IJDC | *Conference Pre-print*

Long-Term Data Preservation Data Lifecycle, Standardisation Process, Implementation and Lessons Learned

Mirko Albani
ESA

Iolanda Maggio
ESA

Data Stewardship Interest Group
CEOS

Abstract

Science and Earth Observation data represent today a unique and valuable asset for humankind that should be preserved without time constraints and kept accessible and exploitable by current and future generations. In Earth Science, knowledge of the past and tracking of the evolution are at the basis of our capability to effectively respond to the global changes that are putting increasing pressure on the environment, and on human society. This can only be achieved if long time series of data are properly preserved and made accessible to support international initiatives. Within ESA Member States and beyond, Earth Science data holders are increasingly coordinating data preservation efforts to ensure that the valuable data are safeguarded against loss and kept accessible and useable for current and future generations. This task becomes increasingly challenging in view of the existing 40 years' worth of Earth Science data stored in archives around the world and the massive increase of data volumes expected over the next years from e.g., the European Copernicus Sentinel missions. Long Term Data Preservation (LTDP) aims at maintaining information discoverable and accessible in an independent and understandable way, with supporting information, which helps ensuring authenticity, over the long term. A focal aspect of LTDP is data Curation. Data Curation refers to the management of data throughout its life cycle. Data Curation activities enable data discovery and retrieval, maintain its quality, add value, and allow data re-use over time. It includes all the processes that involve data management, such as pre-ingest initiatives, ingest functions, archival storage and preservation, dissemination, and provision of access for a designated community.

The paper presents specific aspects, of importance during the entire Earth observation data lifecycle, with respect to evolving data volumes and application scenarios. These particular issues are introduced in the section on 'Big Data' and LTDP. The Data Stewardship Reference lifecycle section describes how the data stewardship activities can be efficiently organised, while the following section addresses the overall preservation workflow and shows the technical steps to be taken during Data Curation. Earth Science Data Curation and preservation should be addressed during all mission stages - from the initial mission planning, throughout the entire mission lifetime, and during the post- mission phase. The Data Stewardship Reference Lifecycle gives a high-level overview of the steps useful for implementing Curation and preservation rules on mission data sets from initial conceptualisation or receipt through the iterative Curation cycle.

Submitted 14 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to Iolanda Maggio, Galileo Galilei, Frascati. Email: Iolanda.maggio@esa.int

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The paper presents specific aspects, of importance during the entire Earth observation data lifecycle, with respect to evolving data volumes and application scenarios. These particular issues are introduced in the section on 'Big Data' and LTDP. The Data Stewardship Reference lifecycle section describes how the data stewardship activities can be efficiently organised, while the following section addresses the overall preservation workflow and shows the technical steps to be taken during data curation. The paper concludes with introducing international collaboration for developing coordinated and harmonised lifecycle concepts.

Big Data and LTDP

'Big Data' indirectly addresses long-term data preservation issues: very large data sets handling, their curation, valorisation, retrieval, manipulation and finally visualization.
One of the most relevant 'Big Data' aspects is a new way of carrying out scientific research. Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive data sets.

Following experimental, theoretical, and computational science, a 'Fourth Paradigm' is emerging in scientific research. This refers to the data management techniques and the computational systems needed to manipulate, visualize, and manage large amounts of scientific data.

The main challenge is not only the volume of data, but its diversity, e.g. in format and type. Other major challenges are data structure and 'data on the move' i.e. transferring data through networks. This latter issue is a big inhibitor to jointly using data across distributed archives. Older Science and EO data are recorded on various devices, in different formats. A huge task represents the recovery, reformatting, reprocessing of such data, as well as the transcription of various associated information, necessary to understand and use the data. Challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization. A large proportion of users are not domain experts anymore, therefore data discovery tools, documentation and support are also needed.

Data Stewardship Reference Lifecycle

Earth Science data curation and preservation should be addressed during all mission stages – from the initial mission planning, throughout the entire mission lifetime, and during the post-mission phase. The Data Stewardship Reference Lifecycle (Figure 1) gives a high-level overview of the steps useful for implementing curation and preservation rules on mission data sets from initial conceptualisation or receipt through the iterative curation cycle.

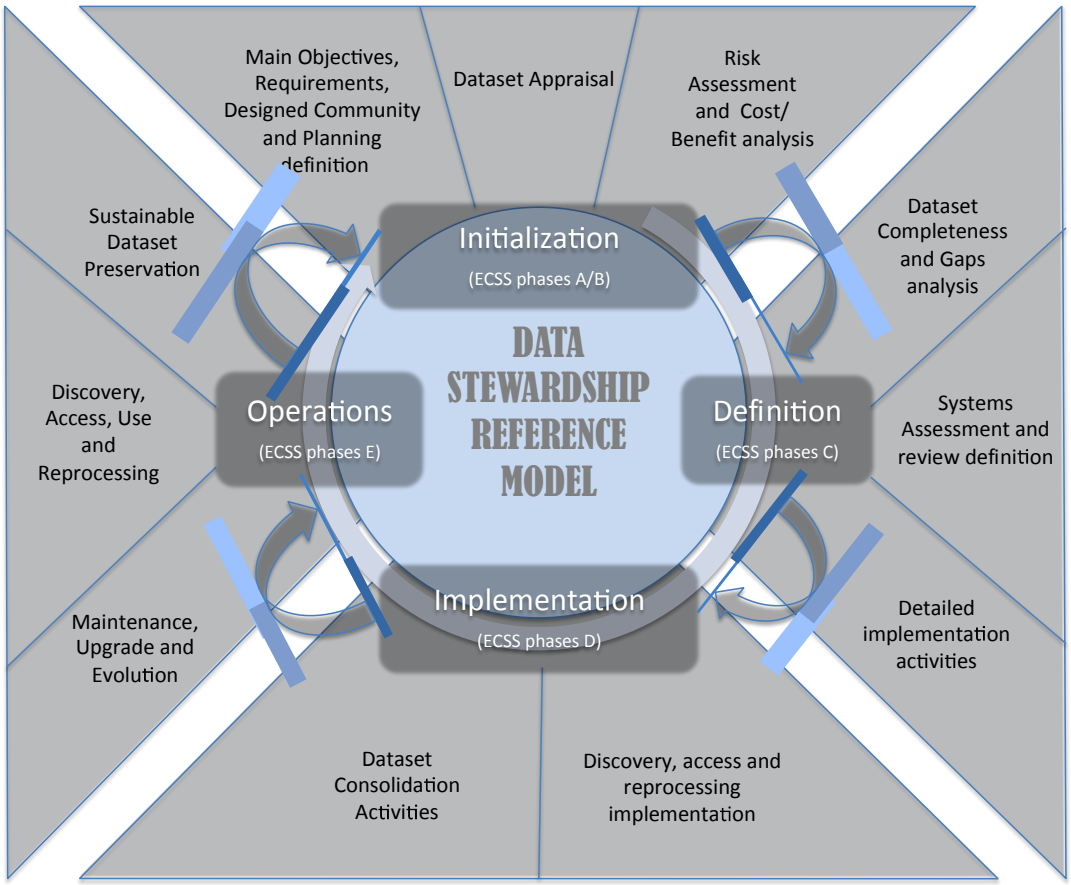


Figure 1. LTDP Data Stewardship Reference Lifecycle.

The core target of the LTDP lifecycle is the preserved data set, composed of consolidated:

1. Data records: these include raw data, Level 0 data and higher-level products, browses, auxiliary and ancillary data, calibration and validation data sets, and descriptive information.

2. Associated knowledge: this includes all the processing software used in the product generation, quality control, the product visualization and value adding tools, and documentation needed to make the data records understandable to the designated community. This includes among others mission operation concept, products specifications, instruments characteristics, algorithms description, Cal/Val procedures, mission/instruments performances reports, quality related information, etc. It is necessary to ensure data remain understandable and usable.

The final, consistent, consolidated, and validated "data records" are obtained by applying a consolidation process consisting of the following main steps:

1. Data collection

2. Cleaning/pre-processing

3. Completeness analysis

4. Processing/reprocessing

In parallel to the data records consolidation process, the data records knowledge, associated information and processing software are also collected and consolidated.

Data stewardship implements and verifies, for the relevant preserved data sets, a set of preservation and curation activities on the basis of a set of requirements defined during the initial phase of the curation exercise. Data preservation activities focus on Earth observation data sets long-term preservation, and are tailored according to its mission specific preservation/curation requirements. They consist of all activities required to ensure the "preserved data set" bit integrity over time, its discoverability and accessibility, and to valorise its (re)-use in the long term (e.g. through metadata/catalogue improvement, processor improvement for algorithm and/or auxiliary data changes and related (re)-processing, linking and improvement of context/provenance information, quality assurance). Preservation activities for digital data record acquired from the space segment and processed on ground embrace ensuring continued data records availability, confidentiality, integrity and authenticity as legal evidence to guarantee that data records are not changed or manipulated after generation and reception over the whole continuum of data preservation (archival media technology migration, input/output format alignment, etc.), valorisation and curation activities. The usage of persistent identifier for citation is part of the agency long term data preservation best practices.

Data curation activities aim at establishing and increasing the value of "preserved data sets" over their lifecycle, at favouring their exploitation, possibly through the combination with other data records, and at extending the communities using the data sets. These include activities such as primitive features extraction, exploitation improvement, data mining, and generation/management of long time data series and collections (e.g. from the same sensor family) in support to specific applications and in cooperation with international partners.

Data stewardship activities refer to the management of an EO Data set throughout its mission life cycle phases and include preservation and curation activities. It includes all the processes that involve data management (ingestion, dissemination and provision of access for the designated community) and data set certification.

Preservation Workflow

The LTDP data stewardship reference lifecycle is also represented through the preservation workflow, which defines a recommended set of actions to be sequentially implemented for the preservation of a "data set", with the goal of ensuring and optimizing its (re)-use in the long term. This preservation workflow, collaboratively developed with European space data holders, ensures that Earth observation mission data sets remain accessible and useable in the long term. Applying this workflow will produce a consolidated, accessible and useable Earth observation data set – consisting of the data records and the associated knowledge – and comprehensive documentation of the preservation procedure. While best initiated during the early mission planning phases, the preservation workflow can also be applied to data sets of current and historic Earth observation missions. The preservation workflow recommended actions/steps are the following:

1. EO missions/sensors data set appraisal, definition of designated community & preservation objective (with preservation/curation requirements)
2. Tailoring of mission specific consolidation process (on the data records)
3. EO missions/sensors data set PDSC tailoring and inventory table filling (including dependencies: Inventory Data Model)
4. Tailored PDSC consultation with designated community

5. Implementation of tailored consolidation process and collection of documentation and processing software
6. Update of EO missions/sensors data set PDSC & inventory table
7. Archive & ingestion, master inventory and catalogue population
8. Dissemination & Web configuration
9. Risk & cost assessment, preservation & cost planning, implementation.

The Schema presented below indicates the order in which these steps should be applied:

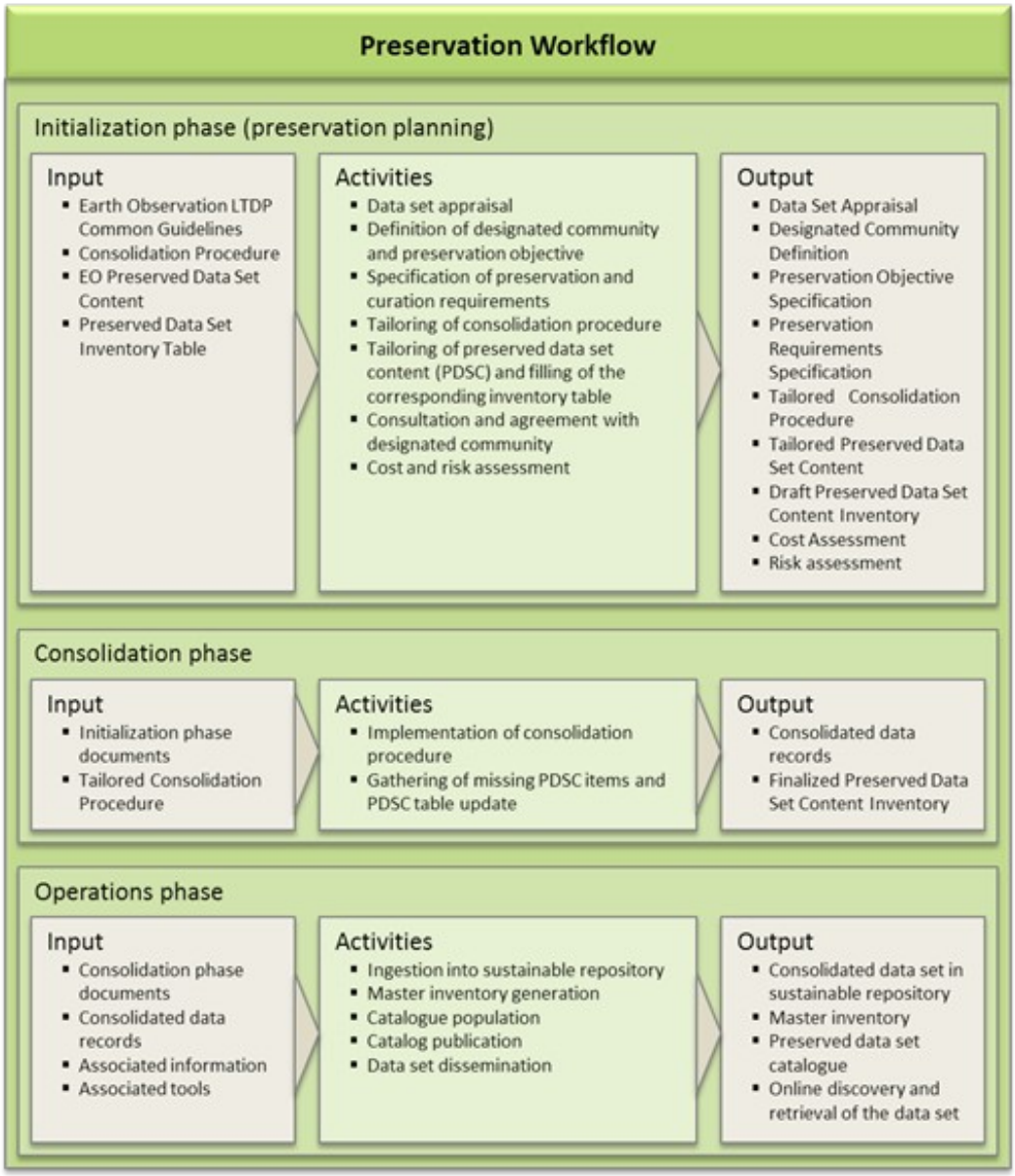
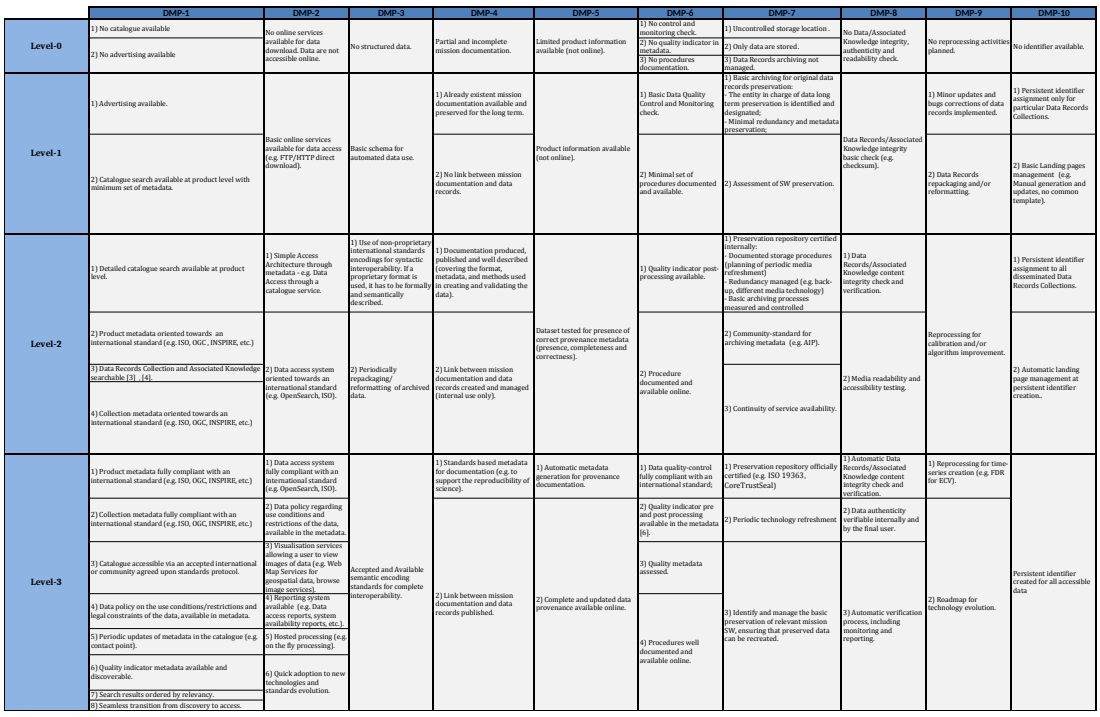


Figure 2. Preservation Workflow Steps.

WGISS Data Management and Stewardship Maturity Matrix

The scope of the on-going WGISS Data Management and Stewardship Maturity Matrix definition is to measure the overall preservation lifecycle and to verify the implemented activities needed to preserve and improve the information content, quality, accessibility, and usability of data and metadata. It can be used to create a stewardship maturity scoreboard of dataset(s) and a roadmap for scientific data stewardship improvement; or to provide data quality and usability information to users, stakeholders, and decision makers.

In the extended environment of Maturity Matrices and Models, the Maturity Matrix for "Long-Term Scientific Data Stewardship", of Ge Peng and Jeffrey L. Privette (2015), represents a systematic assessment model for measuring the status of individual datasets. In general, it provides information on all aspects of the data records, including all activities needed to preserve and improve the information content, quality, accessibility, and usability of data and metadata. This was used as a starting point of the WGISS Data Management and Stewardship Maturity Matrix. In parallel, the GEO Data Management Principles Task Force was tasked with defining a common set of GEOSS Data Management Principles (DMP-IG). These principles address the need for discovery, accessibility, usability, preservation, and curation of the resources made available through GEOSS.



	DMP-1	DMP-2	DMP-3	DMP-4	DMP-5	DMP-6	DMP-7	DMP-8	DMP-9	DMP-10
Level-0	1) No catalogue available 2) No advertising available	No online services available for data download. Data are not accessible online.	No structured data.	Partial and incomplete mission documentation.	Limited product information available (not online).	1) No control and monitoring check. 2) No quality indicator in metadata. 3) No procedures documentation.	1) Uncontrolled storage location . 2) Only data are stored. 3) Data Records archiving not managed.	No Data/Associated Knowledge integrity, authenticity and readability check.	No reprocessing activities planned.	No identifier available.
Level-1	1) Advertising available. 2) Catalogue search available at product level with minimum set of metadata.	Basic online services available for data access (e.g. FTP/HTTP direct download).	Basic schema for automated data use.	1) Already existent mission documentation available and preserved for the long term. 2) No link between mission documentation and data records.	Product information available (not online).	1) Basic Data Quality Control and Monitoring check. 2) Minimal set of procedures documented and available.	1) Basic archiving for original data records preservation: - The entity in charge of data long term preservation is identified and designated; - Minimal redundancy and metadata preservation; 2) Assessment of SW preservation.	Data Records/Associated Knowledge integrity basic check (e.g. checksum).	1) Minor updates and bugs corrections of data records implemented. 2) Data Records repackaging and/or reformatting.	1) Persistent identifier assignment only for particular Data Records Collections. 2) Basic Landing pages management (e.g. Manual generation and updates, no common template).
Level-2	1) Detailed catalogue search available at product level. 2) Product metadata oriented towards an international standard (e.g. ISO, OGC, INSPIRE, etc.) 3) Data Records Collection and Associated Knowledge searchable [3] , [4]. 4) Collection metadata oriented towards an international standard (e.g. ISO, OGC, INSPIRE, etc.)	1) Simple Access Architecture through metadata - e.g. Data Access through a catalogue service. 2) Data access system oriented towards an international standard (e.g. OpenSearch, ISO).	1) Use of non-proprietary international standards encodings for syntactic interoperability. If a proprietary format is used, it has to be formally and semantically described. 2) Periodically repackaging/ reformatting of archived data.	1) Documentation produced, published and well described (covering the format, metadata, and methods used in creating and validating the data). 2) Link between mission documentation and data records created and managed (internal use only).	Dataset tested for presence of correct provenance metadata (presence, completeness and correctness).	1) Quality indicator post-processing available. 2) Procedure documented and available online.	1) Preservation repository certified internally: - Documented storage procedures (planning of periodic media refreshment) - Redundancy managed (e.g. back-up, different media technology) - Basic archiving processes measured and controlled 2) Community-standard for archiving metadata (e.g. AIP). 3) Continuity of service availability.	1) Data Records/Associated Knowledge content integrity check and verification. 2) Media readability and accessibility testing.	Reprocessing for calibration and/or algorithm improvement.	1) Persistent identifier assignment to all disseminated Data Records Collections. 2) Automatic landing page management at persistent identifier creation..
Level-3	1) Product metadata fully compliant with an international standard (e.g. ISO, OGC, INSPIRE, etc.) 2) Collection metadata fully compliant with an international standard (e.g. ISO, OGC, INSPIRE, etc.) 3) Catalogue accessible via an accepted international or community agreed upon standards protocol. 4) Data policy on the use conditions/restrictions and legal constraints of the data, available in metadata. 5) Periodic updates of metadata in the catalogue (e.g. contact point). 6) Quality indicator metadata available and discoverable. 7) Search results ordered by relevancy. 8) Seamless transition from discovery to access.	1) Data access system fully compliant with an international standard (e.g. OpenSearch, ISO). 2) Data policy regarding use conditions and restrictions of the data, available in the metadata. 3) Visualisation services allowing a user to view images of data (e.g. Web Map Services for geospatial data, browse image services). 4) Reporting system available (e.g. Data access reports, system availability reports, etc.). 5) Hosted processing (e.g. on the fly processing). 6) Quick adoption to new technologies and standards evolution.	Accepted and Available semantic encoding standards for complete interoperability.	1) Standards based metadata for documentation (e.g. to support the reproducibility of science). 2) Link between mission documentation and data records published.	1) Automatic metadata generation for provenance documentation. 2) Complete and updated data provenance available online.	1) Data quality-control fully compliant with an international standard; 2) Quality indicator pre and post processing available in the metadata [6]. 3) Quality metadata assessed. 4) Procedures well documented and available online.	1) Preservation repository officially certified (e.g. ISO 19363, CoreTrustSeal) 2) Periodic technology refreshment 3) Identify and manage the basic preservation of relevant mission SW, ensuring that preserved data can be recreated.	1) Automatic Data Records/Associated Knowledge content integrity check and verification. 2) Data authenticity verifiable internally and by the final user. 3) Automatic verification process, including monitoring and reporting.	1) Reprocessing for time-series creation (e.g. FDR for ECV). 2) Roadmap for technology evolution.	Persistent identifier created for all accessible data

Figure 2. WGISS Data Management Stewardship Maturity Matrix.

The content of the WGISS Data Management and Stewardship Maturity Matrix represents the result of a combined analysis performed on the DMP-IG and a consultation at European level, with the Long Term Data Preservation Working Group. The rationales for applying the WGISS Data Management and Stewardship Maturity Matrix are:

1. Providing data quality, usability information to users, stakeholders, and decision makers;
2. Providing a reference model for stewardship planning and resource allocation;

3. Allowing the creation of a roadmap for scientific data stewardship improvement;
4. Providing detailed guidelines and recommendations for preservation;
5. Evaluating if the preservation follows best practices;
6. Giving a technical evaluation of the level of preservation and helping with self-assessment of preservation;
7. Providing a status of the preservation, but doesn't offer information on numbers or averages related to preservation;
8. Helping to break down problems related to preservation, and to understand the costs associated with each preservation level;
9. Funding agencies can define certain goal levels that they would.

Cooperation Activities

ESA is cooperating in the LTDP domain in Earth observation with European partners through the LTDP Working Group, formed within the Ground Segment Coordination Body (GSCB), and with other international partners, through participation to various working groups and initiatives. The EO LTDP framework international context is shown below:

Figure 3. Long-term Data Preservation Framework international context.

The LTDP core documents have also been reviewed and approved at international level within the Committee on Earth Observation Satellites (CEOS) and the Group on Earth Observations (GEO). A review of the Preservation Workflow document is currently on going in the frame of the CCSDS Data Archive Ingestion (DAI) working group.

Media Rescue Activity: Lessons Learned

Heritage data preservation activities include the preservation of unique data that can only be recovered from historical media. Therefore, the preservation of these media, together with the hardware that could read the media, should be ensured. During the rescue activity of JERS-1 mission media, some lessons learned were collected. Having no inventory available for the JERS-1 media at the Fucino ground station, several trips to the facility were undertaken in order to manually generate the media inventory. This was later compared against the JERS-1 data already available at ESA, which allowed to identify the missing data. However, this was not a simple task, as a large part of the media labels were either missing crucial information or this information could not be easily read, due to deterioration over time, as the storage environment was not systematically monitored.

The main lesson learned from this media rescue activity is that long-term preservation should be considered, and planned for, from the initial stages of a mission, in order to ensure that long-term data preservation policies are followed throughout the mission lifetime. Preservation of the main information on media labels and in local, digital, inventories should also be ensured, together with other Associated Knowledge. Furthermore, the original media, hardware and software should be preserved until it is certain that all unique data that could be recovered, was retrieved from the historical media. This also implies that the physical archiving storage must be located in a well-controlled environment that would prevent deterioration of the media labels or the media itself.

ERS and ENVISAT Consolidation Activities

The ERS-1, ERS-2 and Envisat missions constitute the European Space Agency's heritage in Earth Observation. Extending over many years, and covering numerous aspects of the Earth's systems, from atmosphere and ocean, to land and ice measurements, the EO data sets resulting from these missions hold significant scientific value and constitute a humankind asset. The main aim is to preserve these digital assets and to ensure their accessibility and usability for future generations.

The REAPER (REprocessing of Altimeter Products for ERS) project is performing a full reprocessing of both the ERS-1 and the ERS-2 Altimetry missions. The reprocessed data set spans from the start of the ERS-1 mission in July 1991 to June 2003, when the loss of the ERS-2 on-board data storage capability occurred causing the end of the ERS-2 global mission coverage.

The ERS-1/2 Low Bit Rate (LBR) data consolidation, gap filling and master dataset generation project (to be completed in mid 2015) is further refining the existing Level 0 master datasets for all ERS-1/2 LBR Instruments. This activity is also requiring the re-transcription of data from heritage media to fill identified gaps. The consolidated datasets will then be at the basis of further reprocessing campaigns in the future.

A similar project is addressing the consolidation of the ERS-1/2 SAR Level 0 master dataset including the repatriation of SAR data from National and Foreign stations in order to complete the master dataset available at ESA facilities.

The (A)ATSR SWIR Calibration and CLOUD Masking project aims at investigating the Long-Term stability of the ATSR instrument series, by building on early work on AATSR data and analysing the complete dataset to be available for the final users.

The project will provide a SWIR channel correction option and a set of calibration correction functions applicable to the currently available ATSR dataset.

The ERS/Envisat MWR recalibration project aims at deriving a homogeneous and fully error-characterised water vapour thematic climate data record (TCDR) based on the entire time

series of available data for ERS-1, ERS-2 and Envisat. This dataset will provide the backbone for atmospheric correction for ESA's critical altimetry missions. As such, the revised dataset will yield a positive impact on long-term stability and accuracy of radar altimetry products as well as provide uncertainty estimates on the tropospheric correction used in the mean sea level retrievals. High accuracy and stability is especially crucial for sea level trend analysis. This dataset will also provide a unique resource for climate research reaching back 20 years. With the launch of the Sentinel-3 altimetry instrument suite, a long-term perspective exists for extending the dataset.

AVHRR Times Series Generation

The session shows some details of our Advanced Very High-Resolution Radiometer (AVHRR)-archive, the importance of pre-processing, validation, and the product retrieval including time series of snow extent, albedo and lake surface water temperature. Finally, some recommendations will be given on the usage of AVHRR for climate applications and the next steps to make our archive accessible via ESA web site will be presented.

Data of the Advanced Very High-Resolution Radiometer (AVHRR) onboard of many NOAA- and since 2006 on EUMETSAT MetOp-satellites is the only source to provide long time series based on an almost unchanged sensor of the last 40 years. This long period fulfills the requirements of the World Meteorological Organization for a statistical sound analysis to study climate change induced shifts of Essential Climate Variables (ECV). Beside the central NOAA CLASS archive exist many local archives, which are partly not accessible, or the data holdings are not well maintained to be used by research teams. A few archives in Europe have data holdings covering Europe on a daily basis but are not homogenized and consolidated. In the frame of ESA's LTDP the data holdings of University of Bern (Switzerland), European Space Agency and Dundee Satellite System (UK) are combined, redundancy removed and validated to proof readability of the AVHRR LAC level 1b data. In a final step metafiles and quick looks are generated to be included in a data container (EO-SIP). The homogenized data set was transferred to ESA to keep the unique AVHRR images alive for the next +50 years and make it accessible for all interested parties via ESA web interfaces. The whole process for consolidation of the AVHRR data was in-line with the recommendations of the CEOS Working Group on Information Systems and Services (WGISS) to fulfil the needs for data management and stewardship maturity.

Conclusions

Data holdings are growing exponentially in Earth Science data archives worldwide. The European Copernicus program will continue to deliver Petabytes of valuable satellite-based Earth observations for many years to come. Only a systematic approach to data preservation during the entire data lifecycle, coordinated between data holders and application communities, will ensure that these data sets will be accessible and useable to current and future generations, for monitoring long-term variations in environmental parameters as a basis for objectively assessing and predicting effects of global change.

References

CEOS, "EO Data Preservation Guidelines Best Practices", http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Data%20Preservation%20Guidelines_v1.0.pdf

CEOS, "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

CEOS, "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

CEOS, "Associated Knowledge Preservation Best Practices",
http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS%20Best%20Practices/CEOS%20Associated%20Knowledge%20Preservation%20Best%20Practices_v1.0.pdf

CEOS, "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

CEOS, "Long-Term Preservation of Earth Observation Space Data: Glossary of Acronyms and Terms",
http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/White_Papers/EO-DataStewardshipGlossary_v1.2.pdf

CEOS, "CEOS Persistent Identifier Best Practices",
http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS%20Best%20Practices/CEOS%20Persistent%20Identifier%20Best%20Practices_v1.2.pdf

GEOSS, "Data Management Principles",
https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf

Peng, G. & Privette, J. (2015). Scientific Data Stewardship Maturity Matrix. Retrieved from
http://www.slideshare.net/gepeng86/scientific-data-stewardship-maturity-matrix