

Draft Recommendation for
Space Data System Practices

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

PROPOSED Draft Recommended Practice

CCSDS 653.0-W-0.17

WHITE Book

* October 2016

AUTHORITY

|  |
| --- |
|  |
|  | Issue: | White Book, Issue 0.17 |  |
|  | Date: | October 2016 |  |
|  | Location: | Rome, Italy |  |
|  |

**(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 |
| CCSDS 653.0-W-13 | Information Preparation to Enable Long Term Use | May 2016 | Re-write using PMBOK as basis |
| CCSDS 653.0-W-14 | Information Preparation to Enable Long Term Use | June 2016 | Update based on comments from Mark Conrad and John Garrett and discussions during DAI meetings. |
| CCSDS 653.0-W-15 | Information Preparation to Enable Long Term Use | August 2016 | Updates based on inputs from DMBOK |
| CCSDS 653.0-W-16 | Information Preparation to Enable Long Term Use | Sept 2016 | Updates based on suggestions from DAI call 20160830 to clarify Collection Groups. |
| CCSDS 653.0-W-16JGG | Information Preparation to Enable Long Term Use | Sept 2016 | JGG Comments |
| CCSDS 653.0-W-17 | Information Preparation to Enable Long Term Use | Sept 2016 | Includes updates from the DAI WG meeting in Rome |

CONTENTS

Section Page

[DOCUMENT CONTROL v](#_Toc467487439)

[CONTENTS vii](#_Toc467487440)

[1 Introduction 1](#_Toc467487441)

[1.1 purpose and scope 1](#_Toc467487442)

[1.2 Context 1](#_Toc467487443)

[1.3 Applicability 3](#_Toc467487444)

[1.4 Rationale 3](#_Toc467487445)

[1.5 conformance 4](#_Toc467487446)

[1.6 document structure 4](#_Toc467487447)

[1.7 definitions 4](#_Toc467487448)

[1.7.1 acronyms and abbreviations 4](#_Toc467487449)

[1.7.2 terminology 5](#_Toc467487450)

[1.8 NOMENCLATURE 6](#_Toc467487451)

[1.8.1 NORMATIVE TEXT 6](#_Toc467487452)

[1.8.2 INFORMATIVE TEXT 6](#_Toc467487453)

[1.9 References 6](#_Toc467487454)

[2 Overview of the Framework 8](#_Toc467487455)

[2.1 Groupings of Activities 8](#_Toc467487456)

[2.1.1 PMBOK Process Groups 8](#_Toc467487457)

[2.1.2 DAMA-DMBOOK Activity Groups 9](#_Toc467487458)

[2.1.3 Collection Groups 9](#_Toc467487459)

[2.2 Areas 9](#_Toc467487460)

[2.2.1 PMBOK Knowledge Areas 9](#_Toc467487461)

[2.2.2 DMBOK Functions 10](#_Toc467487462)

[2.2.3 Additional Information Areas 10](#_Toc467487463)

[3 Collection Groups 11](#_Toc467487464)

[3.1 Overview of Collection Groups 11](#_Toc467487465)

[3.2 Details of the Collection groups 11](#_Toc467487466)

[3.2.1 The Initiating Collection Group 11](#_Toc467487467)

[3.2.2 The Planning Collection Group 12](#_Toc467487468)

[3.2.3 the Executing Collection Group 12](#_Toc467487469)

[3.2.4 the Closing Collection Group 12](#_Toc467487470)

[3.2.5 the Control Collection Group 13](#_Toc467487471)

[4 Additional Information Areas – to Ensure Long-Term Usability 14](#_Toc467487472)

[4.1 Information Areas Derived from OAIS DEFINED Information Objects 15](#_Toc467487473)

[4.1.1 Content Information 15](#_Toc467487474)

[4.1.2 Preservation Description Information (PDI) 17](#_Toc467487475)

[4.1.3 Package Description 18](#_Toc467487476)

[4.1.4 Packaging Information 18](#_Toc467487477)

[4.2 Information Areas Derived from Issues Outside the Information Model 18](#_Toc467487478)

[4.2.1 Publications 18](#_Toc467487479)

[4.2.2 Related datasets 18](#_Toc467487480)

[4.2.3 Potential other uses of the data 18](#_Toc467487481)

[4.2.4 Suggestions about the appropriate Designated Community 19](#_Toc467487482)

[4.3 Issues Suggested by PMBOK 19](#_Toc467487483)

[4.4 Issues Suggested by DMBOK 20](#_Toc467487484)

[5 Framework - Activities detail 21](#_Toc467487485)

[6 ANNEX USE CASE: LTDP Workflow (Informative) 28](#_Toc467487486)

[7 ANNEX USE CASE: Large Project (Informative) 36](#_Toc467487487)

[7.1 Mapping to Stages 36](#_Toc467487488)

[7.1.1 Formulate Stage Example 36](#_Toc467487489)

[7.1.2 Planning Example 36](#_Toc467487490)

[7.1.3 Executing Example 37](#_Toc467487491)

[7.1.4 Closing Example 37](#_Toc467487492)

[7.2 Examples of Topics 37](#_Toc467487493)

[7.2.1 Data Objects Example: 37](#_Toc467487494)

[7.2.2 Representation Information Example: 38](#_Toc467487495)

[7.2.3 PDI Example: 39](#_Toc467487496)

[7.2.4 Reference Information Example: 39](#_Toc467487497)

[7.2.5 Provenance Example: 39](#_Toc467487498)

[7.2.6 Context Example: 39](#_Toc467487499)

[7.2.7 Fixity Example: 40](#_Toc467487500)

[7.2.8 Access Rights Information Example: 40](#_Toc467487501)

[7.2.9 Package Description Information Example: 40](#_Toc467487502)

[7.2.10 Packaging Information Example: 41](#_Toc467487503)

[8 ANNEX USE Case: Mapping to Generally Accepted RecordKeeping Practices (GARP) (Informative) 42](#_Toc467487504)

[9 ANNEX POTENTIAL USE Case (Informative) 44](#_Toc467487505)

[9.1 Small research project 44](#_Toc467487506)

[9.2 Large research project 44](#_Toc467487507)

[9.3 Aircraft Manufacture 44](#_Toc467487508)

[10 Informative References (Informative) 46](#_Toc467487509)

[11 Security Considerations (Informative) 47](#_Toc467487510)

[11.1 Introduction 47](#_Toc467487511)

[11.2 security concerns with respect to the CCSDS document 47](#_Toc467487512)

[11.3 Data privacy 47](#_Toc467487513)

[11.4 Data integrity 47](#_Toc467487514)

[11.5 Authentication of communicating entities 47](#_Toc467487515)

[11.6 Control of access to resources 47](#_Toc467487516)

[11.7 Availability of resources 47](#_Toc467487517)

[11.8 Auditing of resource usage 47](#_Toc467487518)

[11.9 Potential threats and attack scenarios 47](#_Toc467487519)

[11.10 Consequences of not applying security to the technology 47](#_Toc467487520)

# Introduction

## purpose and scope

There is a well-recognized need to capture digital information associated with a great variety of endeavours in virtually all areas of society. However it is widely recognized that many such endeavours are not able, for one reason or another, to leave a sufficient legacy of information so others can reuse and fully leverage the effort that has gone into the endeavour. The purpose of this Recommended Practice is provide guidance for projects about the metadata (the term Additional Information is used below) that needs to be captured and/or generated, and retained in order to ensure that the information created by the project, either as part of its main objectives or as a by-product of achieving those objectives, can be exploited over the long term.

This Recommended Practice deals with the aspects of a project, 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 *Reference Model for an Open Archival Information System (OAIS)* provided a starting point and inputs from a variety of other sources were used to arrive at the terms used within this standard.

This Recommended Practice accomplishes the following:

* identifies the Additional Information to be collected or improved at various points;
* forms a basis for the specification 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;

## Context

This Recommended Practice extends the approach taken by the widely used *Project Managers Book of Knowledge (PMBOK) Guide* [1] and the related *Data Management Body of Knowledge (DMBOK)*[2][3].

The PMBOK [1] defines a project as an endeavour which is temporary, i.e. having a beginning and an end, undertaken to create a unique product, service or result and focusses on the information and techniques required to manage the project so that it achieves its objectives. The DMBOK [2][3] focuses on all aspects of data management within such an activity, while noting that “*Data, and information created from data, are now widely recognised as enterprise assets*”, and furthermore “*Data has value only when it is actually used, or can be useful in the future*”.

This document is focussed on the information that needs to be captured and/or generated and retained in order to ensure that the information created by the project, either as part of its main objectives or as a by-product of achieving those objectives, can be exploited over the short, medium and long term. It is expected that by collecting this Additional Information projects can significantly improve their information legacy to the benefit of the wider community.

At various times in the project, and for various reasons, data is captured or created. There is Additional Information associated with this data that also needs to be captured. The types of Additional Information are informed by the Open Archival Information System (OAIS) reference model that provides a conceptual view of long term information preservation in an archive.

This Recommended Practice fits into the overall context defined by a number of other standards. Some relationships between the documents are illustrated in Figure 1-1.



Figure ‑ Relationship between CCSDS standards

The OAIS Reference Model [4] 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.

The *Producer-Archive Interface Methodology Abstract Standard (PAIMAS)* [5] 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 *Producer-Archive Interface Specification (PAIS)* [6] 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 [7] provides metrics for use in assessing the trustworthiness of digital repositories or archives.

In addition there are other CCSDS/ISO standards that may be used to create Representation Information (the *Parameter Value Language (PVL)* [8], *the Data Description Language EAST Specification* [9] and the *Data Entity Dictionary Specification Language (DEDSL)* [10]) and also to package information the *XML Formatted Data Unit (XFDU)* [11]). 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 that underpin this document is described in Annex C.

## Applicability

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. It is applicable to information created by an individual, in an individual project or by an organisation as a whole. It is applicable to projects where the data already exists as well as projects where data is to be created in the future. It is also applicable to projects where the associated data is not the main focus of the project.

This document should be of use to funders, information creators, a role which may be played by multiple actors such as researchers or manufacturers, archive managers and end-users. It will be of use by helping to increase the effectiveness of preservation activities and the exploitation of information and by informing the practices and standards these users define in their communities.

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.

## 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. The amount and content of the 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 project. For example Provenance Information should originate at data creation and will accumulate over time, recording the things that 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. This standard encourages the active management of these plans to continue to address the communities’ needs and uses for the data. These Data Management Plans can also be captured as part of the Additional Information.

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

There are a small number of generally applicable processes in a project 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. There is a need to specify the information to be captured within and at the interfaces between each of those Project Phases. Improvements or changes to the Additional Information must be considered as the work proceeds. Therefore there is a need for guidance as to what Additional Information should be captured or improved through the various Project Phases.

This document should enable:

* the Producer 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 is collected as described in Section 5.

## document structure

Section 2 gives an overview of the document concepts which are expanded in the following sections. The Process Groups in the project are described in more detail in Section 3. Section 4 defines the areas about which information should be collected and identifies the major pieces of information related to eventual re-use and exploitation which need to be collected. Section 5 shows the way in which that information may evolve through the project.

## definitions

### acronyms and abbreviations

|  |  |
| --- | --- |
| **CCSDS** | Consultative Committee for Space Data Systems |
| **CRC** | Cyclic(al) Redundancy Check |
| **CRIS** |  |
| **DAMA** | Data Management Association International |
| **DEDSL** | Data Entity Dictionary Specification Language |
| **DMBOK** | Data Management Body of Knowledge |
| **DMP** | Data Management Plan |
| **EO** | Earth Observation |
| **ESDIS** | Earth Science Data and Information System |
| **GARP** | Generally Accepted Recordkeeping Principals |
| **ICD** | Interface Control Document |
| **LTDP** | Long-Term Data Preservation |
| **OAIS** | Open Archival Information System |
| **PAIMAS** | Producer-Archive Ingest Methodology Abstract Standard |
| **PAIS** | Producer-Archive Ingest Specification |
| **PDSC** | Preserved Data Set Content |
| **PMBOK** | Project Management Book of Knowledge |
| **PVL** | Parameter Value Language |
| **SIP** | Submission Information Package |
| **URL** | Uniform Resource Locater |
| **XFDU** | XML Formatted Data Unit |
| **XML** | eXtensible Markup Language |

### terminology

Apart from the extra terms below, the definitions provided by OAIS and the other standards described in section 1.2 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 Data to ensure that it can be preserved and exploited. This will include Representation Information and Preservation Description Information (PDI), as defined by OAIS.

**Data Management Plan**: A document that describes how Data will be handled throughout the project and what will happen to it when the project ends.

**Additional Information Area**: A concept or term about which information is needed to support long-term exploitation of data.

**Project Phase**: A collection of logically related project activities that culminates in the completion of one or more outputs.

Note: Here are a few selected definitions from OAIS:

**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 1 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. A Guide to the Project Management Body of Knowledge (PMBOK® Guide)—Fifth Edition, 2013, see <http://www.pmi.org/pmbok-guide-and-standards/pmbok-guide.aspx>
2. The DAMA Guide to the Data Management Body of Knowledge (DMBOK Guide) First Edition, 2009, <https://www.dama.org/content/body-knowledge>
3. DMBOK Version 2 see [http://dama-dach.org/dmbok2-DMBOK-version-2/](http://dama-dach.org/dmbok2-dama-dmbok-version-2/)
4. *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>
5. *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>
6. *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>
7. *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>
8. *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 of the Framework

The basic approach of PMBOK, DMBOK and this document is to identify a framework in the form of a matrix of groupings of activities/processes for a number of areas. This will act as a checklist to help to ensure that data is useable over the long term.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Activities→Areas↓ |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Section 2.1 describes the groupings of activities while section 2.2 describes the areas.

## Groupings of Activities

There are many system lifecycle descriptions [xxx] which the describe stages or phases of systems. The PMBOK [1] and the DMBOK [2][3] instead use a more flexible approach of describing groups of activities which are appear repeatedly in these various stages or phases.

PMBOK and DMBOK use slightly different groupings and terminology, the former uses the term Process Groups while the latter uses the term Activity Groups. The next sub-sections describe these two groupings plus the variant, termed Collection Groups, which best fits the requirements of this document.

### PMBOK Process Groups

PMBOK describes five Process Groups:

* **Initiating** - *consisting of processes performed to define a new project or new phase of an existing project by obtaining authorization to start the project or phase*
* **Planning** – *consisting of those processes performed to establish the total scope of the effort, define and refine the objectives, and develop the course of action required to attain those objectives*
* **Executing** – *consists of those processes performed to complete the work defined in the project management plan to satisfy the specifications*.
* **Closing** – *consists of those processes performed to conclude all activities across all Project Management Process Groups to formally complete the project, phase or contractual obligations*.
* **Control** – *consists of those processes required to track, review, and orchestrate the progress and performance of the project*.

These Process Groups are independent of application areas or industry. They are not Project Phases; all Process Groups may be conducted within a single phase; the individual Process Groups and processes are often iterated within a project and may not need to be performed in a particular order.

### DAMA-DMBOOK Activity Groups

DMBOK uses four Activity Groups which have obvious overlaps with PMBOK Process Groups:

* **Planning** – activities that set the strategic and tactical course for other data management activities. Planning Activities may be performed on a recurring basis.
* **Development** – activities undertaken within implementation projects and recognized as part of the systems development lifecycle (SDLC), creating data deliverables through analysis, design, building, testing, preparation, and deployment.
* **Control** – supervisory activities performed on an on-going basis.
* **Operational** – service and support activities performed on an on-going basis.

Each data management activity fits into one or more data management activity groups.

### Collection Groups

This document uses Collection Groups because the concern is to collect Additional Information about the data being created. The Collection Groups identify where Additional Information may be collected and is a specialisation of the PMBOK terminology, in particular including “Closing” because when the process to create data closes, steps must be taken to ensure its usability after the end of that process. The Collection Groups are:

* **Initiating** – justification for creating the data
* **Planning** – planning for the data creation and encoding
* **Executing** – creating/collecting/encoding the data. At each point there may be deviations from the planned results, including instrument effects and unexpected influences.
* **Closing** – completing the data creation/collection/encoding to satisfy the requirements of the project, phase or contractual obligations.
* **Control** - track, review, and orchestrate the progress and performance of the activities.

These will be discussed in section 3.

## Areas

The areas identify general headings which are or should be used in most projects most of the time. PMBOK and DMBOK provide lists of these areas in general terms. PMBOK uses the term **Knowledge Areas** while DMBOK currently uses the term **Functions**. This document uses the term **Additional Information Areas**.

### PMBOK Knowledge Areas

The PMBOK identifies ten Knowledge Areas including Project:

* Integration Management
* Scope Management
* Time Management
* Cost Management
* Quality Management
* Human Resource Management
* Communications Management
* Risk Management
* Procurement Management
* Stakeholder Management

These each represent a “*complete set of concepts, terms, and activities that make up a professional field, project management field, or area of specialisation. These ten areas Knowledge Areas are used in most projects most of the time.”*

### DMBOK Functions

DMBOK identifies ten functions:

* Data Governance
* Data Architecture Management
* Data Development (to be renamed in the next version [3] of DMBOK to Modelling and Design)
* Data Operations Management (to be renamed Data Storage and Operations)
* Data Security Management
* Reference and Master Data Management
* Data Warehousing and Business Intelligence Management
* Document and Content Management
* Meta-data Management
* Data Quality Management

The next version of DMBOK will add Data Integration and Interoperability and renames Functions to “Knowledge Areas”.

### Additional Information Areas

The Additional Information Areas of course have a strong relationship to the PMBOK and DMBOK Knowledge Areas/Functions. PMBOK and DMBOK deal with all aspects of a project but this document is concerned only with those aspects which ensure long term usability of the data.

PMBOK Knowledge Areas of Scope, Cost, Quality, Risk and Stakeholder Management have relevance as do almost all the DMBOK functions.

The DMBOK Metadata Management function clearly is of particular significance but for our purposes it is important to provide greater granularity by using the concepts provided by the OAIS Information Model. OAIS defines the information objects required for Long-Term Preservation. All or part of these must therefore be created/collected through the project.

These are discussed in more detail in section 4.

# Collection Groups

Section 3.1 outlines the Collection Groups. Section 3.2 provides a brief description of each of the Collection Groups.

## Overview of Collection Groups

The project activity is divided into five Collection Groups - Initiating, Planning, Executing, Closing, and Control. The Collection Groups are not project phases but rather groups of processes.

The Collection Groups, following in the style of PMBOK’s Knowledge Areas, are illustrated as follows. The groups are linked by the outputs they produce; they are overlapping activities that occur throughout a project or phases of a project. They are not expected to be either one time or discrete events.

Figure 3‑1 illustrates these Collection Groups in a project with three phases.



Figure ‑ Example Collection Area application within a project with three phases

## Details of the Collection groups

In the following sections the focus is on the Information created by the project – including the Project Management information – which may need to be retained as Additional Information so that the data (created or collected by the project) will remain understandable and usable.

### The Initiating Collection Group

The Initiating Collection Group consists of processes performed to define a new project or new phase of an existing project by obtaining authorization to start the project or phase.

This could include proposing the project, perhaps responding to solicitations and funding information available. It would be reasonable to expect the following types of information to be created:

* the aims of the project 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

This, along with any solicitations, is likely to be important Additional Information that should be preserved as documentation of the project. The participants in this group of processes will almost certainly include sponsors and proposers and may also include data managers and archivists. Examples of documents to begin managing during project initiation include the list of project participants and organizations represented,

the criteria for repositories where the project data and documentation will be preserved, agreements among participants regarding authorship ownership of intellectual property produced by the project as well as relevant policies of participating organizations regarding such rights.

### The Planning Collection Group

The Planning Collection Group consists of those processes performed to establish the total scope of the effort, define and refine the objectives, and develop the course of action required to attain those objectives.

In the Planning Collection Group 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;
* the development of the associated procedures for data collection;
* the establishment of a data dictionary

This is important Additional Information that should be preserved as documentation of the project. Examples of documents to be managed during project planning include the project management plan, the communication plan, the risk management plan, assignments for roles and responsibilities of team members, the list of project deliverables, the list of candidate repositories and how they meet the established criteria for managing data and documents produced by the project.

### the Executing Collection Group

The Executing Collection Group consists of those processes performed to complete the work defined in the project management plan to satisfy the specifications.

Activities are carried out which:

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

These processes will produce data that needs to be preserved for the long term either as a product or by-product. Examples of documents to be managed during project execution include signed contracts and approvals received from stakeholders or other authorities.

### the Closing Collection Group

The Closing Collection Group consists of those processes performed to conclude all activities across all Collection Groups to formally complete the project phase, or the entire project.

The data which may be part of the legacy of the project and which can be exploited in various ways include:

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

There may also be ideas for exploitation in future.

The Closing Collection Group is performed by the project team to use/re-use and exploit the information and, if appropriate, prepare it for handing over for long-term preservation, re-use and exploitation. Examples of documents to be managed during project closing include invoices, signed acceptances, procurement documents, and evidence of payments received.

### the Control Collection Group

The Control Collection Group consists of those processes performed to ensure the project is on track or to identify areas which need attention. This process group provides information needed to manage the other process groups. The information collected during the monitoring and controlling processes is part of the legacy of the project and therefore may need long term preservation.

This could include:

* Programmatic changes
* Configuration management materials
* Changes in development or execution schedules
* Program or design review materials
* Changes in scope
* Test results

Examples of documents to be managed during project monitoring and closing include Configuration Change Requests and other documents describing proposed changes, and documented decisions of the Change Management Board or other decision bodies.

# Additional Information Areas – to Ensure Long-Term Usability

An archive must create AIPs as part of the preservation process. Many of the components that form the AIPs may only be known by the participants in the project. This Recommended Practice provides guidance for the project participants to help ensure that the information is captured, as part of the required Additional Information.

The Information Areas 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 project.

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



Figure ‑ Archival Information Package (Detailed View)

There are other pieces of information that are 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 – planned, current and actual

• ideas about the Designated Community – previous and current

• ideas about the ways in which the data may be exploited – planned, previous, and current

These Additional Information Areas are discussed in more detail in the following sections.

## Information Areas Derived from OAIS DEFINED Information Objects

OAIS defines several major categories of information that make up the Archival Information Package (AIP): Content Information and Preservation Description Information (PDI). Packaging Information needs to be available to clearly identify and delimit what makes up the AIP. The OAIS standard also defines the Package Description, which is needed to provide visibility and access into the contents of an Archive, but is not required for the Long Term Preservation of the Content Information.

The next sections provide additional information about each of these Information Areas.

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

Duplicates?

Scope changes of project would change the above lists

#### Representation Information

The Representation Information includes

* structure,
* semantics
* other Representation Information such as analysis and display software.

In some projects 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.

Quality flags and magic values (care needed when transformed) or NULL or missing values



Figure ‑ Example of a RIN

### Preservation Description Information (PDI)

The PDI is information that is necessary to preserve the Content Information. It includes Reference Information, Provenance Information, Context Information, Fixity Information and Access Rights Information.

#### Reference Information

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

* Identifiers used in publications
* Naming conventions used in internal systems
	+ How versions/editions are dealt with e.g. numerical or time tagged versions

#### Provenance Information

Provenance Information provides information including

* specific aspects of the project origins and history,
	+ from what it was derived i.e. previously collected data
	+ what data is related
* data custody – who was in control of the data at various point in the project,
* version control – what, if any, version control was used for the data,
* calibration and test.

#### Context Information

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 project origins and history
* Funders
* Current Research Information Systems (CRIS) information

#### Fixity Information

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 - where they are kept and who can change them..

#### Access Rights Information

Access Rights Information including

* ownership,
* copyright and licensing or access restrictions.

### Package Description

The Package Description 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 project could 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 project will provide information to help in this unless the archive and the project have a close relation and the archive has chosen to maintain an AIP structure that maps directly to the project data structure.

## Information Areas Derived from 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
* Annotation systems used with the data (if any)
* Community tagging (Web 2.0) e.g. quality tags held by 3rd parties

### 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 project may only be interested in, or may only have funding for, exploiting the Primary Data in those ways.

However the project members may recognize that the Primary Data may have potential other uses. For example the Primary Data may have been overhead imagery captured to monitor changes in infrastructure to aid mapping functions and another data project may be able to make use of that same imagery for weather domain cloud cover studies.

### Suggestions about the appropriate Designated Community

The project 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 project provides. For example there may be a general area of scientific expertise or a type 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 project.

## Issues Suggested by PMBOK

**NOTE: the PMBOK and DMBOK sections will be integrated into the document and not called out separately**

* Integration Management (Not Relevant?)
	+ How are the datasets related as an integrated product eg Level 0, Level 1, Level etc, “final”
	+ Hardware/software changes/switchover
* Scope Management
	+ Changes in scope of data collection.
	+ Scope Changes (if applicable to preserving the data) would in many cases be context data. If the scope changes to gather more content, that would generate new or different content. That would likely also be noted in context.
* Time Management (Not Relevant?)
* Cost Management
* Quality Management
	+ How the quality of the data can be checked, quality flags
	+ How was it certified
	+ Reprocessing needed?
* Human Resource Management (Not Relevant?)
	+ Record of people responsible for the data e.g. who could change it or who made decisions about authenticity
	+ o Who was involved in handover -/ acceptance
* Communications Management
	+ Communications Management (Not Relevant?)
	+ Publications about the data
	+ Embargoes on data publication
	+ Confidentiality/sensitivity/security constraints
* Risk Management
	+ Legal implications if data is released
	+ Security of the data - risk of loss
	+ Schedule of deliveries - h/w and s/w
* Procurement Management
	+ Specifics of hardware & software
	+ Licences

## Issues Suggested by DMBOK

* Data Governance (Related to OAIS Access Information?)
	+ Just Monitoring and Control?
	+ Identification of the "assets" - what should be preserved
	+ Ownership
	+ Overall responsibility for the data (related to Authenticity of the data and of good quality) or who to report error to get things corrected/improved
	+ Management of Data Governance
* Data Architecture Management (OAIS Representation Info: Record the overall data architecture and relationship between data elements)
	+ Selection of context metadata and Representation Information
	+ Relationship between data elements
	+ Selection of formats and semantics, ontologies etc
* Data Development (to be renamed in the next version [3] of DMBOK to Modelling and Design) (OAIS Representation Info: Information about data modelling)
	+ Relationship between data elements
* Data Operations Management (to be renamed Data Storage and Operations) (See Content Data above)
	+ Register of formats, semantics
	+ Migration management
	+ Management of copies of the data
	+ Synchronisation policy of copies
* Data Security Management (See OAIS Access Control Information)
	+ related to OAIS Access Rights Information
	+ Defense against hacking
	+ Which anti-virus checks performed
	+ Roles of people e.g. who can change/delete
* Reference and Master Data Management
	+ data used in calibration and quality checking?
	+ WHAT DOES THIS COVER?
* Data Warehousing and Business Intelligence Management
	+ Related to potential uses of data
* Document and Content Management
	+ Classification systems used – related to OAIS Reference Information?
* Meta-data Management
	+ EVERYTHING
* Data Quality Management
	+ See above

# Framework - Activities detail

The table below indicates the minimum useful status of information capture for each of the areas in each of the Collection Groups. The Control Collection Group is not included in the table because those processes would ensure that the information is captured.

Typically information to address each issue and to document the decisions made in regard to each of these areas will begin to be accumulated early in the project. 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 project, 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.

The terms used in this section could be mapped to equivalent terms used in the local environment. For examples of this type of mapping see the Annexes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Additional Information Area** | **Detailed area** | **Initiating** | **Planning** | **Executing** | **Closing** |
| Data Object | 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 |
| Calibration and system test tools and system test data that 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 Information | 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 Information | 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 |
| The owner, and the restrictions on access (licenses), and the 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 |
| The target archives and designated community for the solicitation. | Fairly firm | Complete | Complete, but may Evolve | Complete, but may Evolve |
| The budget for archiving. | Fairly firm | Complete | Complete, but may Evolve | Complete, but may Evolve |
| 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 |
| 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 in Process Groups

# 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:

* [EO Preserved Data Set Content](http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content_v1.0.pdf) available from http://ceos.org/document\_management/Working\_Groups/WGISS/Interest\_Groups/Data\_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content\_v1.0.pdf
* [EO Data Preservation Guidelines](http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Data%20Preservation%20Guidelines_v1.1.pdf) available from http://ceos.org/document\_management/Working\_Groups/WGISS/Interest\_Groups/Data\_Stewardship/Recommendations/EO%20Data%20Preservation%20Guidelines\_v1.1.pdf
* [Preview Image Principle](http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Best_Practices/WGISS_DSIG_Preview-Image-Principle-v2_Jan2013.pdf)
* [Data Management Statement](http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Best_Practices/WGISS_DSIG_Data-Management-Statement-Apr2012.docx) available from
* [CEOS Persistent Identifier Best Practices](http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Best_Practices/CEOS%20Persistent%20Identifier%20Best%20Practices_v1.0.pdf)
* [Generic Earth Observation Data Set Consolidation Process](http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Best_Practices/GenericEarthObservationDataSetConsolidationProcess_v1.0.pdf)
* [Long Term Preservation of Earth Observation Space Data:  Preservation Workflow](http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Best_Practices/Preservation%20Workflow_v1.0.pdf) available from http://ceos.org/document\_management/Working\_Groups/WGISS/Interest\_Groups/Data\_Stewardship/Best\_Practices/Preservation%20Workflow\_v1.0.pdf

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ProposedILF Stage** | **LTDP Workflow Phase** | **Activity** | **Description** | **Input** | **Output** |
| **Formulation** | **Implementation** | **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 formExample: 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** | **Implementation** | **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. |
| **Implementation** | **Implementation** | **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) |
| **Implementation** | **Implementation** | **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 |
| **Implementation** | **Implementation** | **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) - draftTailored completed preserved data set content inventory (table) - draft |
| **Implementation** | **Implementation** | **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) - draftTailored completed preserved data set content inventory (table) - draft | Final preserved data set content (document) Revised tailored completed preserved data set content inventory (table) |
| **FormulationImplementationOperationsExploitation** | **Implementation** | **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 recordsAssociated 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 repositoriesMaster inventoryPreserved 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 recordsPreserved data set cataloguePreservation 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 inventoriesArchive, 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 inventoriesArchive, access , and management infrastructure | Improved accessibility and usability |

# ANNEX USE CASE: Large Project (Informative)

The following provides an 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

### Formulate Stage 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 fulfil 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 evaluation and definition of new missions. 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).

### Planning 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).

### Executing 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.

### Closing 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 may be 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.

## 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 project. 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, and 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. Otherwise 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 encyclopaedia 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‑‑ 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, accessed 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, which include appropriate Data Management Plans. 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)