Long Term Preservation of Earth Observation Space Data

Generic Earth Observation Data Set Consolidation Process

CEOS-WGISS

Data Stewardship Interest Group

Doc. Ref.: CEOS/WGISS/DSIG/GEODSCP

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, R. Cosac, 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. Scope of Document

This document represents the Generic EO Consolidation Process and it is intended to be used as input to the "Tailoring of mission specific E2E consolidation process" step, described in the Preservation Workflow, to produce the **mission-specific consolidation process**. It consists of a series of recommendations and advice focused on the implementation of actions for the consolidation of the Data Records and their Associated Knowledge, for a given mission. These recommendations are meant to be used as guidance for the mission requirement definition, ground system implementation, data centres operations services, for the preservation of their data holdings.

The EO Consolidation process produces a consistent, consolidated and validated[1] set of **"Data Records"** and **"Associated Knowledge"**. It can be applied to level 0 data but also to higher level products which require a consolidation. The process is to be tailored for each Mission/Sensor according to its specific preservation and curation requirements, and consists of all the activities needed for Data Records and Knowledge Collection, Analysis, Cleaning, Gap Analysis/Filling, Pre-processing, Processing/Reprocessing (including software integration steps), Completeness Analysis and Cataloguing. "Consolidated Data Records" represent the basic input for any further higher level re-processing and for the long-term preservation.

This document is an input of the Initialization Phase of the Preservation Workflow **Error! Reference source not found.**, during which an appraisal of the Data Set is performed. The Generic EO Consolidation process is consequently defined taking into account at least the following:

- Mission/Sensor Category
- Mission Objectives and Requirements
- Mission Operations Concept
- Mission Designated Community
- Preservation and Curation requirements

1.3. Definitions

The following definitions apply to this document:

An **Archiving and Consolidation Centre** is an entity in charge of archiving and consolidating EO Missions/Sensors Data Set(s) for the purpose of long term EO data preservation. This entity is responsible for the implementation of the archive-related activities, the EO Consolidation Process and Preservation Workflow.

[1] The term "Validated" refers to quality and operative activities aimed at defining and declaring the "Data Records" as Master (i.e. usable for any future higher level re-processing campaign).

The **EO Consolidation Process** is an input to the Preservation Workflow, which defines a recommended set of actions/steps to be implemented for the preservation of the EO "Data Set" with the goal to ensure its preservation, valorisation, curation and exploitation in the long term.

An **EO Missions/Sensor Data Set** consists of data records and associated knowledge.

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*.

1.4. Related Documents

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

- Preservation Workflow
- EO Data Stewardship Definitions
- EO Data Preservation Guidelines
- Preserved Data Set Content
- 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

1.5. Document Overview

This document is divided into:

Section 1: Introduction, which includes definitions, abbreviations and related documents.

Section 2: Describes the EO Consolidation Process approach and objectives.

Section 3: Process Activities. This section describes the macro activities of the Consolidation Process.

Section 4: Stages Requirements, providing details on the stages of the EO Consolidation Process.

Section 5: Defines the requirements for statistics, completeness and quality assessment.

2. CONSOLIDATION PROCESS APPLICABILITY AND APPROACH

The Generic EO Consolidation Process focuses on the consolidation of the Data Records and their Associated Knowledge. The data record type covered by this process is potentially any data record including raw data, Level 0 data and higher-level products, browses, auxiliary and ancillary data, calibration and validation data sets, and metadata.

The EO Consolidation is a process consisting of different stages. Before starting these steps, the pre-requirement for this process is represented by the "Data Appraisal, Definition of Designed Community & Preservation Objective" activity (described in the Preservation Workflow), which consists in performing an assessment of an EO space data set under evaluation. When performing the analysis, an assessment report should be produced in order to properly document the different steps performed during the appraisal activities and the final results (see **Error! Reference source not found.**).

The EO Consolidation Process should be tailored on mission specific basis (mission objectives, budget availability, operational constraints) and according to the sensor category, quality and completeness mission specification requirements.

The EO Consolidation process should be put in place during each phase of the Mission life cycle. (New Current and Historical), in order to ensure that the preserved data set is acquired, processed, preserved, validated and certified for completeness, for the long term preservation and exploitation during each phase of the mission.

For new missions, it is of outmost importance to gather high level consolidation requirements for the specific sensor category and to refine the mission specific quality and long term preservation requirements for the data records in the mission requirements documentation, which will impact consolidation activities during the various phases.

These requirements should in turn be reflected in the mission operational concept (e.g. particularly for distributed data sets), in the System Requirements for the implementation of the flight and ground segment, and in the operational services requirements during the routine and post mission phase consolidation activities.

3. EO CONSOLIDATION PROCESS STEPS

The Consolidation Process is composed of the following recommended steps, depending on the mission phases:

- **Step 1: Data set analysis and collection:** in accordance with the initial phase of the Preservation Workflow procedure.
- **Step 2: Cleaning/pre-processing:** Data set cleaning (production of a master data set which is devoid of corrupted and duplicate files, aligned to the same naming convention and file format).
- **Step 3: Completeness analysis:** analysis of mission data set completeness to identify and define gap recovery[2] activities.
- **Step 4: Processing/reprocessing:** Regarding level 0, this step consists in the production of any missing or erroneous L0[3] products from raw data or L0 unconsolidated products. Regarding higher level products, reprocessing campaigns could be pursued if the need arises, e.g.to fill identified gaps, or if a new processor, new AUX data version, and/or any new inputs to the processing become available.

If needed, any ad-hoc interventions, which require advanced instrument and scientific knowledge of the data, could be foreseen and pursued.

The long term preservation mandatory requirement for the consolidation process is that each of the above steps is properly documented.

The data record inherent knowledge, as collected during data set appraisal, together with knowledge gathered from the consolidation process steps, should be maintained and validated consistently with the newly generated master data records and any processing chain, as part of the EO data record consolidation process.

The following should be produced as output during the consolidation process.

- Individual Data Records resulting from each of the Consolidation Stages and candidates for being labelled as Master.
- EO Consolidation Technical Notes, documenting the various steps, trade-offs, decisions etc.
- Ancillary tools used or developed during the course of the Consolidation exercise to read/convert/assess redundancy/assess completeness/stitch the data, etc.
- Verification documentation.

The following paragraphs offer a detailed description of each recommended step.

3.1. Step 1 – Data Collection

1.1 Analysis and Collection - analysis of the available Mission/Sensor Data Set, in accordance with the initial phase of the Preservation Workflow procedure:

- Identify the data set content to be preserved: data records, processing software, mission documentation)

[2] Gap Recovery: specific processing which aims at reconstructing the missing data from the raw data, if thisis still available or, keeping/deleting the data according to its quality (typically 'hole' processing in the raw data).

[3] If the relevant raw data is available

- Generate and (during the process) maintain a complete inventory of the archived preserved data set content, defined for each mission/instrument, with the following items as a minimum:
    - Description and availability of the data records
    - Coverage and volumes of the data records
    - Physical locations of the data records
    - Media of storage and archive formats.
    - Processing software (if maintained or simply archived) information: versioning, IPR/licenses, etc.
    - Mission related documentation: versioning, repository, etc.

1.2 Transcription: applies the Transcription procedure to retrieve and recover Data Sets. The transcription procedure should also foresee the conversion/decoding of the data format. Sometimes a Transcription/Media Migration should be pursued, in order to transcribe historical media to ready readable format (DLT, LTO, EXABYTE) and to allow decommissioning of original mission-specific hardware and software applications.

This recommendation is useful for the Gap Recovery activity (foreseen in Step 3) and to complete a Reprocessing Campaign.

3.2. Step 2 – Cleaning/Pre-processing

2.1 Merging: if data is identified and collected from different data sources, a merging of data is required. The activity also foresees the exclusion/isolation of exact duplicates.

2.2 Decode: if needed, or particularly for Historical missions, when a process of decoding is to be put in place. In particular:

2.2.1 Aligning/homogenizing data, including File Naming and Format/Packaging, possibly by means of ancillary tools developed for the purpose

2.2.2 Identify and harmonize metadata, improving them (in line with the SIP Standard)

2.3 Cleaning:

2.3.1 Detection and exclusion/isolation of corrupted files

2.3.2 Reduction of redundancy across the files

2.3.3 Analysis of overlaps, either by means of ancillary tools specifically developed for the purpose or manually

2.3.4 Ad-Hoc (Mission/Sensor Based) intervention on the files to reactivate their usability for Recovery or Reprocessing Campaign

2.3.5 Classification of Bad Data to allow partial reprocessing/recovery if possible (Sensor based)

3.3. Step 3 – Completeness Analysis

3.1 Monitoring: for operational missions this consists of periodic checks (Mission/Sensor based) on the production. This step is composed of:

3.1.1 Analysing and **collecting** the unavailability for instrument, station and satellite in order to declare the real and fixed holes/gaps

3.1.2 Gap Analysis & Recovery: for operational missions, periodic recovery procedure (Mission/Sensor based) to be implemented in order to cover any gaps in the coverage, not relating to the above mentioned unavailability periods.

3.2 Control: quality and integrity check to be performed on the output. This step is composed of:

3.2.1 Integrity check: validation of the production packaging before ingestion in the archive (systematic 1st level of validation)

3.2.2 Post-processing check: quality check, such as Image Quality Control

The assessment of the completeness of the data records shall be, whenever available, based on the original mission operation planning, taking into account ground and flight unavailability and data records processing auxiliary files orchestration. The assessment shall foresee:

a) For Future missions: to define the mission specific scientific quality requirements affecting completeness, sensor specific quality measure (uncertainty, inter-orbit minimum gap, intra-orbit missing data minimum thresholds, etc.) and its implementation in the operations mission concept and relevant ground and system specification and design. The consolidation process can be used to help in defining what information should be archived/monitored during the mission, how this should be shared and with whom.

b) For Current missions: to routinely apply systematic data analysis and recovery procedure at the operational acquisition centers and to routinely assess the data set against quality and completeness requirements.

c) For Historical missions: to collect all available data record at all level, retrieve original mission planning if available, catalogue metadata, original operations concept and any useful information necessary to implement the data coverage Gap Recovery procedure or in order to complete a reprocessing campaign.

3.4. Step 4 – Processing/Reprocessing

Regarding Level 0, it consists in the production of any missing or erroneous L0[4] (the processing to the corresponding higher level product could also be useful for checking the quality of the L0 produced). Regarding higher level products, reprocessing campaigns could be pursued if the need arises e.g.to fill identified gaps, or if a new processor, new AUX data version, and/or any new inputs to the processing become available. The following steps are addressing mainly level 0 data but can be applied also to higher level products.

4.1 Packaging Standard definition

4.2 Alignment of the Data Records format with respect the SIP standard

4.3 Generate or manage the quality flag in the reprocessed products, in order to have clear evidence about the quality of the Data sets

4.4 Generate and harmonize metadata

4.5 Regeneration of missing L0: any missing Level 0 data is regenerated from raw data, using the latest version of the Level 0 builder (the respective higher level product could also be processed if needed, in order to attest the Level 0 quality).

4.6 Reprocess/bulk-process data records to higher-level products, using the last version of the processor for the higher levels. Data sets already generated by previous processor versions are

[4] If the relevant raw data is available

reprocessed and aligned to the last one. This activity is driven only by real and important changes in algorithm or AUX data.

4.7 Validation test definitions for archiving: to ensure full agreement on both sides, some initial submissions should be performed on the 'test data' prior to the beginning of the data delivery.

4.8 Ad-hoc interventions (Data Expertise) if advanced instrument and scientific knowledge of the data is requested.

4.9 Reconciliation: to match the output of the exercise to its input in order to recognise any mismatch.

4. STEPS REQUIREMENTS

The diagram shown in Figure 1, below, describes the recommended activities to be conducted during each individual step.

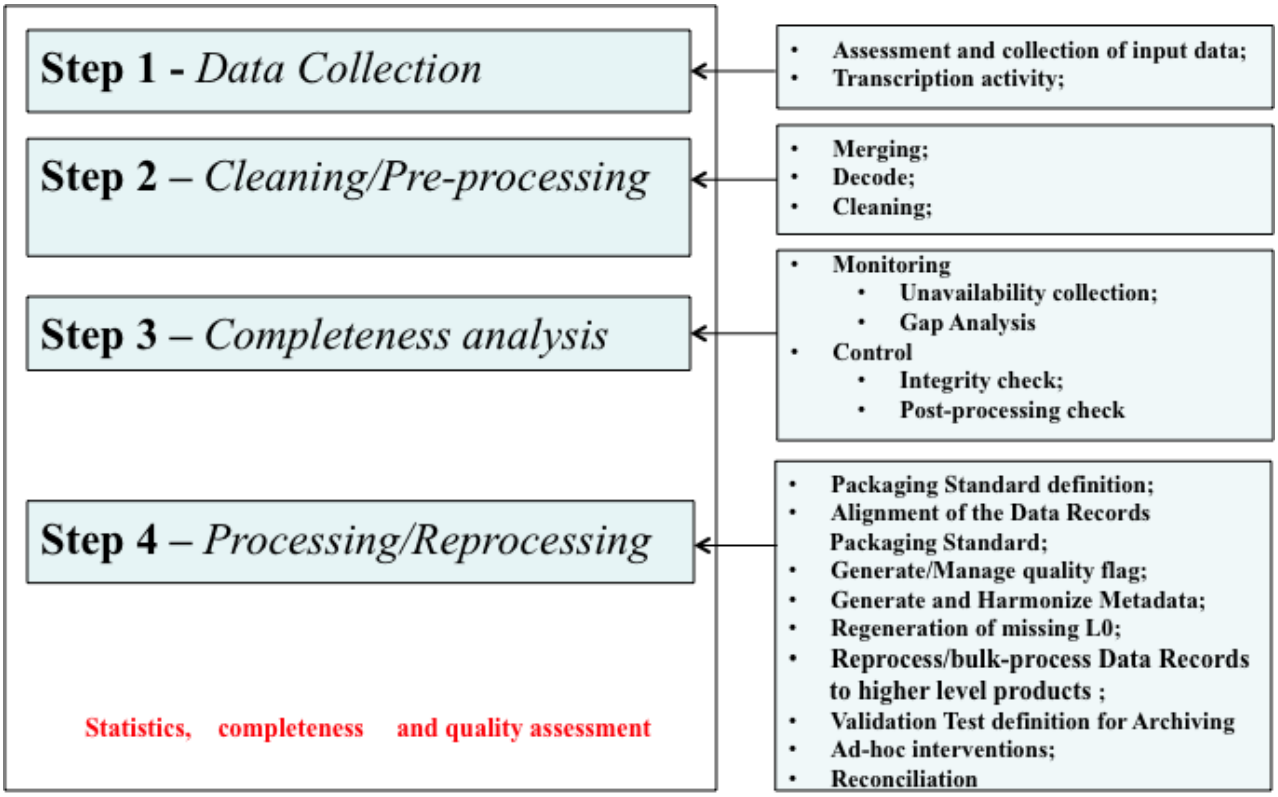

Figure 1 - EO Consolidation Process Steps

4.1. Step 1: Data Collection

The recommended activities are described in the following paragraphs. Depending on the Sensor category, some activities may not be applicable.

4.1.1. Assessment and Collection of input data and resources

Occasionally there are many sources available for the same data set, including:

- Main archives, hosting a complete data set, should be targeted first, such as Archiving and Consolidation Centres
- Alternative archives, hosting a backup or a large percentage of a data set, in particular for low Level data, ground stations, re-transcription centres, Archiving and Consolidation Centres, using the files as input

- Other sources hosting a limited amount of data such as, backup ground stations, users holding copies of files, etc.

It is important to identify all the possible sources because:

- In some cases no archive contains all the data, and completeness can only be achieved by merging several sources
- Some files may be corrupted and have to be replaced by another copy existing in another identified source
- Some files may be stored on obsolete media, for which reading devices are no longer available or, that are not compatible with the latest hardware.
- The access or restoration of the products may involve more time or manpower, depending on the source.

The Archiving and Consolidation Centre should:

- Identify any other sources in addition to the input data sets provided, which may be used in the consolidation process and improve the quality and completeness of the output.
- Recommend a prioritisation of the sources, indicating which should be ingested for consolidation and which, if any, should be left aside, on the basis of a cost/benefit trade-off between the cost of extracting the data and the amount of data expected from each source.
- Provide an estimate of the completeness and redundancy levels that can be expected.
- Collect and ingest into its storage infrastructure the data from the various input sources (or selected portions).

Note: no data should be discarded and any files not involved in the consolidation exercise should be kept in the service storage and/or inventoried until the Owner of the Consolidated Data Records declares it as "Master", after which, any erroneous or duplicate products could be deleted.

Data consolidation requires some critical information, particularly for assessing the contents of the data, in order to homogenize the data and to evaluate its completeness.

The following documentation is typically necessary:

- Accurate and complete format description, for all versions of files: this must be available both for the analysis of the file content and extent, and for end-usage of the files.
- Unambiguous and extensive file naming convention.
- History of processing software changes is required to understand the differences between redundant or various generations of files, eventually setting up a selection strategy during the Cleaning phase.

In addition, some mission- and spacecraft-related ancillary data is necessary for the homogenization and completeness analysis. This may typically include:

- Mission Phases: a clear definition of the mission phases (start/end orbit and start/end times)
- If applicable, mission cycles: proper and clear definition of the cycles (start/end orbit and start/end times)
- Orbits: clear definition of all mission orbits (start/end times)
- Information on payload activity and anomalies. Unavailability or anomaly notifications for platform, instrument, telemetry or ground stations are essential to assess the level of completeness of the data.

- Results of previous assessments of the data (e.g. for completeness analysis)
- Detailed Mission Operations Plan (DMOP): reporting the planning input against actual time series. This can explain when periods with no measurements were planned.
- Emergency planning requests

The Archiving and Consolidation Centre should:

- Determine the list of additional resources (documentation and ancillary data) needed for the consolidation exercise and the authoritative sources for these resources.
- Collect the additional resources from their authoritative sources.

4.1.2. Transcription Activity

The Transcription activity ensures first and foremost that the data are protected from immediate loss, by means of copying them into a safe environment, such as a robotic tape library. A proper ingestion using a specific data model - including the generation of catalogue information and browse images - may or may not be done in this step. If this is not included in the transcription activity, proper ingestion will have to be done later since this ensures discoverability and accessibility of the data.

Transcription may also involve the following activities:

- Technology migration: if a data record resides on old media, migration to new media should be considered, in order to avoid any data loss
- Data Formatting: when a new data format is needed, the transcription allows to adapt/align the format of the data set

4.2. Step 2: Cleaning/Pre-processing

The recommended activities are described in the following paragraphs. Depending on the Sensor category, some activities may not be applicable.

4.2.1. Merging

If different data sources are identified and collected, a merging[5] of data is needed. This activity also foresees the exclusion/isolation of exact duplicates (i.e. when acquisition start and stop time and data record content are the same).

The Archiving and Consolidation Centre should:

- Merge the data from the various sources into a single data set, taking into account the versions and the quality of the data (Note that the quality of the data depends on the sensor category)
- Detect files with the same file name but different content, and rename the files in compliance with the adopted file naming convention
- Detect files which have exactly the same content (bit-for-bit comparison), but a different name, and consequently eliminate the redundant files which do not respect the adopted file naming convention

In case the adopted file naming convention allows identical files to have different names, or files with different contents to have the same name, the Archiving and Consolidation Centre should then extend

[5] Merging refers to when one or more products overlap, the highest quality of these products is chosen for creating a unique, better product

the file naming convention to eliminate such cases, and document the resulting extended file naming convention.

4.2.2. Decoding

If needed, or particularly for Historical missions, when a process of decoding is to be put in place. In particular:

4.2.2.1. Alignment/Homogenisation

The Archiving and Consolidation Centre should align the data from all the sources contributing to the consolidation, to the chosen file names, format and packaging.

In case the reading/conversion tools are not fully available among the collected material, the data homogenisation may require the development of such tools.

The files collected for the consolidation may be heterogeneous in terms of:

- File naming
- Format and Packaging e.g. including or not an XML metadata file for each file present in a compound product. (e.g. Internal Archiving format, Multi Mission Facility Infrastructure)

The data needs to be homogenised in order to facilitate the consolidation, as well as further utilization of the data.

4.2.2.1.1. File Naming

The Archiving and Consolidation Centre should define and use a file naming convention for the consolidated data, based on:

- The OAIS and PAIS/PAIMAS standards
- The expectations of the processor envisaged for subsequent processing of the data
- The most frequently used naming convention for the sources being consolidated
- The most recent file naming convention used for the mission
- The file naming convention used most widely by the data users
- The most suitable file naming convention

The proposed naming convention should represent an existing standard naming convention. Unless it is necessary to satisfy the required criteria, no new file format or packaging should be developed.

The file naming convention proposed by the Archiving and Consolidation Centre should:

- Have no dependency on other resources. All the fields of the file name (e.g. orbit number, start and stop time, product type code, checksum) should be extracted from the contents of the file itself.
- Code the receiving stations according to internal station codes.
- Be based on the standard definition of orbit, and corresponding time frame, i.e. "from equator crossing in ascending node (ANX) to next equator crossing ascending node (ANX+l)". Other definitions (e.g. from pole to pole) should be avoided.
- Be unambiguous (e.g. in case a checksum is included in a file name, the choice of calculation algorithm should be explicitly indicated).
- Be fully documented.

- Ensure that files with differences in their contents are assigned different names, and that files with identical contents are assigned identical names.

Note: this is expected to involve a checksum being part of the file name.

4.2.2.1.2. Format and packaging

The Archiving and Consolidation Centre should propose, in line with the Data Preservation and other Standards in use, a format and packaging for the consolidated data, based on:

- The expectations of the processor envisaged for subsequent processing of the data
- The most frequently used format and packaging for the sources being consolidated
- The most recent file format and packaging used for the mission
- The file format and packaging favoured by the data users
- The most suitable file format and packaging for the long term preservation of the consolidated data

The proposed file format should be an existing standard file format. Unless it is necessary to satisfy the required criteria, no new file format or packaging should be developed.

4.2.2.2. Identify and harmonise metadata

The related recommendations suggested for the data, in the previous paragraphs, should also be applied to the metadata, improving them in line with the metadata standards used.

4.2.3. Cleaning

4.2.3.1. Detecting and exclusion/isolation of corrupted files

The Archiving and Consolidation Centre should detect any files which may be corrupted, including:

- Files with contents or size not compliant with the header meta-data
- Files with sizes that are obviously not in line with the reasonable expectations, considering the average file size
- Files with field values incoherent with the nature of the field
- Files with checksum that is not equal to the checksum contained in the file name (if available).

The Archiving and Consolidation Centre should:

- Detect and separate the corrupted files from the target Data Set.
- Replace the corrupted files with intact versions, if available from another source.

4.2.3.2. Reduction of redundancy across files

The Archiving and Consolidation Centre should make an assessment and a recommendation concerning the redundancies in the resulting data set, and propose a course of action, such as:

- No further action, e.g. in case there are few redundancies
- A file-based approach, where full files are kept or discarded without modifying/stitching the contents of the files, e.g. in case the redundancies are concentrated in some of the files it is recommended to justify the proposed criteria for selecting highly redundant files to be discarded. To show this, the example of two Level 0 files, generated using two different Level

0 processor versions, is proposed: in this case, the decision could be to keep the Level 0 file generated with the latest version, if the file is not corrupted.

- A record-based approach involving the modification/stitching of the contents of the files

4.2.3.3. Overlap Analysis

This activity foresees an analysis of the overlaps within the files, possibly by means of ancillary tools developed for the purpose.

This kind of analysis depends on the sensor/mission category. The overlap management is a recommendation of the "Tailoring of mission specific E2E Consolidation process" step, described in the Preservation Workflow.

The example reported below is related to a procedure which has been used during a consolidation of data from ESA radar altimeter and microwave instruments:

The overlaps could be classified according to the following categories:

- Products with exactly the same start and stop time: these products are considered as fully redundant and one has to be selected over the other. Since both files exist in the merged Master Data Set, an investigation will have to be carried out in order to select the best one.
- Products including other products: one product fully includes the coverage of another (and extends it). The including product could be assumed as the best one however, this should be confirmed by content analysis, as mentioned above.
- Products having almost exactly the same coverage (more than 90% each, or more than 70% each, if none of the products is longer than 15 minutes): these products are considered as redundant despite them being slightly shifted from each other. Selection of the best product will also be performed through content analysis. Some products are shifted, although they have the exact same duration. This is suspected to be related to time issues with the records. If the shift is lower than 10 seconds, these products are classified as having almost the same coverage.
- Products with less than 10% coverage overlap are not considered to be overlapping (there is typically a 10 to 15 seconds overlap at each end of consecutive products) and both are kept in the master data set.
- Products with up to 50% coverage overlap are also not considered to be redundant, although they do have some significant overlap (in most cases from about 70 to 100 seconds). However, they also have significant extra content, and thus both are kept in the Master data set.
- Products with significant overlap have between 10% and 90% overlap for at least one of the files. These cover multiple cases:
    - 1st case: one product is almost completely included in another product (more than 90%), but smaller, since the including product overlap is less than 90%, and only extends it by a few records (less than 60 seconds). In this case, the smaller product is rejected from the Master data set.
    - 2nd case: when both products have significant overlap, but each file also has a lot of extra content (some products seem to be shifted by about a quarter of orbit). If this is at least 5 minutes for each product, then both are kept. They are not considered redundant.

The Archiving and Consolidation Centre should:

- Check that the presumed file start and stop times, according to the file name/header, match the actual contents. If this is not the case, then the start and stop times should be corrected.
- Compute, for each file, the overlap length and percentage with respect to the previous and next files in chronological order.
- Analyse the overlapping sections of two files to determine if the corresponding data is fully redundant or not and, on this basis, produce an index of redundancy. The analysis should focus on the similarity of the two overlapping files/file sections, using as a minimum, the following criteria:
    - Number of data records
    - Record times
    - Acquisition station / processing centre
    - Processor version
    - Data in the overlapping records: metadata, ancillary information, quality information, measurements

All suggested and recommended activities could be performed by means of ancillary tools, specifically developed for the purpose, or manually.

On the basis of this analysis, the Archiving and Consolidation Centre should define a strategy (e.g. stitching of files and removal of overlapping records) for the elimination/reduction of the redundancy in the data.

The proposed strategy should:

- Be executable through a defined algorithm applied to the full data set
- Include rules for associating the newly formed primary and auxiliary data files resulting from the stitch-and-remove process
- Show any trade-off proposed between the removal of redundant data and the loss of unique data

4.2.3.4. Ad-Hoc (Mission/Sensor Based) intervention

Ad-hoc interventions could be applied on the files, in order to reactivate their usability, e.g. for a Recovery or Reprocessing Campaign.

These interventions depend on the sensor/mission category and the recommendations represent the output of the "Tailoring of mission specific E2E Consolidation process" step, described in the Preservation Workflow.

E.g. for the SEASAT mission, the following activities were applied to the Raw data set:

- Validation and bit alignment
- Cleaning of bit error
- Reducing discontinuity
- Linear regression
- Creation of geo-referenced ground range amplitude images

4.2.3.5. Classification of Bad Data

In spite of all the work done to decode and clean the data, many errors could remain in the supposedly fixed files.

The classification of Bad Data is useful to allow partial reprocessing/recovery, if possible (Sensor based), as described in Step 4. This classification depends on the internal infrastructure present in the dissemination center. The flag to highlight the Bad Data could be useful to distinguish the right products to be provided to the end user (e.g. to be used in the catalogue search).

4.3. Step 3: Completeness Analysis

The recommended activities are listed in the following paragraphs. Depending on the Sensor category, some activities may not be applicable.

4.3.1. Monitoring

For Current or Future missions, the periodic checks (Mission/Sensor based) on the production, could be performed weekly or twice per week. The tailoring of the EO Consolidation Process should define the suitable timeframe, taking into account the mission operations scenario. In line with the LTDP Common Guidelines, the operators should follow approved procedures in order to ensure the whole coverage of the Data Records.

If the mission is Historical, the check should be done on the whole data set in order to highlight any gaps in the coverage. The coverage check on the whole data set should be done taking into account the acquisition planning.

Tools and ad-hoc software could be used to monitor the Data Records' completeness.

4.3.1.1. Unavailability Collection

Analysis and collection of any unavailability for instrument, station and satellite, in order to declare the real and fixed holes/gaps.

The list of unavailability should be declared for:

- Instrument unavailability
- Platform unavailability
- Station(s) unavailability

For Current missions, this collection should be performed every time the unavailability occurs.

For Future missions, if the preservation standards are followed, the unavailabilities will already be available when the mission passes into the Post Mission Phase.

The Archiving and Consolidation Centre should:

- Produce a "net" list of the gaps, each associated with its duration, orbit identification and time, and the corresponding geographical location, limited to the gaps not explained by unavailability.
- Compute an index of completeness of the consolidated data set, based on:
    - The data which was expected, considering the unavailability and the planning
    - The data which is actually present in the data set
- Provide a justification for any gap in the data, that is not explained by a known unavailability or planned gap, including:
    - Possible reason why the sources of the input data do not contain the data in question
    - Possible additional sources of data able to fill the gap

  - Possible presence of the data in the sources rejected, following the cost benefit trade-off performed at the start of the exercise
  - Possible additional unavailability not officially communicated
- Use catalogue information, from the various organisations managing and distributing the data being consolidated, in order to support the differentiation between data which was never acquired and data which was lost.
- Propose a revised (enlarged) input set of data to be consolidated, in the form of a delta consolidation, including additional data previously excluded, together with an estimate of the gain expected from a new iteration, including this extra data.
- Fully document the completeness assessment.

4.3.1.2. Gap Filling Analysis

For Current or Future missions, the Gap Filling Analysis corresponds to the periodic recovery procedure (Mission/Sensor based) to be implemented in order to fill any gaps in the coverage, not related to the above-mentioned unavailability.

If the mission is Historical, the Gap Filling Analysis should be done on the whole data set, in order to cover and highlight all the gaps. This should be done after the Data Collection and Cleaning.

The Data Gap Filling allows the identification of any hole in the coverage (not relating to any unavailability for the instrument, station and satellite, or planned instrument pause). This activity is useful for the regeneration of the missing L0 (if possible) and of the higher-level products (if needed, for checking the L0 Data quality).

The Gap Analysis will produce a technical note, highlighting all the missing products that need to be recovered. This document represents the input for the next step.

4.3.2. Control

This step defines a quality and integrity check that should be performed on the output data. It includes:

- Integrity Check
- Post-Processing Check

4.3.2.1. Integrity Check

The Integrity Check allows the validation of the production packaging before ingestion in the archive (systematic 1st level of validation). This activity could be performed using tools developed specifically for the purpose, or manually.

4.3.2.2. Post-Processing Check

This kind of quality check, such as Image Quality Control, should be performed if it is included in the requirements collected during the Initial Tailoring (High Level Requirements in Preservation Workflow). This activity could be performed using tools developed specifically for the purpose, or manually.

4.4. Step 4: Processing/Reprocessing

Regarding level 0, this step consists in the production of any missing or erroneous L0 products from raw data or L0 unconsolidated products. Regarding higher level products, reprocessing campaigns

could be pursued if the need arises, e.g.to fill identified gaps, or if a new processor, new AUX data version, and/or any new inputs to the processing become available.

4.4.1. Packaging standard definition

Definition of Packaging solution, in line with the relevant packaging standard (e.g. OAIS/SIP), through the exchange of requirements and identification of constraints.

4.4.2. Alignment of the Data Records format

Alignment of the Data Records format, in line with the OAIS and LTDP standards.

4.4.3. Generate or manage quality flag

During the reprocessing/bulk-processing activity a quality flag or report should be generated and linked to the reprocessed products, in order to have clear evidence regarding the data quality. A quality information summary or statistics could be produced and included in the metadata or in the SIP package.

4.4.4. Generate and harmonize metadata

Generate and Harmonize the Metadata in line with the OAIS and other data preservation standards (e.g. OGC O&M) to improve the services and functions (data search, discovery, retrieval and dissemination), which make the archival information holdings accessible to users.

4.4.5. Regeneration of missing L0

During this activity any missing or erroneous Level 0[6] data is regenerated from raw data, using the latest version of the Level 0 builder (the respective higher level product could also be processed if needed, to attest the Level 0 quality). This step should also allow for any Level 0 data, originating from a previous version, to be regenerated with the latest version (in order to obtain a consistent Level 0 data set). Any regenerated data should also pass through the same checks as the original collected data, i.e. through Step 2.

This step forces the Recovery procedure, useful for retrieving and transcribing the raw data relating to the missing L0, in order to process and generate the missing product. Assuming recovery is successful, the results of the completeness analysis will need to be updated in order to account for the gaps that have been partially or totally filled.

4.4.6. Reprocess/bulk-process data records to higher level products

Using the last version of the Processor, all data sets already generated by previous processor versions are reprocessed and aligned to the last one. This activity is driven only by real and important algorithm changes, and it should be analysed on a case by case basis. If the data have never been systematically processed (although the processor was used only for on-demand processing) then a bulk-processing campaign could be triggered (to be analysed case by case).

Note: The decision to implement this step depends on the high level requirements, collected during the initial phase of the Preservation Workflow, and on the volume of the data, the available hardware resources, the available human resources, etc.

4.4.6.1. Software Integration

- **Processor Integration**

[6] If the relevant raw data is available

Processing system should be installed, integrated and tested in the Archiving and Consolidation Centre environment, prior to reprocessing/bulk-processing.

- **Quality Control Tool Integration**

If available, a Quality Control tool that generates a quality flag (or report), for every processed product, should be installed, integrated and tested in the Archiving and Consolidation Centre environment, ready for running.

4.4.7. Validation Test Definitions for Archiving

In order to ensure full agreement on both sides, some initial 'test data' submissions should be performed prior to the delivery of the data.

4.4.8. Ad-hoc interventions

If needed, any ad-hoc interventions, which require advanced instrument and scientific knowledge of the data (i.e. data expertise), could be foreseen and pursued.

4.4.9. Verification

After the re-processing/bulk processing, a confirmation process should take place, to match the output of the exercise to its input (e.g. if the output has 20 fewer files that the input, the reason should be known, mismatches should be eliminated if possible and documented if not). As a result of the verification, the output (higher level) data set inherits the attribute "Master" from the input (lower level) data set. This can then be used to produce even higher level data.

5. STATISTICS, COMPLETENESS AND QUALITY ASSESSMENT

These activities should be performed as part of the different steps, as applicable.

5.1. Production of statistics on anomalies and redundancies

The Archiving and Consolidation Centre should produce a series of statistics, including (before and after the Step 2 - Cleaning/Pre-processing exercise):

- List, number and percentage of files detected as corrupted
- List, number and percentage of files detected as clones or exact duplicates (including clones revealed as a result of the renaming operation)
- List of renamed files, showing both the old and the new name), by source, number and percentage of files renamed per source/overall
- List of different formats encountered per source, and list, number, percentage of files per format and of files converted into a different format
- If applicable, as part of the approved redundancy-reduction policy: list, number and percentage of files rejected due to overlaps with other files and due to having a lower quality or fewer records, amount of overlap before and after the overlap reduction, data completeness, before and after the overlap reduction.
- List, number and percentage of overlapping portions of files, before and after the redundancy-reduction exercise
- Proportion of overlaps (index of redundancy) in the output data set

The Archiving and Consolidation Centre should produce the statistics in a way that will allow automatic filtering for specific statistics:

- Per phase of the mission
- Per type of input media
- Per input source

The Archiving and Consolidation Centre should include/append the statistics produced, to the Consolidation Technical Note.

5.2. Completeness Assessment

The data set completeness assessment consists in detecting the missing data, with respect to mission acquisition planning, and qualifying the gaps either as platform/instrument anomalies or as data loss.

In principle, all missing data, with respect to the theoretical mission acquisition plan, should be related to a platform or instrument or ground segment unavailability (such as manoeuvre, orbit or phase change) or anomaly. The process should be tailored following the initial analysis, and should also take into account the mission ground segment or the mission specific rules.

Any “lost” details should be considered suspect and should be solved either by improving the list of anomalies or by recovering the data that is still missing. In some cases it could be justified that there

may be a remaining part of "unexplained gaps", as sometimes, the corresponding information cannot be retrieved anymore). An index of completeness should then be calculated to qualify the output data set.

The Archiving and Consolidation Centre should take as the basis for the completeness assessment, the actual payload activity, as reported by the FOS.

The Archiving and Consolidation Centre should detect:

- Inter-file gaps, i.e. elapsed time between the end and the start of two chronologically consecutive files
- Intra-file gaps, time between two consecutive records within the same file, larger than a threshold commensurate with the mission/instrument/type of file

The Archiving and Consolidation Centre should:

- Determine the duration and geographical location of each gap
- Produce a list of the gaps, each associated with its duration and the corresponding geographical location.

5.3. Verification of the Data Consolidation activities

The Archiving and Consolidation Centre should prepare and perform:

- Verification plan (including the test specifications, test scenarios and any tools used in the testing exercise, covering non regression tests if applicable).
- Verification exercises on the basis of the verification plan, and document the outcome in a Verification Test Report.

The test scenarios of the Verification Plan should cover at least the following:

- Sample checks and statistics
- Evidence of the presence of corrupted/duplicated files
- Inconsistent file naming/ inconsistent file format
- Sample checks and statistics to highlight the presence of unjustified gaps
- Sample checks and statistics to highlight the presence of overlaps among the files
- The recalculation of the selected samples of the statistics produced, using the same tools as in the consolidation exercise or alternative tools
- Sample checks of the data storage contents vs. records in the Data Information System
- Any other approach depending on the specific project and data.

The Archiving and Consolidation Centre will pronounce the successful completion of the Verification exercise. After the successful execution of the Verification plan, the Data Release Management process should be activated.

5.4. Data Configuration Management

Throughout the process, it is important to implement configuration management processes for all data sets involved. This comprises:

- Ascertaining and documenting the versions of all input data, e.g. Level 0 data sets

- Controlling and recording the configuration of any auxiliary or ancillary data to be used for processing
- Configuration of the processing environment itself (processing software, QC tools, etc.)
- As a final step, the output data sets (higher level products and metadata) versions must be set and controlled