

Draft Recommendation for  
Space Data System Practices

|  |
| --- |
| Information Lifecycle And Long Term Usage |

PROPOSED Draft Recommended Practice

CCSDS 653.0-W-0.8

WHITE Book

* January 2016

AUTHORITY

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
|  | Issue: | White Book, Issue 0.7 |  |
|  | Date: | January 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 telecon |
|  |  |  |  |

CONTENTS

Section Page

[DOCUMENT CONTROL v](#_Toc440206550)

[CONTENTS vii](#_Toc440206551)

[1 Introduction 1-1](#_Toc440206552)

[1.1 purpose and scope 1-1](#_Toc440206553)

[1.2 applicability 1-2](#_Toc440206554)

[1.3 rationale 1-3](#_Toc440206555)

[1.4 conformance 1-4](#_Toc440206556)

[1.5 document structure 1-4](#_Toc440206557)

[1.6 definitions 1-4](#_Toc440206558)

[1.6.1 acronyms and abbreviations 1-4](#_Toc440206559)

[1.6.2 terminology 1-4](#_Toc440206560)

[1.7 NOMENCLATURE 1-5](#_Toc440206561)

[1.7.1 NORMATIVE TEXT 1-5](#_Toc440206562)

[1.7.2 INFORMATIVE TEXT 1-5](#_Toc440206563)

[1.8 References 1-5](#_Toc440206564)

[2 Overview 2-1](#_Toc440206565)

[3 INFORMATION TOPICS of Interest for Long-Term Perservation 3-1](#_Toc440206566)

[3.1 Information topiCS TABLE 3-1](#_Toc440206567)

[3.2 topiCS High Level Description 3-1](#_Toc440206568)

[3.2.1 Content Data 3-1](#_Toc440206569)

[3.2.2 Representation Information 3-3](#_Toc440206570)

[3.2.3 PDI – REFeFerence INformation 3-3](#_Toc440206571)

[3.2.4 PDI – Provenance Information 3-3](#_Toc440206572)

[3.2.5 PDI – Context Information 3-3](#_Toc440206573)

[3.2.6 PDI – Fixity Information 3-3](#_Toc440206574)

[3.2.7 PDI – ACCESS Rights INformation 3-3](#_Toc440206575)

[3.2.8 DESCRIPTIVE INFORMATION 3-3](#_Toc440206576)

[3.2.9 PACKAGING INFORMATION 3-3](#_Toc440206577)

[3.2.10 – Issues Outside the Information Model 3-3](#_Toc440206578)

[4 THE LIFECYCLE FRAMEWORK: THE MAIN STAGES and Information to be gathered 4-1](#_Toc440206579)

[4.1 Information Lifecycle stages 4-1](#_Toc440206580)

[4.1.1 Formulate Stage 4-2](#_Toc440206581)

[4.1.2 Implement Stage 4-2](#_Toc440206582)

[4.1.3 Operate Stage 4-2](#_Toc440206583)

[4.1.4 Exploit Stage 4-2](#_Toc440206584)

[5 LIfecycle Framework - Activities detail 5-4](#_Toc440206585)

[5.1 TBD 5-10](#_Toc440206586)

[6 [SECTION TITLE] 6-1](#_Toc440206587)

# Introduction

## purpose and scope

The purpose of this Recommended Practice is to provide guidance, building on OAIS concepts and terminology, for gathering information throughout the information lifecycle, from the proposal stage to the long-term re-use of the resulting information, focusing on the issues that need to be addressed at each stage which will help to ensure that the data can be optimally exploited over the long term. In this way information generated by a data producer can be properly preserved and be usable by designated communities and others into the future.

This document should be of use to funders, researchers, archive managers and end-users by helping to increase the effectiveness of preservation activities and the exploitation of data. It should reduce the effort required by the archive if preservation related information has been gathered at the earlier stages. The guidance can form the basis on which plans, including Data Management Plans, can be constructed, updated and monitored, to achieve this. It should be applicable to projects where the data already exists as well as where data is to be created in the future.

The Recommendation does not cover all aspects of the lifecycle and 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.

It will describe stages of the information lifecycle, and within each stage, this Recommended Practice identifies the information which should be collected, created or improved in order to be able to preserve and utilize information objects for the long-term.

This guidance takes the view that curation and preservation are not separate activities to be considered at the end of an information production project, but as a set of actions that must be conducted throughout the information lifecycle.

Other aspects, such as costing, risk management, metadata management, data formats, policies and workflow, provenance-adding and service architectures, which are clearly of interest in preservation and curation, are not addressed here except at a high-level, because 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:

* identifies the main stages in the information lifecycle and the information to be collected at each stage;
* forms a general methodological framework, which should be applicable and usable in any information stewardship, curation or preservation context (this general framework should provide sufficient flexibility to be applied to individual user’s situations);
* 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 lifecycle.

This Recommended Standard fits into the context of other standards as follows.

The OAIS Reference Model (see reference [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. It meets a set of such responsibilities as defined in this document. The model provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access, and for describing and comparing architectures and operations of existing and future archives. It also guides the identification and production of OAIS related standards.

The PAIMAS Standard (see reference [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.

The PAIS Recommended Standard (see reference [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 Recommended Practice (see reference [4]) provides metrics for use in assessing the trustworthiness of digital repositories and the Guidelines for Bodies Recommended Practice (see reference [5]) provides procedures to be followed when conducting audits of digital repositories using those metrics.

## 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 lifecycle of the data. 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 data lifecycles 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 information lifecycle 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 lifecycle.

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 unfamiliar 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 the terminology defined here, in terms of stages and information to be collected, can be mapped to existing practice.

## document structure

Section 3 defines the topics about which information should be collected. Section 4 outlines the stages of the lifecycle and identifies the major pieces of information related to curation which need to be collected while section 5 shows the way in which that information evolves through the lifecycle.

## definitions

### acronyms and abbreviations

|  |  |
| --- | --- |
| **ADMP** | Active Data Management Plan |
| **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

Unless defined 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..

**Project :** an activity planned and designed to achieve a particular aim ranging from the creation of new data to the preservation of existing data with a particular preservation aim.

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

## 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;
* Example.

## 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. *Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories*. Recommendation for Space Data System Practices, CCSDS 652.1-M-2. Magenta Book. Issue 2. Washington, D.C.: CCSDS, March 2014. [Equivalent to ISO 16919:2014.] Available from: <http://public.ccsds.org/publications/archive/652x1m2.pdf>

# Overview

This Recommended Practice deals with the aspects of the information and data lifecycle, in particular the terminology used. Many of these terms are already used with various definitions 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 used within those communities. The OAIS Reference Model provided a starting point and inputs from a variety of other sources were used to arrive at the terms used within this standard.

## Stages

The lifecycle is divided into four stages: formulation, implementation, operation and exploitation.

* The formulation stage solicits and funds information gathering projects.
* The implementation stage designs and assembles the components of the information system.
* The operation stage carries out the information gathering, processing and analysis activities, publishes results, and delivers project data and documentation to the archive.
* During the exploitation stage the archives receives and validates the project data and documentation, integrates it into the archive collection and provides for long term preservation and dissemination.

Rationale:

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

## Topics

The topics for which additional information is gathered largely follow the OAIS information model, supplemented by other topics.

Rationale:

OAIS defines the information objects required for Long-Term Preservation and which must therefore be created/collected through the lifecycle.

## Lifecycle Participants

The major participants’ roles include the sponsor, the project, the archive and the user.

* The sponsor is an entity that provides funding and/or resources for a project.
* The project is an entity that proposes and carries out information gathering projects.
* The archive is an entity that receives the information from projects, validates, preserves and provides it for long term use.
* The user 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 lifecycle, 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 lifecycle 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 a subset of the information from information gathering projects.

# INFORMATION TOPICS of Interest for Long-Term Perservation

In the following we use the term “project” to mean quite generally an activity planned and designed to achieve a particular aim ranging from the creation of new data to the preservation of existing data with a particular preservation aim.

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 lifecycle.

Table 3‑1 indicates the topics under which the curation information is grouped.

Table 3‑1 Curation information to be collected

|  |  |
| --- | --- |
| **Topic** | **Brief description** |
| **Content Data** | The data created or collected which may be re-used and preserved |
| **Representation Information** | See OAIS Representation Information |
| **Reference Information** | See OAIS Reference Information |
| **Provenance Information** | See OAIS Provenance |
| **Context Information** | See OAIS Representation Information |
| **Fixity Information** | See OAIS Fixity |
| **Access Rights Information** | See OAIS Access Rights |
| **Packaging Information** | See OAIS Packaging |
| **Descriptive Information** | See OAIS Descriptive Information |
| **Issues Outside the Information Model** | Issues that do not fit cleanly into the OAIS Information Model |

## 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.

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

The Representation Information includes structure, semantics and dependencies on other 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.

### 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.

Example:

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

#### Reference

Reference information provides a unique identification for each product.

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

Provenance provides information about the project history, data custody, version control, calibration and test.

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 lifecycle. 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

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

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

Fixity information allows verification of the integrity of data objects.

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

Access rights are related to ownership, copyright and licensing or access restrictions.

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.

Descriptive information is used to provide a search capability to identify collections or products of interest. It includes finding aids and browse data.

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.

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

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.

### Issues Outside the Information Model

\*\*\*

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

In the following we use the term “project” to mean quite generally an activity “planned and designed to achieve a particular aim ranging from the creation of new data to the preservation of existing data with a particular preservation aim.

Section 4.1 outlines the stages of the lifecycle. Subsections 4.2 to 4.5 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.

## Information Lifecycle stages

The lifecycle is divided into four stages, formulation, implementation, operation and exploitation. The formulation stage solicits and funds information gathering projects. The implementation stage designs and assembles the components of the information system. The operation stage carries out the information gathering, processing and analysis activities, publishes results, and delivers project data and documentation to the archive. The exploitation stage receives and validates the project data and documentation, integrates it into the archive collection and provides for long term preservation and dissemination. The sponsor is mainly involved in the formulation stage, though it should also be receiving regular status reports from the project and the archive during the other stages. The project is initiated in the formulation stage and has the lead role in the implementation and operation stages. The archive is mainly responsible for the exploitation stage, but provides inputs to the earlier stages as well.

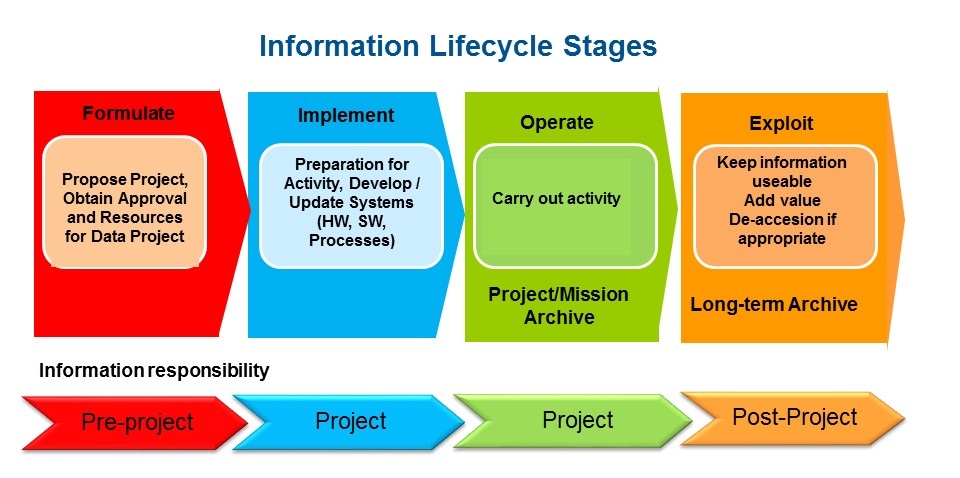


Figure 4‑1 Stages of the Information Curation Lifecycle

Figure 4‑1 above identifies the stages of the information lifecycle:

1. The Formulate stage is where those responsible for domains call for, review and chose among proposed projects to acquire or generate, manage or support exploitation of data or information gain approval and resources to carry out that activity and where those responsible for proposing the projects gain approval and resources to carry out that activity.

2) The Implement stage is where those responsible prepare to carry out the activity. The project develops (or updates) systems (hardware, software, data and processes to meet the needs of the proposed activity.

3) The Operate stage is where those responsible carry out the activity to acquire or generate, manage, and support exploitation of data or information from this activity

4) The Exploit stage is where those responsible maintain the data or information in a useable form and add Provenance if possible and proposed as part of activity. It also handles de-accessioning of the data if appropriate.

### Formulate Stage

The participants in the Formulate stage will almost certainly include Sponsors and Proposers and may also include Data Managers and Archives.

Example:

Sponsors typically have a well-defined charter that identifies their area of interest, within which, for example, programs are identified for specific research topics. Sponsors may bring in domain experts from existing projects and interest groups to develop strategic plans and objectives for the programs as well as long term schedules to fulfill the plans and objectives.

Proposers develop proposals that describe potential projects The proposals are likely to provide estimates of the cost and a preliminary risk assessment. The proposer may also prepare a data management plan that identifies the information gathering system, data processing system, and the types of products that will be produced. The data management plan will include estimates of data volumes and a schedule of activities and deliveries.

The proposer will often have to work with an archive to understand the standards in place and the mechanisms for delivering information objects. The archive may be a co-signatory for the proposal or at least the Data Management Plan. The archive uses the data management plan to develop a support plan that identifies the cost and resources that will be required to support the Project. These resource requirements need to be integrated into the project’s and the archive's long term budget and schedule.

Some projects may be able to use the Proposal and Data Management Plan as their system design. But for large and complex projects the selection is only the beginning of a substantial design process. We identify a System Design document to represent what may actually represent a sequence of requirements and design documents. The initial project design includes all elements of the acquisition, transfer, processing, distribution and analysis systems. For large projects there may be substantial negotiation and iteration between the sponsor and the project during the design process.

“Information produced during this stage provides a snapshot of the scientific and technical framework in which the project was born. Project … requirements, assessment studies, technology readiness reviews and cost analysis are performed during this stage. Preserving this information – both for approved and not approved projects – would allow future users to have reference material for new missions evaluation and definition. Traceability of this information is also useful to compare initial expectations to what was actually achieved by the mission and to understand which changes occurred between the pre-mission and the next stages.” (LTDP PDSC document).

### Implement Stage

The implement stage is performed by the project team.

Example:

Once the project proposal is approved, the project implementation team develops the information acquisition systems or provides access to existing data sources. This could be as complicated as building a spacecraft or as simple as accessing an online database. The project implementation team may provide command and control systems to interact with the information collection systems. The project implementation team develops processing software and scripts. It provides data processing and distribution systems which include analysis and access software. It documents the information content of all these systems via interface specifications, information models and data dictionaries.

“This stage produces the entire project and data detailed definition documents. It includes Sensor/Instrument requirements, characteristics, calibration methods, etc. Preserving this information is fundamental to understand changes that may have occurred over time while in operation. It is also needed to understand procedural impacts relative to instrument, algorithm and product implementation. Data acquired during the calibration and validation campaigns of the instrument under construction (e.g. in a laboratory or dedicated campaigns) is of critical importance as a reference for the future use of the data.” (LTDP PDSC document).

### Operate Stage

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

Example:

The project team receives and processes the data to support both command and control and data analysis. The project team maintains a record of all acquisition and processing activities that might be useful for understanding and interpreting the data. It performs data analysis and publishes results. The project team operates the implemented distribution mechanisms for transferring data to the project team and other users. The project team prepares Submission Information Packages (SIPs) for delivery to the long-term archive and participates in peer reviews.

### Exploit Stage

The exploit stage is performed by the long-term archive possibly with support from the project.

Example:

The archive receives and validates SIP deliveries from the project. It performs any conversion or consolidation necessary to integrate the data and metadata content of the SIPs into the archive collection. It transfers the data contents of the SIPs to the archive repository. It extracts metadata from the SIPs and loads it into the archive registry or database. It provides an access capability for users to be able to search, select and download both metadata and data from the archive collection. It provides for long term preservation by insuring the integrity, usability and accessibility of the data.

The differences between the operation and the exploitation stages are in the source of funding, the designated community and maturity of the products. The project and the archive may have completely separate funding sources. The project users may be limited to a small team whereas the archive may represent a large community. Special hardware and software capabilities may be provided to project teams but not to archive users. The archive may provide capabilities for exploiting the project data in concert with other data products.

# LIfecycle Framework - Activities detail

The table below indicates the minimum useful status of information capture for each of the topical issues at each of the lifecycle 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 lifecycle. 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 information lifecycle, 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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Topic** | **Issue** | **Formulation** | **Implementation** | **Operation** | **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 data | 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 (JGG for what?) | 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 Lifecycle Stages

1. Mapping to LTDP Workflow  
     
   (Informative)

[Annexes contain ancillary information. Normative annexes precede informative annexes. Informative references are placed in an informative annex. See CCSDS A20.0-Y-4, *CCSDS Publications Manual* (Yellow Book, Issue 4, April 2014) for discussion of the kinds of material contained in annexes.]

1. Mapping to TBD (Others besides LTDP)  
     
   (Informative)

[Annexes contain ancillary information. Normative annexes precede informative annexes. Informative references are placed in an informative annex. See CCSDS A20.0-Y-4, *CCSDS Publications Manual* (Yellow Book, Issue 4, April 2014) for discussion of the kinds of material contained in annexes.]

1. Security Considerations  
     
   (Informative)
   1. 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.

* 1. security concerns with respect to the CCSDS document

This document provides guidance on additional information to be collected.

* + 1. Data privacy

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

* + 1. Data integrity

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

* + 1. 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.

* + 1. 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.

* + 1. 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.

* + 1. 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.

* 1. Potential threats and attack scenarios

None.

* 1. 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.