Long Term Preservation of Earth Observation Space Data

Preservation Workflow

CEOS-WGISS
Data Stewardship Interest Group

Doc. Ref.: CEOS/WGISS/DSIG/PW
Date: March 2015
Issue: Version 1.0

Change Record

Comments	Issue	Date
First issue reviewed by LTDP working group and CEOS WGISS	1.0	March 2015

Authors

Role	Name
Editors	I. Maggio, K. Molch, M. Albani

Table of Contents

List of Figures

1. INTRODUCTION

1.1. Intended Audience

This document is intended to assist data managers in Earth observation (EO) data centers in the task of preparing Earth observation space data sets for long-term accessibility and usability.

1.2. Background

Earth observation data are unique snapshots of the condition of the Earth or atmosphere at a specific point in time. As such they constitute a humankind asset, which needs to be preserved, i.e. safeguarded against loss and kept accessible and useable for current and future generations. This task becomes more important in the view of over 40 years' worth of data available in Earth observation archives around the world – and the increasing demand for monitoring long-term variations of environmental parameters, such as sea surface temperature or global ozone distributions, which require long time series of data. Moreover, with the advent of new, high resolution Earth observation missions and programs, data volumes are expected to grow significantly over the next years.

However, the main challenge is not the sheer volume of data, but its diversity, e.g. in format and type. Historic EO data, in particular, may be stored in different formats on various types of – possibly out-of-date - media. Recovery, reformatting, and reprocessing of such data, as well as the recuperation of the associated knowledge – whether it be representation information for structural and sematic understanding, mission documentation for context, or visualization and processing capacity - is problematic if attempted many years after the mission has ended.

Therefore, data stewardship is the responsibility to curate the Earth observation data during all mission stages, starting during the mission planning phases and extending beyond the mission lifetime, when the - then 'historic' - data have to be kept accessible and useable for an – ideally – unlimited timespan.

To accomplish this task a systematic and coordinated approach to data stewardship - and within to data curation and data preservation - is needed. The preservation workflow presented in this document proposes a series of actions to be carried out before, during, and/or after the end of the Earth observation mission to ensure Earth observation space data sets are preserved in a sustainable manner.

1.3. Scope of Document

Data curation, as part of an overall data stewardship, includes data preservation, which is targeted at protecting data integrity and ensuring sustainable accessibility and usability, and is the scope of this document. This document provides Best Practices for a workflow which ensures sustainable preservation of Earth observation (EO) data. A number of supporting Best Practices documents and templates have been prepared to assist the data manager in designing and conducting individual preservation tasks suggested in this document, e.g. carrying out a data consolidation exercise. This document refers to the supporting material where relevant. The associated documents are indicated in *italics* and will be available via the CEOS website (www.ceos.org).

1.4. Definitions

This document focuses on data preservation. Data preservation is one of the data curation activities within the responsibility of data stewardship. In line with the CEOS Data Preservation Definitions document, the following paragraphs provide the understanding of the terms data stewardship, curation, preservation, and consolidation as they are used in the framework of Earth observation data preservation. Figure 1 illustrates the nested concepts.

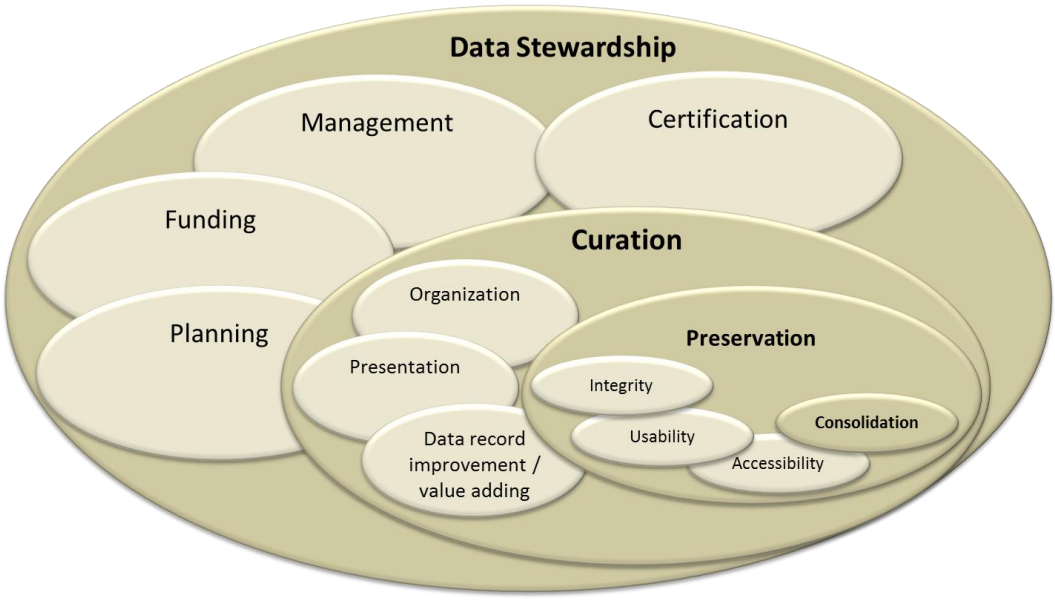

Figure 1. Content and relationship of concepts within data stewardship.

Data stewardship is the responsibility for planning, management, certification, and adequate funding for EO data sets throughout the mission phases and data life cycle. It includes curation and preservation activities.

Data curation consists of value adding, organization, presentation and preservation activities, which aim at establishing and increasing the value of EO data sets over their life cycle, at favoring their exploitation, possibly through the combination with other data records, and at extending the communities which are using the data sets. Curation is one of the tasks of data stewardship.

Data preservation consists of actions on individual or multi-mission EO data sets with the goal to ensure their integrity over time, their discoverability and accessibility, and to facilitate their (re)-use in the long term. One example is data record consolidation, which is the process to generate a canonical set of products for long-term archiving and further processing. Preservation is one of the tasks of data curation.

The preservation workflow is aimed at generating a complete EO data set for preservation. The set in this context consists of the data records and the associated knowledge. The associated comprises to tools and information which facilitate the usability of the data records. The terms

defined below and

Figure 2 provides a graphic representation of the concepts.

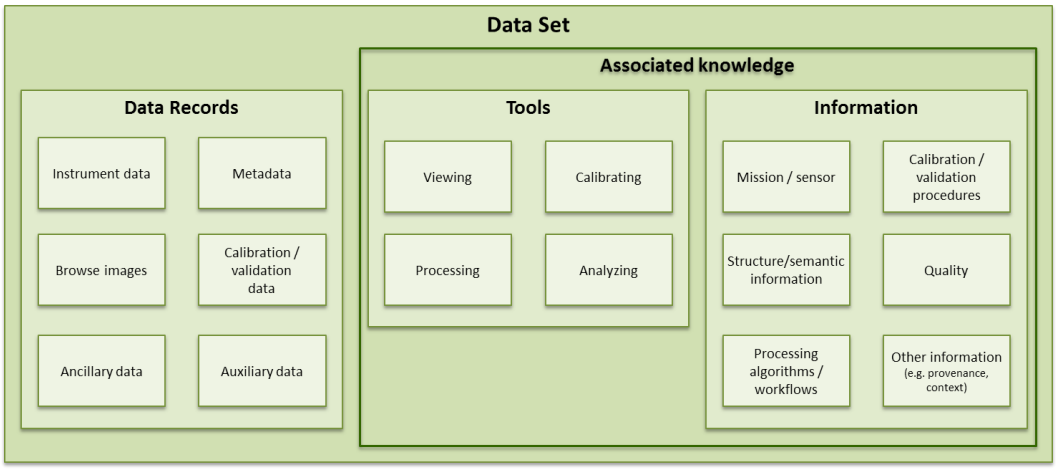

Figure 2. Components of an Earth observation data set.

Data records include the instrument data (raw data, Level-0 data, higher-level products), browse images, auxiliary and ancillary data, calibration and validation data, and descriptive metadata

The **associated knowledge** includes all the **tools** used in the generation of the data records, calibration, visualization, and analysis, and all the **information** needed to make the data records understandable and usable by the designated community. The latter includes, mission / sensor information, calibration procedures, structure and semantic information, quality information, processing algorithms/workflows and all other information as needed. The OAIS information model refers to this associated knowledge as e.g. representation information and preservation descriptive information.

For a comprehensive list of definitions related to data stewardship please refer to the *CEOS EO data stewardship definitions*.

1.5. Related Documents

The following CEOS documents are related to this preservation workflow procedure:

- EO Data Stewardship Definitions

- EO Data Preservation Guidelines
- Preserved Data Set Content
- Generic EO Data Set Consolidation Process
- Persistent Identifiers Best Practice
- EO Data Purge Alert Procedure

These documents can be found at
http://ceos.org/ourwork/workinggroups/wgiss/interest-groups/data-stewardship/

Additional documents of relevance may be found e.g. at:

- http://earth.esa.int/gscb/ltdp/
- https://earthdata.nasa.gov/standards/preservation-content-spec
- http://public.ccsds.org/publications/default.aspx

2. OVERVIEW OF THE PRESERVATION WORKFLOW

The preservation workflow defines a procedure recommended to be applied to digital data for their preservation with the objective to optimize their reuse in the long term. The procedure starts when a decision on preserving a specific EO space data set is pending and has to be taken on the management level. The procedure consists of the set of actions preferably but not necessarily to be conducted in sequence. The output of applying the procedure will be a complete, discoverable, accessible, and useable Earth observation data set, including the data records and the associated knowledge, and a series of documents describing the preservation strategy pursued, the implementation plan, and individual activities conducted.

The generic workflow presented in this document will have to be tailored to meet the preservation needs of the individual Earth observation data set at hand. The procedure is foreseen to be applied at the collection level.

The preservation workflow has been subdivided into four phases:

- Initialization (preservation planning)
- Consolidation
- Implementation
- Operations

Figure 3 provides a graphic overview of the preservation workflow. The remainder of the document will describe in detail each individual step, specifying input and output, and point to additional documentation and relevant templates.

The preservation workflow can be applied to data sets of historic, current, and future Earth observation missions alike. For historic missions, difficulties may arise in recovering all the relevant information and tools (the 'preserved data set content'). For current missions, preservation activities should be initiated while the mission is still in operation in order to recuperate all relevant information. For future missions the definition of long-term preservation strategies and implementation aspects should ideally be planned for or initiated during the mission preparation phases. This will facilitate the availability and usability of data and associated knowledge during the mission and for the long-term and will reduce associated consolidation and preservation costs.

The preservation workflow includes a cost and risk assessment, introduced during the initialization phase, which should accompany and likely be updated throughout the entire preservation process.

The preservation workflow will - in addition to a consolidated data set, generated during the consolidation activity within the preservation workflow - generate a comprehensive set of documentation on the data set and on the preservation procedure conducted. The use of a consistent file naming scheme is highly recommended. An example is provided below:

File naming scheme:	Example:
<Mission_Sensor_ProductType_DocumentName_Date>	ENVISAT_ASAR_L1b-PRI_DataSetAppraisal_2014-10-09.doc

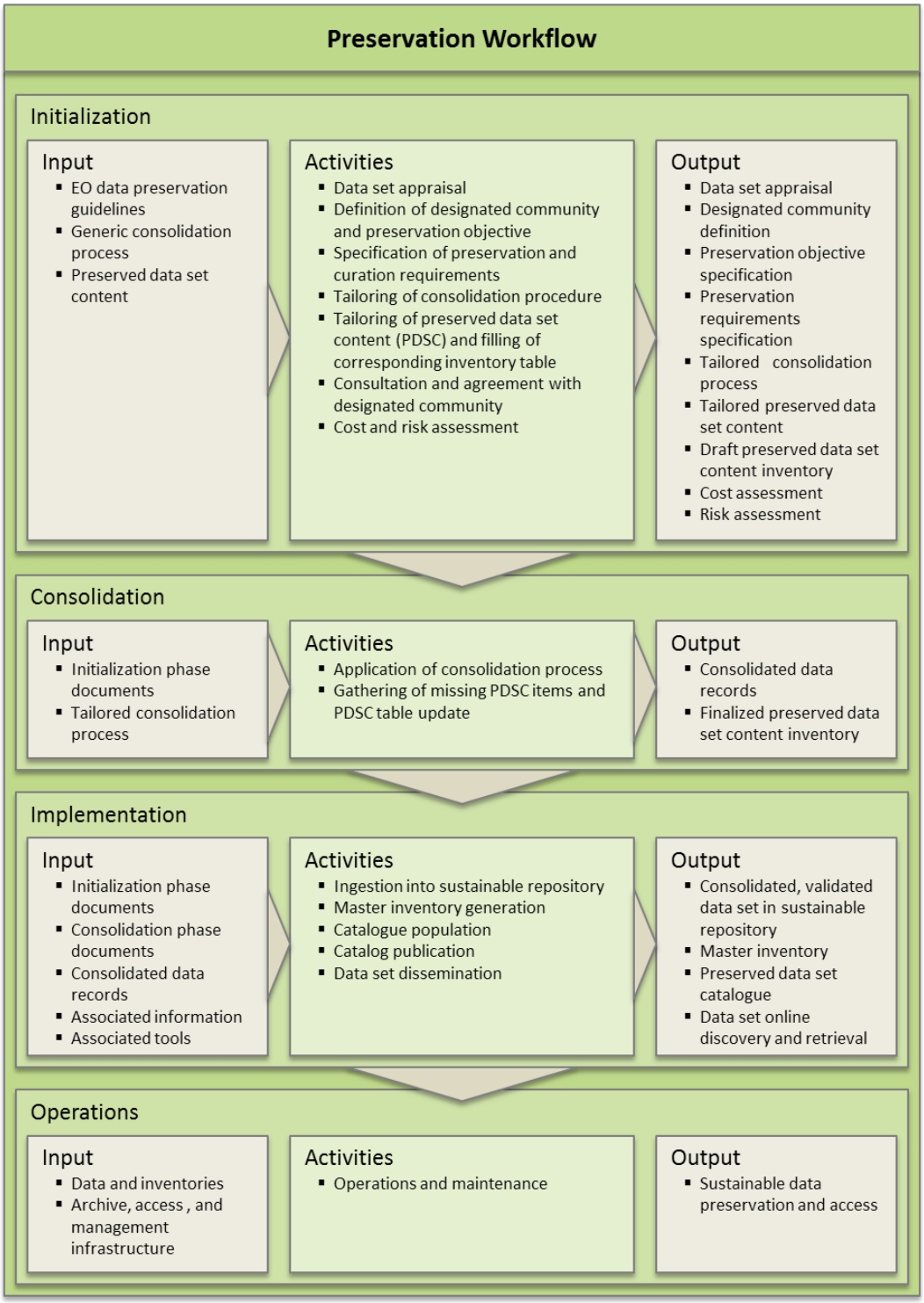

Figure 3. Overview of the preservation workflow.

3. PRESERVATION WORKFLOW IN DETAIL

3.1. Initialization Phase - Preservation Planning

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).	
Input	Example: http://eros.usgs.gov/government/ratool/
Output	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)

Definition of designated community and preservation objective	
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.	
Input	Data set appraisal (document)
Output	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.

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.	
Input	*EO Data preservation guidelines*[1] Designated community definition (document) Preservation objective specification (document)
Output	Preservation requirements specification (document)

Definition of the consolidation process	
The consolidation 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 Data Set Consolidation Process* document helps define a tailored procedure for the specific data records at hand. The tailored consolidation process will be applied to the data records during the consolidation phase.	
Input	Preservation requirements specification (document) *Generic EO Data Set Consolidation Process*
Output	Tailored consolidation process (document) addressing the following topics, as specified in the *Generic EO Data Set Consolidation Process*: • Data collection • Cleaning and pre-processing • Completeness analysis • Processing and re-processing

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 manager should generate and fill a preserved data set content inventory table to assess which

[1] Documents in *italics* are available as CEOS best practice documents on www.ceos.org

data records, information, and software is available and should be preserved. The table facilitates the assessment of completeness against the tailored preserved data set content document. For the data records, the information should be collected at both collection level and at scene/pass level. The detailed scene/pass-based 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.	
Input	Designated community definition (document) Preservation objective specification (document) Preservation requirements specification (document) *Preserved data set content document* Preserved data set content inventory table
Output	Tailored preserved data set content (document) - draft Tailored completed preserved data set content inventory (table) - draft

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.	
Input	Designated community definition (document) Preservation objective specification (document) Tailored preserved data set content (document) - draft Tailored completed preserved data set content inventory (table) - draft
Output	Final preserved data set content (document) Revised tailored completed preserved data set content inventory (table)

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 ten years. Since the temporal scope of the preservation activity extends over several decades, risks may change considerably.	
Input	Data set appraisal (document) Preservation objective specification (document) Revised tailored completed preserved data set content inventory (table)
Output	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

3.2. Consolidation Phase

Implementation of consolidation process	
The tailored consolidation process, defined earlier, is being implemented. As specified, the set of activities will be applied to the data records to be preserved.	
Input	Tailored consolidation process (document)
Output	Consolidated, clean data records, ready for preservation and user access

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.	
Input	Final preserved data set content (document) Revised tailored completed preserved data set content inventory (table)
Output	Final tailored and completed preserved data set content inventory (table)

3.3. Implementation Phase

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.	
Input	Consolidated data records Associated information (as specified in revised tailored completed preserved data set content inventory) Associated tools (as specified in revised tailored completed preserved data set content inventor)
Output	Complete, consolidated data set (data records and associated knowledge) ingested into sustainable repositories Master inventory Preserved data set catalogue

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.	
Input	Consolidated data records Preserved data set catalogue

	Preservation objective specification (document) Preservation requirements specification (document)
Output	Online discovery and retrieval (download, ordering) of the data records and (selected) associated knowledge

3.4. Operations Phase

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. 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.	
Input	Data and inventories Archive, access, and management infrastructure
Output	Sustainable data preservation and access

In order to add value to or to improve accessibility and usability of the preserved data set, curation activities should be 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. These activities, however, are outside the scope of the preservation workflow.