival Information System |
| **PAIMAS** | Producer-Archive Ingest Methodology Abstract Standard |
| **PAIS** | Producer-Archive Ingest Specification |
| **XML** | eXtensible Markup Language |

### terminology

Apart from the extra terms below, the definitions provided by OAIS and the other standards described in section 1.1.1. are used; these terms are normally capitalised, following the OAIS convention. It is assumed that the reader has some familiarity with OAIS.

**Additional Information**: The information which should accompany the Primary Data to ensure that it can be preserved and exploited. This will include Representation Information and Preservation Description Information (PDI), as defined by OAIS.

**Information Creation Project :** an activity planned and designed to achieve a particular aim ranging from the creation of new information to the preservation of existing information with a particular preservation aim. For example a project may involve research about a very specific topic over a few years involving one person, or it may be a large multi-national effort collecting information about many topics. Another example would be the manufacture of an airplane, where the primary goal is to create the plane but information such as designs, budgets, and test results are also created. In OAIS terminology the information from the ICP may go through intermediaries which play the role of Producer.

**Primary Data**: the data created by the Information Creation Project which, with the associated Additional Information, is of primary concern in terms of preservation and exploitation. For example this may be the data captured by a scientific satellite or the data associated with the design of an airplane. The Additional Information may itself require preservation and may be re-used, however it is of secondary concern and may be added to, as in the case of Provenance, or replaced, as in the case of Representation Information when the Primary Data is Transformed. All or part of the Primary Data will become OAIS Content Data Objects.

**Data Management Plan**: A data management plan or DMP is a document that describes how data will be handled throughout the project.

Note: OAIS provides the following definitions:

**Content Information:** A set of information that is the original target of preservation or that includes part or all of that information. It is an Information Object composed of its Content Data Object and its Representation Information.

**Data:** A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.

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

**Designated Community**: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time.

**Digital Object**:An objectcomposed of a set of bit sequences.

**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 the 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.

**Representation Information**:The information that maps a Data Object into more meaningful concepts.

## NOMENCLATURE

### NORMATIVE TEXT

The following conventions apply for the normative specifications in this Recommended Standard:

1. the words ‘shall’ and ‘must’ imply a binding and verifiable specification;
2. the word ‘should’ implies an optional, but desirable, specification;
3. the word ‘may’ implies an optional specification;
4. the words ‘is’, ‘are’, and ‘will’ imply statements of fact.

NOTE – These conventions do not imply constraints on diction in text that is clearly informative in nature.

### INFORMATIVE TEXT

In the normative sections of this document (sections 3 to 5), informative text is set off from the normative specifications either in notes or under one of the following subsection headings:

* Overview;
* Background;
* Rationale;
* Discussion;

## References

The following publications contain provisions which, through reference in this text, constitute provisions of this document. At the time of publication, the editions indicated were valid. All publications are subject to revision, and users of this document are encouraged to investigate the possibility of applying the most recent editions of the publications indicated below. The CCSDS Secretariat maintains a register of currently valid CCSDS publications.

1. *Reference Model for an Open Archival Information System (OAIS)*. Recommendation for Space Data System Practices, CCSDS 650.0-M-2. Blue Book. Issue 1. Washington, D.C.: CCSDS, June 2012. [Equivalent to ISO 14721:2012.] Available from: <http://public.ccsds.org/publications/archive/650x0m2.pdf>
2. *Producer-Archive Interface Methodology Abstract Standard*. Recommendation for Space Data System Practices, CCSDS 651.0-M-1. Magenta Book. Issue 1. Washington, D.C.: CCSDS, May 2004. [Equivalent to ISO 20652:2006.] Available from: <http://public.ccsds.org/publications/archive/651x0m1.pdf>
3. *Producer-Archive Ingest Specifications*. Recommendation for Space Data System Standards, CCSDS 651.1-B-1. Blue Book. Issue 1. Washington, D.C.: CCSDS, February 2014. [Equivalent to ISO 20104:2015] Available from: <http://public.ccsds.org/publications/archive/651x1b1.pdf>
4. *Audit and Certification of Trustworthy Digital Repositories*. Recommendation for Space Data System Practices, CCSDS 652.0-M-1. Magenta Book. Issue 1. Washington, D.C.: CCSDS, September 2011. [Equivalent to ISO 16363:2012.] Available from: <http://public.ccsds.org/publications/archive/652x0m1.pdf>
5. Parameter Value Language Specification (CCSD0006 and CCSD0008).. Blue Book. Issue 2. Washington, D.C.: CCSDS, June 2000. [Equivalent to ISO 21962:2003.]

Available from: <http://public.ccsds.org/publications/archive/641x0b2.pdf>

1. The Data Description Language EAST Specification (CCSD0010). Blue Book. Issue 3. Washington, D.C.: CCSDS, June 2010. [Equivalent to ISO 15889:2011.]

Available from <http://public.ccsds.org/publications/archive/644x0b3.pdf>

1. Data Entity Dictionary Specification Language (DEDSL)—XML/DTD Syntax (CCSD0013). Blue Book. Issue 1. Washington, D.C.: CCSDS, January 2002.. [Equivalent to ISO 15889:2011.]

Available from <http://public.ccsds.org/publications/archive/647x3b1.pdf>

1. XML Formatted Data Unit (XFDU) Structure and Construction Rules. Blue Book. Issue 1. Washington, D.C.: CCSDS, September 2008. [Equivalent to ISO 13527:2010.]

Available from <http://public.ccsds.org/publications/archive/661x0b1.pdf>

# Overview

This Recommended Practice deals with many aspects of an ICP. The terminology used is particularly important for effective communication. Many of these terms are already used within the target communities for this standard – e.g. space, science, records management and archival communities. It is expected that other communities can easily map this terminology to the terminology commonly used within those communities.

## Stages

The ICP activity is divided into four stages: **Formulation, Implementation, Operation and Initial Exploitation**. These are described in more detail in section 4.

* The Formulation stage conceives, justifies and seeks resources needed for the ICP.
* The Implementation stage designs and assembles the components needed for the ICP.
* The Operation stage carries out the information gathering, processing and analysis activities.
* The Initial Exploitation stage includes use (and re-use) to extract value from the outputs of the ICP for scientific, commercial, social or economic gain. For example this could involve publication and dissemination of the outputs of the ICP and finalisation of the Additional Information which should, after this stage, be handed over to the archives(s) to ensure long term preservation and dissemination.

Rationale:

Although small projects may involve only one individual through to the point of handing over the information to an archive, other projects may involve large teams which change through the ICP. Information needs to be passed between these teams. The four stages identified above provide a practical minimal subset of the possible stages; information would need to be collected in one stage to be passed to the next stage, and over time even in a single stage as personnel change.

At any of these stages the other stakeholders in the ICP may be consulted, including the archive(s) to which the information may eventually be sent for long term preservation and exploitation.

## Topics

The topics for which additional information is gathered largely follow the OAIS information model, supplemented by other topics. These are discussed in more detail in section 3.

Rationale:

OAIS defines the information objects required for Long-Term Preservation. All or part of these must therefore be created/collected through the ICP.

## ICP Participants

For completeness it is worth indicating the major participants’ roles. Note however that this document contains no further discussion of these roles, except in the Use Cases in Appendices A to TBD.

* The sponsor is an entity that provides funding and/or resources for a project.
* The entity that proposes and carries out information gathering projects.
* The archive is an entity that will eventually receive the information from projects, validates, preserves and provides it for long term use; its views may be taken into account by the other ICP participants
* The Consumer is an entity which uses and/or re-uses the information gathered. The archive guarantees that the Designated Community members will be able to understand and use the data. Other users may also be able to understand and use the data, the archive may not guarantee this but may do what it can to enable this use, for example by adding further Representation Information beyond that required by the Designated Community, in order to increase the exploitation of the data.

There are other supporting participants that are also involved in the ICP, including the standards organizations, software developers, interest groups and publishers. Standards organizations develop and support standards, software developers create tools which can be applied to the various activities of information gathering projects. Interest groups provide a forum for capturing and disseminating expertise across all ICP activities. Publishers provide standards for document submissions, arrange for peer review of submissions and publish and provide long term access to the published results of information gathering projects.

# INFORMATION TOPICS of Interest for Long-Term Perservation

The topics covered in this document are organized around the OAIS information model concepts, in particular the Archival Information Package (AIP) Information Model Components. An AIP should contain all the information required for long term usability and therefore this information must be collected in a timely way throughout the ICP.

OAIS contains the following diagram to show the various components of an AIP.



Figure 3‑1 Archival Information Package (Detailed View)

A archive must create AIPs as part of the preservation process. Many of these components may only be known by the participants in the Information Creation Project (ICP) and this Recommended Practice provides guidance for the ICP participants to help ensure that the information is captured, as part of the required Additional Information.

There are other pieces of information that is not covered by the scope of the OAIS Information Model but may be useful for those preparing to archive the information. These include

• the total volume of data

• ideas about the Designated Community

• ideas about the future ways in which the data may be exploited

These topics are discussed in more detail in the next section.

## OAIS DEFINED Information Objects

OAIS defines several major categories of information that make up the Archival Information Package: Content Information, Preservation Description Information (PDI) and Packaging Information. OAIS also defines Package Description Information, which is needed to provide visibility and access into the contents of an Archive, however is not required for the Long Term Preservation of the Content Information.

The next sections provide additional information about each of these topics.

### Content Information

Content Information includes the Data Objects as well as the Representation Information needed to understand and use the data objects. Representation Information is classified as Structure Information, Semantic Information and Other Representation Information. In broad terms Structure Information describes the physical layout of the data objects, Semantic Information describes the meaning of the values in the data object and Other Representation Information identifies other dependencies that need to be understood to use the data objects including software.

#### Data Objects

Data Objects are the data which will become the primary focus of preservation. Additional Information could include:

* Planned data rates
* Planned volumes of data
* Quality tests which may be performed on the data
* Information Properties which may be of use e.g. accuracy of the data values

#### Representation Information

The Representation Information includes

* structure,
* semantics
* other representation information such as analysis and display software.

In some ICPs the Representation Information may be captured in a number of formal documents. In others, especially those which extend over many years or even decades, there are likely to be a number of pieces of Representation Information which are not formally captured. For example there may be information which “everyone knows” such as:

* the way in which software libraries are named or organized
* the meaning of comments e.g. “will run on Cray-like machines” – may actually mean the software must be built on machines which use double-precision floating point numbers by default.
* Compiler bugs which must be worked-around
* The meaning of elements of the data header (if any)
* The location of documentation for proprietary systems

Each piece of Representation Information will consist of a Data Object and its Representation Information; each piece of this Representation Information will have its own Data Object and Representation Information, and so on. OAIS describes this as a Representation (Information) Network (RIN).

The amount of Representation Information which the archive will eventually require will depend upon the Designated Community which the archive serves. It may be useful to work with the archive to draft the RIN as early and in as much detail as possible.

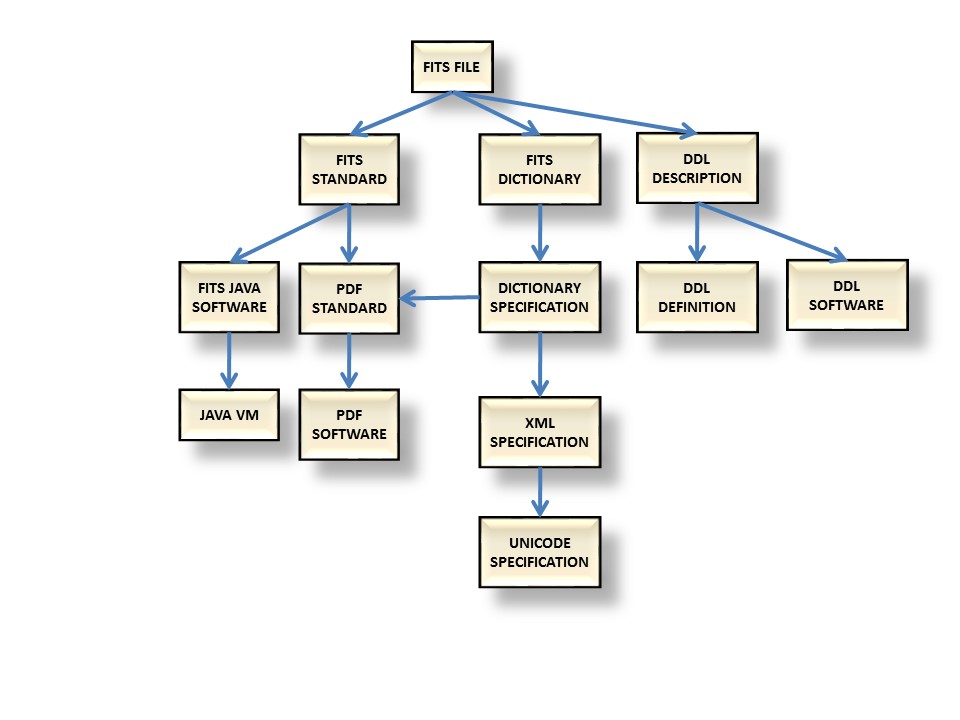


Figure 3‑2 Example of a RIN

### Preservation Description Information (PDI)

The PDI is information that is necessary to preserve the content information. It includes reference, provenance, context, fixity and access rights information.

#### Reference

Reference information provides a unique identification for each product. Useful Additional Information may include:

* Identifiers used in publications
* Naming conventions used in internal systems

#### Provenance

Provenance provides information including

* Specific aspects of the ICP origins and history,
* data custody – who was in control of the data at various stages,
* version control – what, if any, version control was used for the data,
* calibration and test - .

#### Context

Context information identifies or captures the knowledge that is needed to fully understand and interpret the project results. It includes background, publications and relationships.

* Broader aspects of the ICP origins and history
* Funders
* Current Research Information Systems (CRIS) information

#### Fixity

Fixity information allows verification of the integrity of data objects and could include:.

* Digests and Checksums – how they were calculated and where they are kept
* Description of how the digests are safeguarded.

#### Access Rights Information

Access rights including

* ownership,
* copyright and licensing or access restrictions.

### Package Description Information.

Descriptive information is used to provide a search capability to identify collections or products of interest. It includes finding aids and browse data. The archive must create appropriate Package Description Information. The ICP must provide information to allow the archive to do this.

### Packaging Information.

The Packaging Information is the information that is used to bind and identify the components of an Information Package.

The archive creates the AIPs and it is unlikely that the ICP will provide information to help in this.

### Issues Outside the Information Model

#### Publications

There may be many publications associated with the Primary Data including:

* research publications based on the data
* publications containing the data
* documents about the data – some of these documents may also be Representation Information

#### Related datasets

There may be many other data instances which may be related to the Primary Data (with its Additional Information) and which may aid in exploiting the Primary Data, for example

* data in the same discipline, for example astronomical data
* data in a complementary discipline, for example atomic spectral databases and astronomical data
* data about the same object, for example data measured at different wavelengths about a particular star

#### Potential other uses of the data

The Primary Data may have been created for a particular purpose, for example a particular research study or as a record of a step in a manufacturing process. The initial exploitation of the data may then be to produce a research paper or to prove the quality of manufacture.

The ICP may only be interested in, or may only have time of funding, to exploit the Primary Data in those ways.

However the ICP members may recognize that the Primary Data may have potential other uses. For example XXXXX

#### Suggestions about the appropriate Designated Community

The ICP may have some specific ideas about what Knowledge Base would be needed to understand and use the Primary Data, given the Representation Information which the ICP provides. For example there may be a general area of scientific expertise or a types of manufacturing process. This information could be useful for any archive which wishes to preserve and facilitate the exploitation of the Primary Data, given the Representation provided by the ICP.

# THE MAIN STAGES of the ICP and Information to be gathered

Section 4.1 outlines the stages of the ICP. Subsections 4.2 provide a brief description of each stage. The following subsections provide details of activities identified for each stage and relate them to the information topics identified in Section 3.

## Overview of Information ICP stages

The ICP activity is divided into four Stages, Formulation, Implementation, Operation and Initial exploitation.

Figure 4‑1 Stages of the Information Creation Project

Figure 4‑1above identifies the stages of the ICP:

## Details of the Stages

### FormulatION Stage

In the Formulation Stage the ICP is proposed, perhaps responding to solicitations and funding information available. At this stage it would be reasonable to expect

* the aims of the ICP to be clear enough to justify its funding;
* the way in which data would be collected and the kind of data to be collected would be known in general terms;
* the initial exploitation of the data would be outlined

The participants in the Formulation stage will almost certainly include sponsors and proposers and may also include data managers and Archives.

### ImplementATION Stage

In the Implementation Stage the preparations are made to collect or create data. This could include:

* the design and assembly of the components of the information system;
* the development or update of hardware and/or software systems;
* development of the associated processes;

The Implementation stage is performed by the project team.

### OperatION Stage

In the Operation Stage the activities are carried out which:

* should create or collect the data;
* process and analyse data.

The Operation stage is performed by the project team possibly with support from the archive. It operates the acquisition systems to gather data.

### INITIAL ExploitATION Stage

In the Initial Exploitation Stage the data can be exploited in various ways including:

* publication of research findings;
* generation of income
* exchange of social information
* predictions

There may also be ideas for exploitation in future.

The Initial Exploitation stage is performed by the project team to use/re-use and exploit the information and, if appropriate, prepare it for preservation and long-term preservation, re-use and exploitation.

# ICP Framework - Activities detail

The table below indicates the minimum useful status of information capture for each of the topical issues at each of the ICP stages. Typically information to address the issue and to document the decisions made in regard to each of these topical issues will begin to be accumulated early in the ICP. Then as time goes on more information is gained until the needed information is complete. In the case where new information about a topical issue will continue to be generated, then by late in the ICP, the collected information should be up to date. And even once complete, maintenance efforts and periodic reviews should be made to ensure that the information remains up to date to ensure that the data remains understandable as the Designated Community’s knowledge base changes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Topic** | **Issue** | **Formulation** | **Implementation** | **Operation** | **Initial Exploitation** |
| Content Data | Inventory of data produced/ expected | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Types of data (raw, processed, etc.) which should be preserved? | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Type of data e.g. images, tables – which generic interfaces? | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Volume that would require preservation | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Quality constraints | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Quality checks which may be performed on the data by non-experts | Rough idea | Increasingly detailed | Increasingly detailed | Complete |
| Representation Information | Choice of data format | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Format definitions and formal descriptions | Rough idea | Increasingly detailed | Becoming complete | Up to date and accumulating |
| Semantics of the data elements | Rough idea | Increasingly detailed | Becoming complete | Almost complete |
| Data dictionaries and other semantics | Rough idea | Increasingly detailed | Becoming complete | Up to date and accumulating |
| Information Model | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Other Data Documentation | Rough idea | Increasingly detailed | Becoming complete | Up to date and accumulating |
| Applicable standards | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Hardware and Software Dependencies | Rough idea | Increasingly detailed | Becoming complete | Up to date and accumulating |
| Other software which may be used on the data |  | Increasingly detailed | Increasingly detailed | Growing |
| What calibration and system test tools and system test data will be delivered. | Rough idea | Increasingly detailed | Becoming complete | Up to date and accumulating |
| Relationships between data items | Rough idea | Increasingly detailed | Complete | Complete |
| Reference Information | DOI or other unique identifiers | Rough idea | Becoming complete | Up to date and accumulating | Up to date and accumulating; New methods could be introduced |
| Rules, methods, tools for referencing data | Rough idea | Becoming complete | Up to date and accumulating | Up to date and accumulating; New methods could be introduced |
| What standards will be used to ~~format,~~ identify and reference the data and metadata | Rough idea | Becoming complete | Up to date and accumulating | Up to date and accumulating; New methods could be introduced |
| What may be used in future to identify the Information | Fairly firm | Increasingly detailed | Increasingly detailed | Evolving |
| Provenance Information | Record of origins of the project e.g. in a CRIS system | Fairly firm | Complete | Completed | Complete |
| Documentation about the hardware and software used to create the data, including a history of the changes in these over time |  | Rough Idea then Increasingly detailed | Becoming complete | Up to date and accumulating |
| Processing workflow | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Processing inputs |  | Rough Idea then Increasingly detailed | Becoming complete | Complete |
| Processing parameters | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Who was responsible for each stage of processing |  | Increasingly detailed | Becoming complete | Complete |
| When each stage was performed |  | Increasingly detailed | Becoming complete | Complete |
| Record of any special hardware needed | Rough idea | Increasingly detailed | Becoming complete | Complete |
| Calibration | Rough idea | Becoming complete | Complete | Complete |
| System Testing | Rough idea | Becoming complete | Up to date and accumulating | Up to date and accumulating; New methods could be introduced |
| Resident Archives |  |  | Rough idea | Becoming complete |
| Who was responsible for each stage of processing (Fixity) |  | Up to date and accumulating | Up to date and accumulating | Up to date and accumulating |
| Context | Outline of background concepts needed to understand the project | Rough idea | Increasingly detailed | Becoming complete | Up to date and accumulating |
| Publications related to the data (Risk) |  |  | Rough idea | Evolving |
| Publications related to the data (publications) | Rough idea | Increasingly detailed | Up to date and accumulating | Up to date and accumulating |
| Related data which may in the future be combined with this data |  | Increasingly detailed | Increasingly detailed | Growing |
| Potential Value of the data and likely business case for sustainability | Rough idea | Rough idea | Developing | Developing |
| Identification of archives which are likely to be able to host the data | Rough idea | Increasingly detailed | Increasingly detailed | Complete |
| Provide a bibliography of related publications | Rough idea | Increasingly detailed | Up to date and accumulating | Up to date and accumulating |
| Fixity | Fixity (e.g. CRC or digest) of data which may be preserved |  | Complete | Complete, but may Evolve | Complete, but may Evolve |
| How do we verify that all files are intact |  | Complete | Complete, but may Evolve | Complete, but may Evolve |
| Identify any special validation procedures that should be carried out. |  | Complete | Complete, but may Evolve | Complete, but may Evolve |
| Access Rights Information | What are the restrictions on access in the long term |  | Complete | Complete, but may Evolve | Complete, but may Evolve |
| Clear identification of Intellectual Property Rights |  | Complete | Complete, but may Evolve | Complete, but may Evolve |
| Licenses involved | Rough idea | Complete | Complete | Complete |
| Owners of the data – who can authorize hand-over | Rough idea | Complete | Complete | Complete |
| Who is the owner, what are the restrictions on access (licenses), what are intellectual property rights | Rough idea | Complete | Complete, but may Evolve | Complete, but may Evolve |
| Packaging Information | Details of the way components are packaged together for delivery to a repository |  | Increasingly detailed | Complete | Complete |
| Definition of SIPs |  | Developing | Complete | Complete |
|  |  |  |  |  |
| Descriptive Information | Methods for exploration/ quick-look at the data | Fairly firm | Increasingly detailed | Increasingly detailed | Evolving |
| Is browse data needed? | Fairly firm | Complete | Complete, but may Evolve | Complete, but may Evolve |
|  |  |  |  |  |
| Issues Outside the Information Model | Schedule of deliveries | Fairly firm | Increasingly detailed | Complete |  |
| Cost | Fairly firm | Increasingly detailed | Complete, but may Evolve | Complete, but may Evolve |
| Pointers to the components to be transferred to the archive |  | Fairly firm | Complete | Complete, but may Evolve |
| Potential preservation aims of the archive | Rough idea | Increasingly detailed | Increasingly detailed | Complete |
| Potential risks to preservation and exploitation of the data | Fairly firm | Increasingly detailed | Complete, but may Evolve | Complete, but may Evolve |
| What are the target archives and designated community for the solicitation. | Fairly firm | Complete | Complete, but may Evolve | Complete, but may Evolve |
| What is the budget for archiving. | Fairly firm | Complete | Complete, but may Evolve | Complete, but may Evolve |
| What is the schedule for major project milestones and deliveries to the archive. | Fairly firm | Complete | Complete, but may Evolve | Complete, but may Evolve |
| Change Management |  | Complete | Complete, but may Evolve | Complete, but may Evolve |
| What is the mechanism for communication between project and archive. | Fairly firm | Complete | Complete, but may Evolve | Complete, but may Evolve |
|  |  |  |  |  |

Table 5-1: Status of Information Capture for Topical Issues at ICP Stages

# ANNEX USE CASE: LTDP Workflow (Informative)

The following tables provides a mapping between the stages presented in this document with the stages described in the LTDP documentation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Proposed ILF Stage** | **LTDP  Workflow Phase** | **Activity** | **Description** | **Input** | **Output** |
| **Formulation** | **Implemenation** | **Data Set Appraisal** | An appraisal of the data set will provide an initial conception of whether the data set should be preserved and kept accessible and usable for the long term. Topics to be considered include mission relevance, economic considerations, temporal and geographical coverage, size, storage media and archiving format. The United States Geological Survey (USGS) provides helpful information for assessing the 'preservation value' of a data set (see link below). | Appraisal form Example: http://eros.usgs.gov/government/ratool/ | Data set appraisal (document), addressing at a minimum the aspects of the following topics, as proposed by the USGS: ·        Mission alignment with its own mandate, significance ·        General characteristics (including coverage, time span, completeness) ·        Access & distribution characteristics (including users, legal constraints, IP) ·        Physical characteristics (including media, volume, formats, processing level) ·        Metadata characteristics (including mission, sensor, calibration, processing information) ·        Economic characteristics (including preservation costs estimate, cost-benefit analysis) |
| **Formulation** | **Implemenation** | **Definition of designated community and preservation objective (Suggest breaking into 2 activities)** | Defining the designated community will help taking decisions during the preservation planning process. Data formats and access infrastructures may be adapted to the skills, resources and knowledge base that a community has access to. The community should be wide enough to allow for different levels of knowledge, applications and evolving user needs. The challenge lies in foreseeing a future user community and future uses of the data set. The designated community therefore should be re-assessed periodically, e.g. every ten years, to account for any changes in e.g. community composition or data use. The user community should be defined with sufficient detail to permit meaningful decisions to be made, regarding the composition of the data set to be preserved, and to allow derivation of requirements for effective re-use of the data.  The preservation objective can be derived from a dialog with the user community. It should define the level of use that an archive wishes to maintain for the Designated community. It may address topics such data discovery and access, or the provision of visualization, processing and analysis tools and infrastructure. | Data set appraisal (document) | Designated community definition (document) Preservation objective specification (document) addressing e.g.: • Intended use • Temporal scope • Data discovery and access • Visualization, processing, and analysis tools and infrastructure. |
| **Implemenation** | **Implemenation** | **Specification of preservation and curation requirements** | The preservation objective is translated into preservation requirements. These are more specific and may be based on user scenarios and use cases, possibly including detailed system requirements. Requirements for data value adding, across mission data set alignment, access, re-processing, or exploitation may also be included. | *EO Data preservation guidelines*  Designated community definition (document) Preservation objective specification (document) | Preservation and curation requirements specification (document) |
| **Implemenation** | **Implemenation** | **Definition of the consolidation ~~procedure~~ process (Tailoring rather than Definition used in diagram)** | The consolidation ~~procedure~~ process produces, from the input data records (L0 and auxiliary data), the corresponding, consolidated and validated data records, devoid of corrupted and duplicate files, aligned to the same naming convention and file format, and associated quality indicators. This process also impacts the services and functions which make the archival information holdings accessible to users, i.e. data search, discovery, retrieval, and use.  The *Generic EO Consolidation ~~Procedure~~ Process* document helps define a tailored procedure for the specific data records at hand. The tailored consolidation ~~procedure~~ process will be applied to the data records during the consolidation phase. | Preservation and curation requirements specification (document) *Generic EO Consolidation ~~procedure~~ Process* | Tailored consolidation ~~procedure~~ process (document) addressing the following topics, as specified in the *Generic EO Data Set Consolidation ~~procedure~~ Process*: • Data collection • Cleaning and pre-processing • Completeness analysis • Processing and re-processing |
| **Implemenation** | **Implemenation** | **Tailoring of preserved data set content and filling of the corresponding inventory table** | The *preserved data set content* document describes which data records and associated knowledge should be preserved in order to ensure long-term usability of the data set. The composition of the PDSC varies by sensor category and needs to be tailored for the specific data set at hand, taking into consideration the designated community, the preservation objective, requirements and dependencies, if any.  The data set manager should generate a preserved data set content inventory table ~~template is provided to assist the data manager in~~ to assess~~ing~~ which data records, information, and software is available and should be preserved. ~~It~~ The table facilitates ~~an~~ the assessment of completeness against the tailored preserved data set content document. For the data records, the information ~~is~~ should be collected at both collection level and at scene/pass level. The detailed scene/pass data inventory will be used in assessing spatial and temporal gaps in the data records.   Only items listed in this table will be preserved. The tailored and completed preserved data set content inventory table is therefore a critical document in the preservation process. | Designated community definition (document) Preservation objective specification (document) Preservation and curation requirements specification (document) Preserved data set content document ~~Preserved data set content inventory table~~ (Generated by this step. It cannot be an input.) | Tailored preserved data set content (document) - draft Tailored completed preserved data set content inventory (table) - draft |
| **Implemenation** | **Implemenation** | **Consultation and agreement with designated community** | The data set specific tailored PDSC should be discussed and agreed upon with the designated user community. The PDSC inventory table lists all the data set specific data records, information and software which are to be preserved for the future. Items not listed in this table will not be preserved. Hence, acceptance by the user community should be sought before continuing any further preservation activities. | Designated community definition (document) Preservation objective specification (document) Tailored preserved data set content (document) - draft Tailored completed preserved data set content inventory (table) - draft | Final preserved data set content (document)  Revised tailored completed preserved data set content inventory (table) |
| **Formulation Implemenation Operations Exploitation** | **Implemenation** | **Cost and risk assessment** | A cost and risk assessment should accompany the entire preservation process. Periodical re-assessment of both costs and risks helps identify and mitigate upcoming changes and hazards.  Risks assessed should include at minimum semantic risks, technical risks, organizational risks, resource risks and IPR related risks. An assessment of probability and severity/impact, together with a mitigation plan, should be prepared for each risk. The initial planning should extend at least 20 years into the future and be updated regularly - e.g. every 10 years. Since the temporal scope of the preservation activity extends over several decades, risks may change considerably. | Data set appraisal (document)  Preservation objective specification (document) Revised tailored completed preserved data set content inventory (table) | Cost assessment (document) addressing at minimum the following issues: • Updated preservation cost estimate from appraisal  • Resource planning (personnel, investments, operating expenses)  Risk assessment (document) addressing at minimum the following issues: • Risks: semantic, technical, organizational, resource, IPR related  • For each risk: probability, impact, severity, mitigation plan |
|  |  |  |  |  |  |
| **Implementation** | **Consolidation** | **Implementation of consolidation ~~procedure~~ process** | The tailored consolidation ~~procedure~~ process, defined earlier, is being implemented. As specified, the set of activities will be applied to the data records to be preserved. | Tailored consolidation ~~procedure~~ process (document) | Consolidated, clean data records, ready for preservation and user access |
| **Operations** | **Consolidation** | **Gathering of missing PDSC items and update of the PDSC table** | Compiling the knowledge associated with the data set to be preserved, i.e. information and tools, may continue and should be completed during the consolidation phase. The data set specific PDSC inventory table shall be finalized. | Final preserved data set content (document)  Revised tailored completed preserved data set content inventory (table) | Final tailored and completed preserved data set content inventory (table) |
|  |  |  |  |  |  |
| **Operations** | **Implementation** | **Data ingestion, master inventory generation, and catalogue population** | The data set to be preserved, i.e. the consolidated data records and the associated knowledge, are being ingested into the respective repositories. A master inventory should be generated and the catalogue should be populated in preparation for data dissemination. Ideally, these are done automatically during ingestion. | Consolidated data records Associated information (as specified in revised tailored completed preserved data set content inventory) Associated tools (as specified in revised tailored completed preserved data set content inventor) | Complete, consolidated data set (data records and associated knowledge) ingested into sustainable repositories Master inventory Preserved data set catalogue |
| **Operations** | **Implementation** | **Dissemination** | The data set, i.e. the data records and (specific) associated knowledge, is being made available to users, for discovery and retrieval. Providing tools for visualization, analysis, processing and/or corresponding exploitation infrastructure may be provided, as outlined in the Preservation Requirements Specifications. | Consolidated data records Preserved data set catalogue Preservation objective specification (document) Preservation requirements specification (document) | Online discovery and retrieval (download, ordering) of the data records and (selected) associated knowledge |
|  |  |  |  |  |  |
| **Operations** | **Operations** | **Operations and Maintenance** | The data sets, catalogue, and management inventories are being attended to. The archive, access, and management infrastructure is being operated, i.e. monitored for errors with corrective action taken in case of problems. ~~Hardware and software are~~ In accordance with the *EO Data Preservation Guidelines*, the infrastructure is being updated and migration activities are performed as required. In response to reprocessing requirements, e.g. resulting from a processing algorithm update, the preservation workflow may be re-initialized.  As the end of the preservation period, defined in the initialization phase, is approaching, a re-assessment of the preservation planning should be done in order to adjust preservation objectives and priorities. | Data and inventories Archive, access , and management infrastructure | Sustainable data preservation and access |
| **Exploitation** | **Operations** | **~~Curation and Stewardship~~ This activity was deleted** | In order to add value to or to improve accessibility and usability of the preserved data set, curation activities are being conducted. These may include an alignment to generate an across-mission time series, improving data citation and discovery by introducing persistent identifiers, or augmenting the metadata to facilitate content-based image retrieval and data mining. (Note: This paragraph is still included in the Operations section, but is not a description of an Operations Activity.) | Data and inventories Archive, access , and management infrastructure | Improved accessibility and usability |



# ANNEX USE CASE: Large Project (Informative)

The following provides example of the various topics and stages in the context of a large scientific project, for example to collect data from other planets using an orbiting satellite. It also suggests some more detailed guidance for the information to be collected.

## Mapping to Stages

### 

### 

### 

### 

## Examples of Topics

### Data Objects Example:

The data objects topic includes data objects, organization, delivery volume and processing parameters. In addition this may include high level format standards and data model where applicable (which will be supplemented by the detailed Representation Information). The data objects would be described at a high-level in the proposal or data management plan.

The project proposal should identify the types of data products that the project intends to produce in general terms (tables, images, maps).  The Project Data Management Plan should provide more detail, including actual format specifications where available.  If an existing standard is used then the documentation requirements are minimized, as the organization, standard name and version and URL can be referenced.  This will satisfy the need for Structure Information but not the need for Semantic Information.  If a standard format is not used then the project will need to provide documentation which describes the Structure Information.  There are data description languages (DFDL, EAST) and registration schemes (SFDU, XFDU) that can be used to define the explicit structure of data objects.  A Software Interface Specification (SIS) document is often used to provide an explicit description of the byte by byte structure as well as the interpretation of the values in the data object.  There are also data definition standards (DEDSL, ISO11179) which can be used to describe the meaning of data values in a standard way.  The use of these standards will promote interoperability between information systems.  The category Other Representation Information includes software, algorithms, encryption and written instructions.  These items should all be described in the SIS document.  If software is to be included as a deliverable it should follow the guidelines for submission to a public software repository (e.g. GITHUB).  Note that some archives don't accept software, so such a public repository may be the only way to provide software to future data users.

The data objects should be the same objects that are provided to the project team for analysis and should be delivered in the same format as used by the project. Thus it is important that the project understand the archive format requirements in the design stage to avoid having to transform the data prior to delivery. Such transformations are extremely risky and require an extra validation step. The combination of data objects and representation information should provide the capability to recreate any results cited in publications. If not then those products should also be delivered to the archive.

Possible guidelines for data formats.

* 1. Use the established format standards of the designated community throughout the data collection, processing and analysis activities. Wherever possible, use existing community and commercial tools to access and analyze data objects.
  2. Use open, registered, formally-documented formats with defined mime-types and standard file extensions that can be inspected with widely available tools.
  3. Avoid the use of machine or platform dependent data types, interleaving of logical objects and compression or encoding schemes.
  4. Use formats that contain embedded structural information required to view the data object as well as semantic information which identifies the format name and version as well as attributes necessary to interpret the object.
  5. Text-based formats (XML, JSON, CSV) for tabular data and simple binary arrays of 8 or 16 bit integers allow data inspection with common utilities.

### Representation Information Example:

In a large project Representation Information will be generated in the design and develop activities and included in the Interface Control Document (ICD). The structural information provides a description of the physical structure of each object type. If standard data formats are used then the documentation requirements may be minimized by referring to the documentation for the standard format or including it in the SIPs. For non-standard formats the ICD will provide a detailed description of every component of the data object, including the component name, location and data type. Providing a machine readable structure definition using a data description language such as EAST (ISO 15889) or DFDL will promote interoperability and reuse of the data. The semantic information describes the meaning of each data or metadata component. The use of Data Dictionary standards such as Metadata Registries (ISO 11179), Data Entity Dictionary Specification Language (ISO 21961) or the PREMIS Data Dictionary for Preservation Metadata combined with thorough and precise component descriptions will promote interoperability and reuse of the data. The ICD will document other types of representation information needed to access or interpret the data. These include software or hardware required for data processing, algorithms needed to convert data values to physical quantities, and encryption, encapsulation or compression techniques applied to the data.

### PDI Example:

Some of the PDI is described in the Data Management Plan and some in the Interface Control Document.

### Reference Information Example:

The ICD will describe the identification scheme. The reference identifier is often generated during the collection activity (e.g. a time tag, or sequential observation number) and passed along to the processing activity. Reference identifiers often provide the foundation for the directory and file names that are assigned to data objects and may be used to create URLs for accessing those files. ISBN numbers and Digital Object Identifiers (DOIs) are special forms of unique identifiers that can be assigned and registered for products that will be widely accessed.

* Unique Identifiers - Provide a scheme that will result in unique identifiers for all data components.

### Provenance Example:

The project history originates in the Project Description that is provided in the Proposal, DMP or other publications. The material from those sources will be transferred to the Interface Control Document. It may also include a record of commitments and an evaluation of project success. The ICD will record any changes in custody for any project components during the ICP. The ICD will describe the mechanism used to distinguish different versions of products and identify the different inputs and parameters that were used to produce them. The ICD will identify calibration and test procedures and assure that necessary software and data files are retained and delivered to the Archive.

* + History - Describe the history of the project, record of commitments, evaluation of project success.
  + Custody - Identify any changes in custody and their impact during collection, processing or analysis.
  + Version Control - Provide a mechanism to distinguish different versions of data components.
  + Calibration and Test - Provide relevant calibration and system test procedures, software and products.

### Context Example:

Some archives maintain databases of context information and have internal procedures for creating and maintaining these entries. Othewise it should be collected and transferred to the Interface Control Document. Background provides descriptive entries that are needed to understand the content information and can be likened to encyclopedia entries for pertinent topics. Publications includes references to all project and external documents that will be useful in understanding or interpreting the project data. Relationships identify other data collections which might aid in the interpretation or analysis of the project data.

* + Background - Describe all entities or concepts needed to interpret the project results.
  + Publications - Identify relevant publications which would help to understand the data.
  + Relationships - Identify external data collections related to the objects to be archived.

### Fixity Example:

Fixity measures will be described in the Interface Control Document. File creation date and time, file size, computed checksum and are commonly used for verifying the integrity of digital files. These values should be captured or applied during the collection and processing activities and maintained in product labels, logs or a database for eventual transfer to the archive. Care should be taken to avoid transfer mechanisms that change these values in an unexpected way, such as transfers that change directory or file date stamp or add or remove carriage control characters from files. Digital signatures are often used to verify the integrity of documents, especially those which deal with commitments of resources.

* Verification - Provide a mechanism to verify the integrity of project data and documentation.

### Access Rights Information Example:

Access rights will be specified in the Data Management Plan and also included in the Interface Control Document. These documents will identify the owner of the data, the terms of use, permissions or citations that are required and the individuals who can give permission to transfer custody, modify or delete the collection. They will identify any copyright or licensing issues that may be involved with any of the data, software or publications to be transferred to the archive. They will identify any access restrictions (security, government policies, proprietary period, privacy issues) that impact the handling of the data, software or documentation. They will identify how these restrictions apply to the transfer to the archive and to the handling of the data in the archive. They will identify whether the restrictions are applied to entire packages or individual components of the package. If restrictions are needed for specific components then the mechanism for conveying that information to the archive will need to be specified. Most archives do not have the capability to segregate protected content vs public domain content so may not allow on-line access to protected content.

* Ownership - Identify the ownership and the terms of use.
* Copyright and licensing - Identify any copyright or licensing issues.
* Access Restrictions - Identify any access restrictions.

### Package Description Information Example:

Sources of descriptive information will be identified in the Interface Control Document. The primary components used for finding aids are time tags, taxonomic categories (processing level, discipline, wavelength), observation intent, location (lat, lon, elevation), characteristics (resolution), conditions (temp, pressure, wind direction), instrument settings (filter, exposure, gain), data quality and data statistics. Browse products are special renditions of data objects like subsets or supersets that can be used to identify products of interest. Most archives will have requirements for at least high-level search parameters (e.g. Dublin Core) and taxonomic categories for important qualifiers (project name, discipline, instrument).

* Finding Aids - Provide descriptive parameters to support searching for products.
* Browse Data - Provide special renditions of data objects to facilitate browsing.

### Packaging Information Example:

At the point of handing over to an archive, the Packaging Information identifies the media and contents of the Submission Information Packages that are delivered to the Archive. The Packaging Information topics include the delivery format, manifest and delivery security. The Producer and Archive specify the mechanisms for SIP deliveries in the Data Management Plan. If there are no existing standards refer to the Producer-Archive Interface Methodology Abstract Standard (ISO 20652) for a list of steps that are performed in planning for and executing the delivery process. If the project follows archive guidelines for structuring the data collection then the packaging information might be as simple as providing a URL to the top folder of the SIP file structure on the project storage system. If multiple deliveries are scheduled or replacement versions of SIPs are submitted then additional information will be required to indicate how each delivery fits into the overall archive organization. The Producer-Archive Interface Specification (ISO 20104) provides the capability to model, execute, and validate the actual transfer of information from the producer to the archive. Every SIP delivery should be accompanied by a manifest which identifies the url or directory and file names of all the components of the SIP as well as validation parameters (checksum, size, date) for each component. The DMP will identify the security mechanisms (encryption, certificates, password protected accounts) that will be used for transfers between the Project and the Archive.

* Delivery Format - Identify the media and protocol to be used to deliver SIPs to the archive.
* Manifest - Provide a manifest which identifies all the components of the SIP.
* Delivery Security - Identify the security mechanisms for transfers between the Project and the Archive.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# ANNEX USE Case: Mapping to Generally Accepted RecordKeeping Practices (GARP) (Informative)

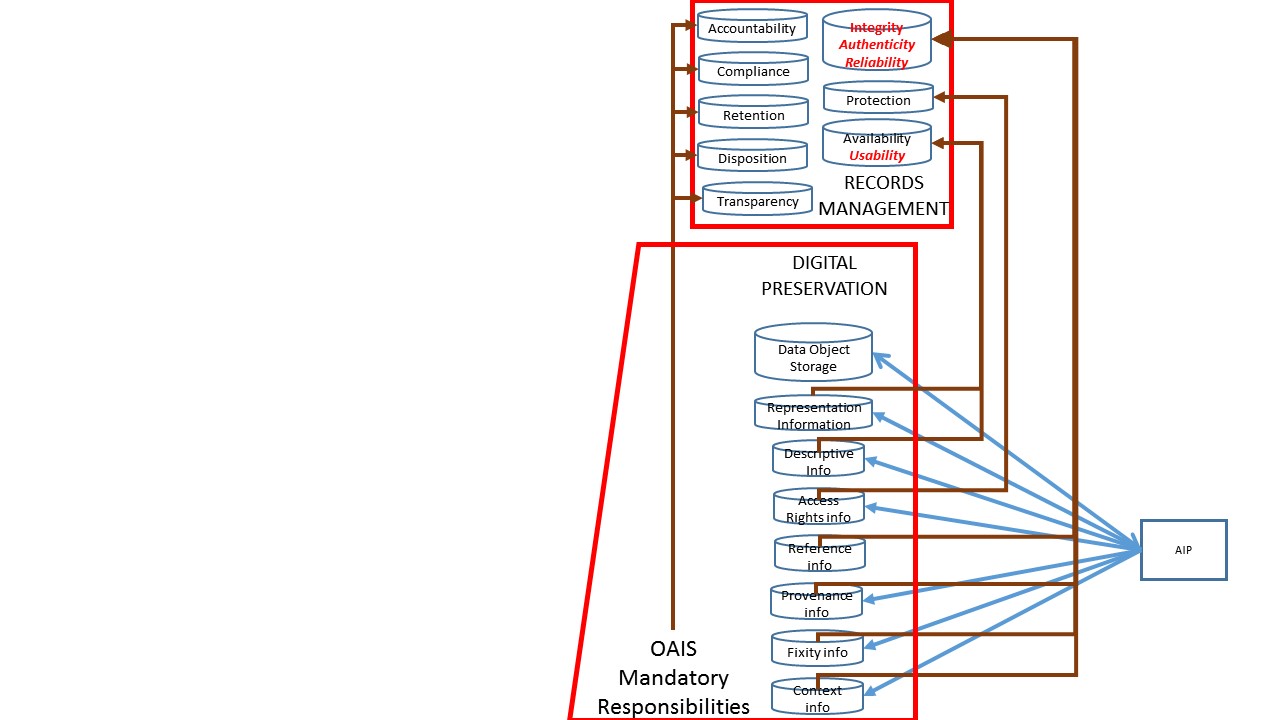


Figure C‑8‑1 Overview of Ingest and relationship between OAIS and Recordkeeping

Recordkeeping, as described in GARP is considered under eight headings, as indicated in the box labelled «Records Management». The information kept under each of these headings by archives tends to be rather general. OAIS tends to involve much more technical detail.

Figure C‑8‑1 indicates the linkages between OAIS and a number of aspects of Recordkeeping:

* **Integrity** (which includes authenticity and reliability) has contributions from:
  + OAIS Fixity Information to capture information such as hashes or digests.
  + OAIS Provenance Information
    - Providing details of who created the object and when, who had custody, who made changes to the object, when and how, how the object was transferred from one system to another or from one organisation to another, and what preservation activities were performed. It the digital object is changed through a Transformation, then the evidence which supports the claim of authenticity should also be captured.
      * There are many ways of recording provenance, and so each piece of provenance should be accompanied by Representation Information which allows one to understand the way in which this is encoded and its meaning.
  + OAIS Reference Information
    - Again this must have Representation Information
  + OAIS Context Information
* **Protection** has contributions from
  + OAIS Access Rights Information
* **Availability** (within which we include usability) has contributions from
  + OAIS Descriptive Information which are used to help discover the information
  + OAIS Representation Information

Note that OAIS allows one to think in more detail about each of the items. For example Fixity Information must include a value plus some Representation Information about that value i.e. how was it calculated and how the value is encoded. The same applies to Provenance and the other OAIS items of information.

The other aspects of RecordKeeping – accountability, compliance, retention, disposition and transparency are related to the OAIS Mandatory Responsibilities

# ANNEX POTENTIAL USE Case (Informative)

## Small research project

An individual researcher applies for funding to perform an experiment. The researcher is successful and sets up the experiment and data collection system and writes a Data Management Plan as required by the funders. He/she performs the experiment, thereby producing data. The researcher analyses the data using software he/she has created and publishes results.

He/she adds appropriate Representation Information and Preservation Description Information and transfers the information to an archive. The archive has defined a Designated Community for this information and ensures that it has adequate Representation Information. The archive may add Representation Information and Descriptive Information to in order that the data can be found, accesses and understood by a wider community beyond the Designated Community, in order to further exploit the data. .

## Large research project

A scientific instrument is required by a space agency for a satellite which is to study the Sun. Several multi-national consortia submit proposals. The proposal from one consortium is selected. The funding is obtained for the various consortium members from the various national funders.

The various consortium members undertake various tasks to build the instrument and the data collection system and associated software. Over a period of 10 years the instrument is built and integrated into the satellite. The satellite is launched and the instrument is deployed and collects data.

The data is collected at a ground station and sent to the researchers who are part of the instrument consortium. Modifications are made from time to time to the on-board software and the data processing software.

The Primary Data and Additional Information is sent to an archive for preservation and re-use by other researchers.

The Additional Information sent to the archive includes Representation Information such as the data format, semantics and processing software, which have been created by various members of the consortium. However many members of the consortium have moved on to other projects or have retired and some relevant information has been lost..

The archive adds Descriptive Information and Representation Information to help other communities, including some suggested by the instrument consortium, to exploit the information in other disciplines.

## Aircraft Manufacture

A large aircraft manufacturer wishes to create and sell a new type of aircraft. The initial design team creates a design which is tested and improved by a number of other specialist teams. A number of sub-systems, such as engines and wings, are sub-contracted to other specialist manufacturers. Over a period of 20 years the aircraft design goes through many stages. A great deal of information is collected to provide evidence for such things as safety and air-worthiness certificates.

The aircraft goes into production and is sold world-wide for the next 40 years.

The information that has been collected, including the design and the evidence about certification are legally required to be kept for 50 years beyond the time of manufacture of this model of aircraft. In addition the information can be used by the manufacturer to develop variants of the aircraft and also entirely new types of aircraft.

# Informative References (Informative)

1. GARP[[1]](#footnote-2) - Generally Accepted Recordkeeping Principles®
2. ESDIS Project. 2013. "NASA Earth Science Data Preservation Content Specification." NASA/GSFC. <https://earthdata.nasa.gov/files/423-SPEC-001_NASA%20ESD_Preservation_Spec_OriginalCh01_0.pdf>
3. LTDP documents

# Security Considerations (Informative)

## Introduction

The use of this Recommended Practice has a potential area of security concern, namely that in the case of data which should be confidential and its use restricted to a specific community, information is collected which allows that data to be found and used.

## security concerns with respect to the CCSDS document

This document provides guidance on additional information to be collected.

## Data privacy

The additional information may itself need to be subject to similar privacy considerations as the data being preserved and exploited.

## Data integrity

The additional information should itself be subject to the same consideration concerning preservation and authenticity as the data being preserved and exploited.

## Authentication of communicating entities

Authentication of communicating entities must be the responsibility of the individuals and organisations responsible to the data holdings and is not covered by this recommended practice.

## Control of access to resources

Control of access to resources must be the responsibility of the individuals and organisations responsible to the data holdings and is not covered by this recommended practice.

## Availability of resources

Availability of resources must be the responsibility of the individuals and organisations responsible to the data holdings and is not covered by this recommended practice.

## Auditing of resource usage

Auditing of resource usage must be the responsibility of the individuals and organisations responsible to the data holdings and is not covered by this recommended practice.

## Potential threats and attack scenarios

None.

## Consequences of not applying security to the technology

Consequences of not applying security to the data to which this recommended practice is applied will depend upon the sensitivity of the data being created/preserved.

1. http://www.arma.org/r2/generally-accepted-br-recordkeeping-principles. ARMA specify that the following citation is included:

   ARMA International (www.arma.org) is a not-for-profit professional association and the authority on information governance. Formed in 1955, ARMA International is the oldest and largest association for the information management profession with a current international membership of more than 10,000. It provides education, publications, and information on the efficient maintenance, retrieval, and preservation of vital information created in public and private organizations in all sectors of the economy. It also publishes Information Management magazine, and the Generally Accepted Recordkeeping Principles®. More information about the Principles can be found at www.arma.org/principles. [↑](#footnote-ref-2)