

## PRACTICE PAPER

# ESA EO Data Preservation System

Mirko Albani<sup>1</sup>, Michel Douzal<sup>1</sup>, Domenico Castrovillari<sup>2</sup>, Paolo Boezi<sup>3</sup>,  
Daniele Iozzino<sup>3</sup> and Iolanda Maggio<sup>3</sup>

<sup>1</sup> European Space Agency, IT

<sup>2</sup> Intecs Solutions, IT

<sup>3</sup> Rhea Group, BE

Corresponding author: Iolanda Maggio (Iolanda.Maggio@esa.int)

The European Space Agency (ESA) has the mandate to assure the long-term preservation, sharing and exploitation of space data and its associated knowledge. ESA's aim is to turn space exploration and space-related activities into an overall societal project involving a wide variety of stakeholders. To this end, it brings together and coordinates as many countries as possible under the banner of space missions. It is a basic principle that ESA deals with its stakeholders openly and with real transparency, an approach that has contributed to its long-term success.

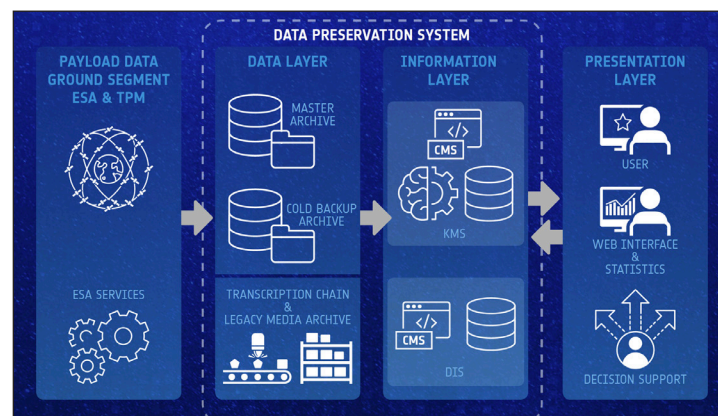
The Earth Observation (EO) Data Preservation System has the main objective of providing the required infrastructure and services to assure ESA and Third Party Missions (TPM) EO Data Records and Associated Knowledge preservation and accessibility, and to support the cooperation activities with national and international organizations in the data preservation domain. The generic "EO Missions/Sensors Preserved Dataset" content includes Data Records and Associated Knowledge.

**Keywords:** Long Term Archive; Provenance; Data Preservation; Preservation Processes

## 1. Introduction

The EO Data Preservation System has the main objective of providing the required infrastructure and services to assure ESA and Third Party Missions (TPM) EO Data Records and Associated Knowledge preservation and accessibility, and to support the cooperation activities with national and international organizations in the data preservation domain. The generic "EO Missions/Sensors Preserved Dataset" content includes Data Records and Associated Knowledge.

The main components of the Earth Observation (EO) Data Preservation System involved in the preservation process are the Management System for Data and Associated Knowledge (KMS), Data Information System (DIS), Master Archive (MAR), Cold Back-up Archive (CBA). The picture below shows the EO Data Preservation System (**Figure 1**).



**Figure 1:** EO Data Preservation System.

The Master Archive and the Cold Back-up archives cover the archiving functionalities whereas the Data Information Service (DIS) provides the information, the history and the provenance of the data archived. The concept of Data Information Service arises from the service requirements ESA included within the Data Service Initiative (DSI) program modelling systems and processes to enable the management of ESA's EO Data as a standard asset. These requirements ensure that metadata for each product is collected in a database and the data element itself is systematically stored with full track of all value added to the data during the service activities. The space data archived are in EO-SIP format being the standard for most ESA Earth Observation missions. EO-SIP represents the package which includes the actual product in its native format, a quality report, a quick-look picture and metadata.

The metadata is an xml file following the OGC standards "Earth Observation profile of Observations & Measurements (OGC EOP O&M)" OGC 10-157. The underlying database storing all information is well suited to report on the activities within the DSI and other services but also to present zoom-able level of detail for any EO Data asset held by ESA, making the tool useful to staff involved in operational support to management and decision-making. In addition, there are other sources of operational data and repositories around the data payload ground segment. The huge value of the information contained in all these systems is enhanced by providing a harmonised, service-independent view and control of the data assets held across system in order to provide end-to-end operational analysis of data assets to pinpoint changes, errors or discrepancies. The crucial factor determining the success of the DIS is the ability to receive the metadata, recognise products across source services, adjust obtained information and produce the unequivocal set of attributes for any data asset. This paper describes the features of the system and any relevant preservation processes.

## 2. Master Archive

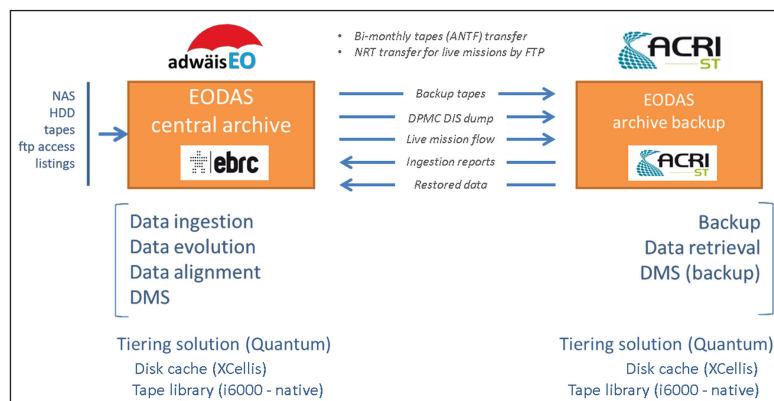
The ESA Master Archive is implemented through a dedicated service (DAS) awarded to industry through an open Invitation To Tender (ITT) in 2016. ESA has outsourced EO data archiving activities to a single provider with the goal to benefit from economy of scales and standardization of data archival & delivery processes and interfaces.

These archiving activities shall cover:

- Historic ESA flagship mission data (ERS-1/2, Envisat)
- Future, active and historic ESA Earth Explorer mission data
- Future, active and historic ESA Third Party Mission data
- Campaign data acquired in the context of ESA EO airborne campaigns
- Input and Output data of services managing EO data (e.g. DSI)
- Any EO data acquired/obtained by the Agency from other sources (e.g. NASA, USGS etc.).

The industrial consortium is made of ACRI-ST (FR), adwäisEO (LU) and KSAT (NO), with the following distribution of roles (**Figure 2**):

- **ACRI-ST**: data archiving (and delivery) service provider – Prime contractor
- **adwäisEO**: data archiving (and delivery) operation provider with its leading edge IT and connected secured infrastructure in Luxembourg
- **KSAT**: data knowledge provider and overall ingestion validation



**Figure 2:** Master Archive service locations and duties in Europe.

In order to guarantee the data safety, the archive is distributed between two locations sited >200 km apart; a master archive in Luxembourg and an archive back up in Sophia Antipolis (France). The archival operational flow between the two facilities is depicted in **Figure 1**, which also lists the technical solutions for the infrastructure in each data centre. Several data quality checks in terms of reliability of the process during data transfer are performed all along the flow (data not corrupted during transcription or during copy to the backup center, data still readable on tape).

The Master Archive infrastructure is mainly based on two similar Quantum iScalar 6000 libraries connected via 10+ GE and 16 Gbps SAN links to DELL M1000e + M6x0 blade enclosure and servers. A StorNext System is used to manage the data in Hierarchical Storage Management (HSM) mode. In this environment, the disk structure containing the data is exported to NFS clients as a classical Linux volume and specific policies determine the way to store the data (keep on disk and/or tape, automatic generation of several copies...). Data are copied to LTO7 tapes in native Quantum format (ANTF):

- 1 main copy for the central archive
- 1 temporary copy for the backup archive for data transfers between main and backup archive centres as shown in **Figure 2**.

Finally, the content of the temporary tape is then restored in the Backup Centre library.

**Figure 3** provides an high-level view of the full process. This technical solution offers full scalability and shall cope with ESA requirements in case of unexpected growth of service needs.

One of the most important activities performed by the Master Archive service is the data quality check, which is performed not in terms of scientific content but in terms of reliability of the process during data transfer. This verification includes the points summarized in the following questions:

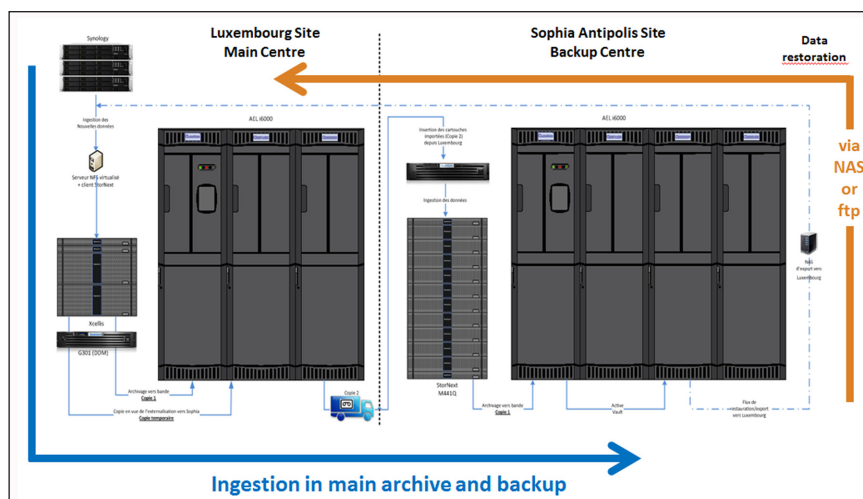
- are all input data archived?
- are we sure that the archived data has not been corrupted during the transcription?
- are we compliant with the requirements related to the minimum distance between the two archive copies?
- are we sure that the stored data is still retrievable and usable?
- are we sure that the retrieved data has not been corrupted?

The following **Figure 4** shows the verifications performed at the various stages of the data ingestion into the archive.

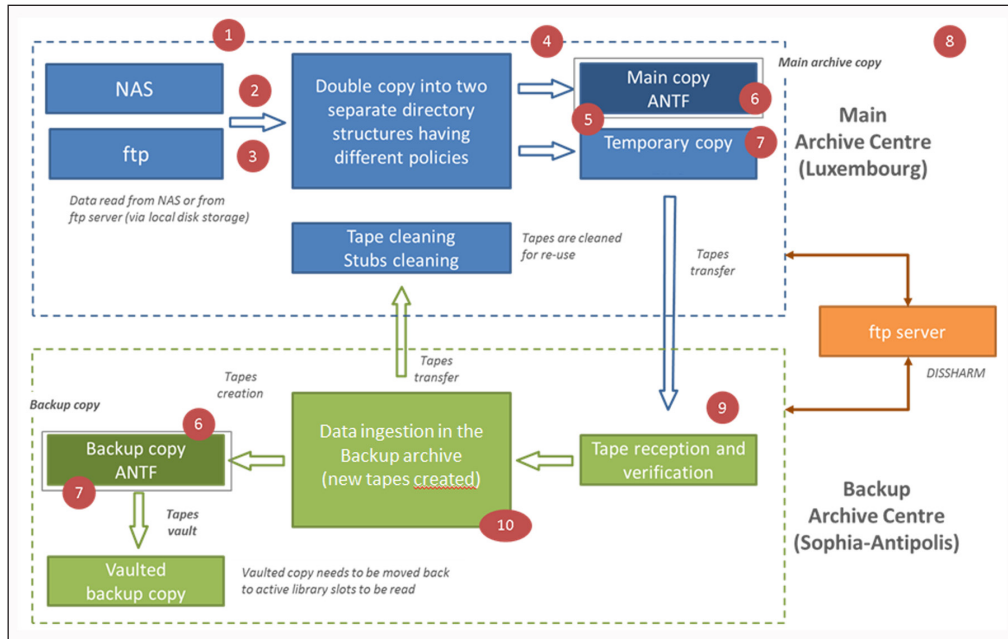
Checks performed during the data ingestion can be found in the following **Table 1**.

The Master Archive (DAS) ingests data from historical missions according to a data ingestion plan constantly maintained by ESA Data Librarian as well as data coming from live missions (currently Cryosat-2, SMOS, ADM-Aeolus and OceanSat-2) upon formal requests by the relevant ESA Operations Managers.

The status and progress of the data ingestion is visible in real time at the service web portal at <http://www.eodas.info> (**Figures 5, 6, 7, 8**).



**Figure 3:** Master Archive infrastructure and high-level data flow between archive centers.



**Figure 4:** Checks performed during the data ingestion process.

**Table 1:** Checks performed by the Master Archive during the data ingestion process.

| #  | Centre | Operation  | Verification  |
|----|--------|------------|---|
| 1  | Main   | Ingestion  | The list of files is compared to any delivered inventory.   |
| 1' | Main   | Ingestion  | The structure and the content (repository, datasets, products, files) is compared with the delivery spreadsheet delivered by ESA.   |
| 2  | Main   | Ingestion  | The total number of products is compared to the delivery information.   |
| 3  | Main   | Ingestion  | The actual file MD5 checksum is compared to the value extracted from the product metadata (manifest file or attached checksum file).  |
| 3' | Main   | Ingestion  | If data is included in a container (zip, tgz, ...), the integrity of the container is verified (i.e. container content can be accessed).  |
| 4  | Main   | Ingestion  | After the generation of the zip container, the files are extracted in a scratch directory and checked with respect to the content of the checksum.txt file. This operation is logged for future use by the global verification of the dataset ingestion.  |
| 5  | Main   | Ingestion  | MD5 hash code computed by Quantum from the products are queried from the StorNext database and compared to the DMPC MD5 hash code before products are set in "TAPE" status.   |
| 6  | Both   | Ingestion  | LTO-7 drives apply an automatic verify-after-write technology to immediately check the data as it is being written.   |
| 7  | Both   | Backend    | Both Quantum libraries include EDLM. The Extended Data Life Management feature ensures that tapes are trouble-free (based on tape scan and tape memory analysis). Tapes scan and analysis is performed following predefined policies (max. 4 tapes per day i.e. 7% of a 10 PBytes archive per month using 2 LTO7 EDLM drives). Suspect tapes are automatically copied to new tapes. |
| 8  | Main   | Validation | After the ingestion process where data has been verified at product level, a global ingestion verification is performed using the DPMC database information, the initial media inventory and the ingestion process log files.   |
| 9  | Backup | Ingestion  | The inventory of the tapes sent to the backup centre is retrieved and used for comparison with the copy process performed to copy the products from temporary tapes to ANTF tapes via disk cache.   |
| 10 | Backup | Ingestion  | The zip container integrity is used to verify that the transferred products have not been corrupted.  |

Figure 5: EODAS Service Web Portal.

| mission | instrument | product_type | dataset | version | nb_prod | size(KB)  | source     |
|---------|------------|--------------|---------|---------|---------|-----------|------------|
| ADEOS   | AVNIR      | L0           | DL228   | 1.0.0   | 2354    | 295958641 | Maspalomas |
| ALOS-1  | AUX        | AUX_COI_PD   | DL236a  | 1.0.0   | 1946    | 288123    | GMV        |
| ALOS-1  | AUX        | AUX_COI_DT   | DL236b  | 1.0.0   | 1933    | 286199    | GMV        |
| ALOS-1  | AUX        | AUX_COR_DT   | DL236c  | 1.0.0   | 1933    | 286199    | GMV        |
| ALOS-1  | AUX        | AUX_COR_PD   | DL236d  | 1.0.0   | 1946    | 288123    | GMV        |
| ALOS-1  | AUX        | AUX_CTM_AX   | DL236e  | 1.0.0   | 1910    | 1331126   | GMV        |

Figure 6: EODAS overall historical data ingestion status.

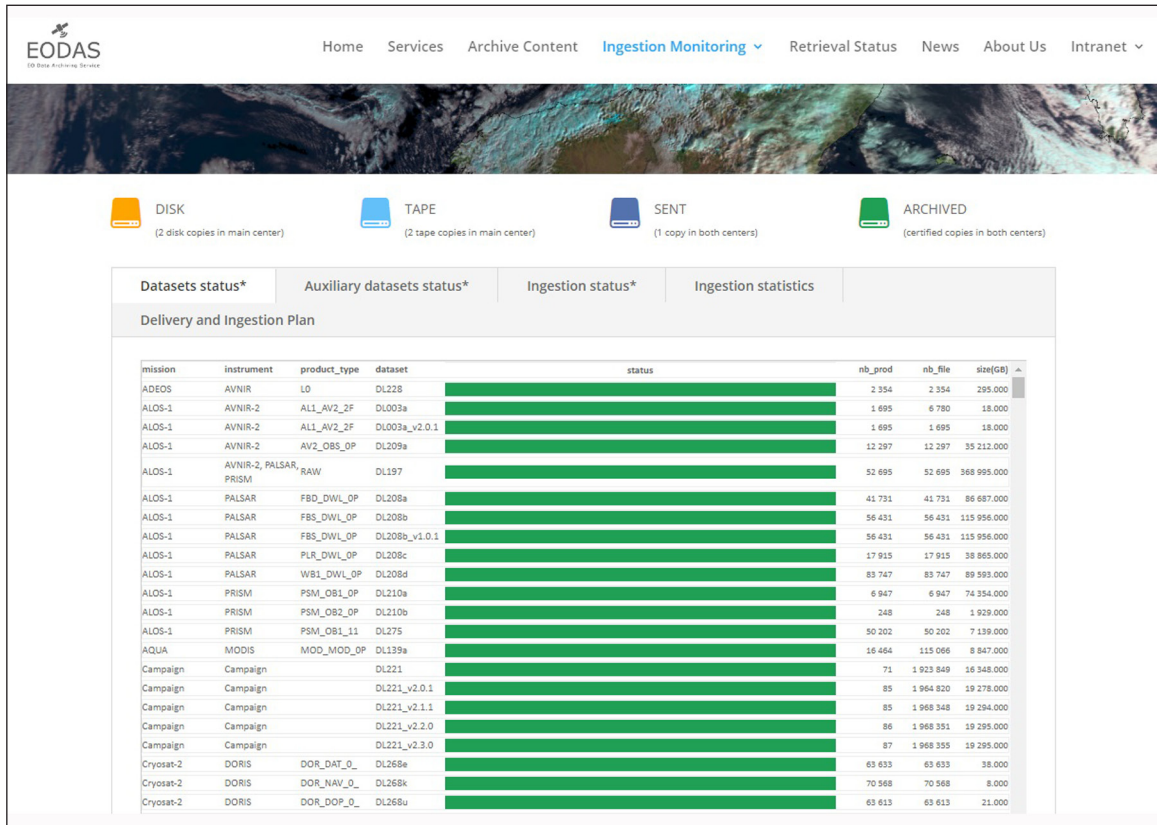


Figure 7: EODAS historical data ingestion detail.

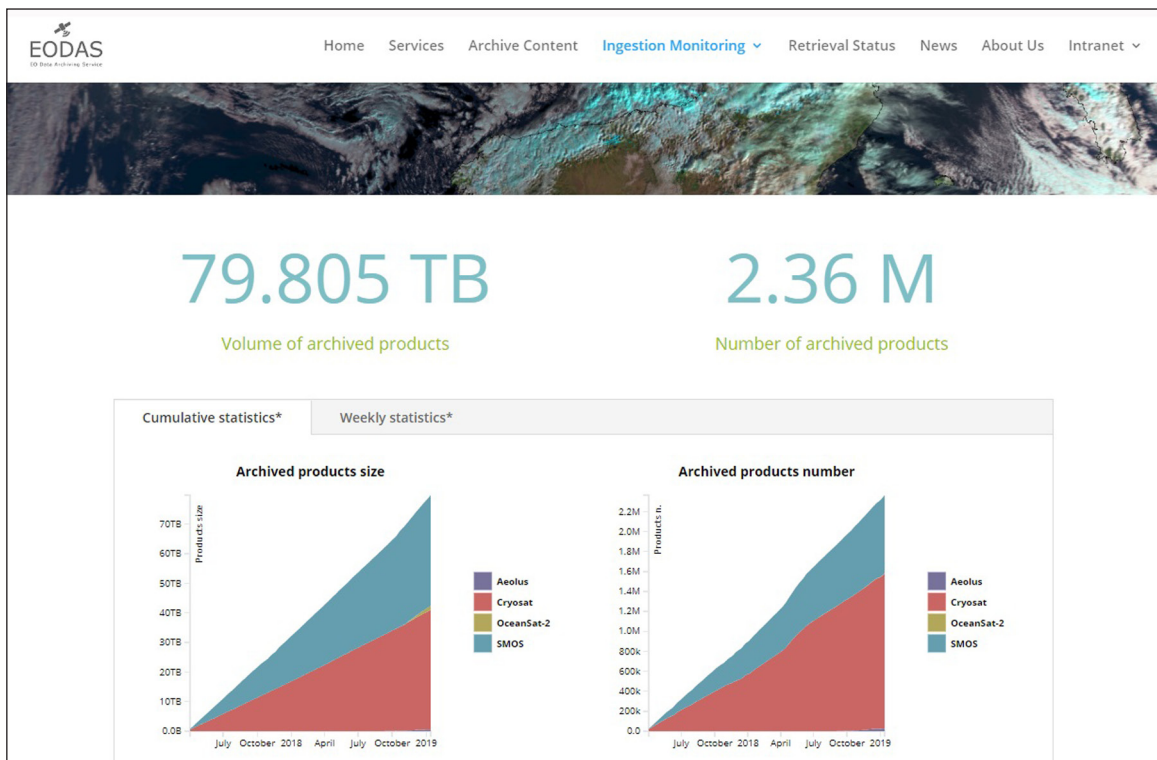


Figure 8: EODAS live mission ingestion status.

The EODAS Service Web Portal is updated on hourly basis with continuous injection of fresh information coming from the core processes that drive the data archiving. Lots of views and filters are available to select the information of interest allowing also making exports in pdf format or excel tables for any further analysis and statistics of the archived products.

### 3. Cold Back-Up Archive

The Cold Back-up Archive (CBA) has been implemented in ESRIN (ESA premise in Frascati - IT) in 2014, throughout the years, it has been upgraded to enhance performances and allow seamless archive and extraction capabilities. It currently contains (**Figure 9**):

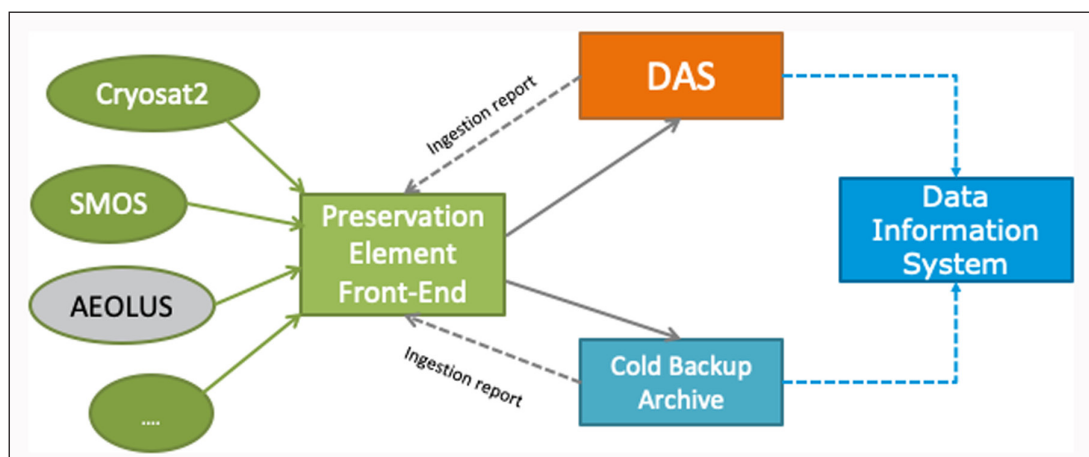
- Unique data
  - Historical data repatriated from ESA Processing and Archiving Centres
  - ESA and/or Third-Party Missions Earth Observation (EO) Unconsolidated data
- ESA and Third-Party EO Live Mission data
  - Aeolus
  - Cryosat-2
  - Oceansat-2
  - Proba-1
  - Odin
  - Scisat
  - Smos
  - Swarm
- Reprocessing campaign data
- Second Copy of the Master Archive

In the scope of the Inter-directorate Joint Activities, a dedicated circulator software has been implemented to allow reception of ESA Science data from both the internet and ESA WAN connected centres. The CBA is therefore being populated with:

- ESA Science active and historical Data with the following missions complete:
  - Cassini
  - Earth
  - Exosat
  - Giotto
  - HST
  - ISO
  - SMART-1
  - Ulysses
  - VEX

ESA and Third-Party active mission data are circulated by the Preservation Element Front-End (PE-FE). Once the data has been ingested and validated, confirmation reports are sent to the Mission Payload Data Ground Segment via standard network transfer protocol. Bulk dissemination of data coming from reprocessing campaigns is circulated on storage devices. The DAS archive is the Master Archive.

The volume of ingested data by the CBA is 7,9 PB of data for a total of 129 million of files (**Figure 10**).

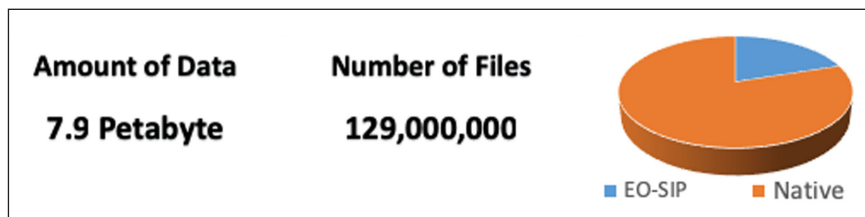


**Figure 9:** Front-End Data Circulation.

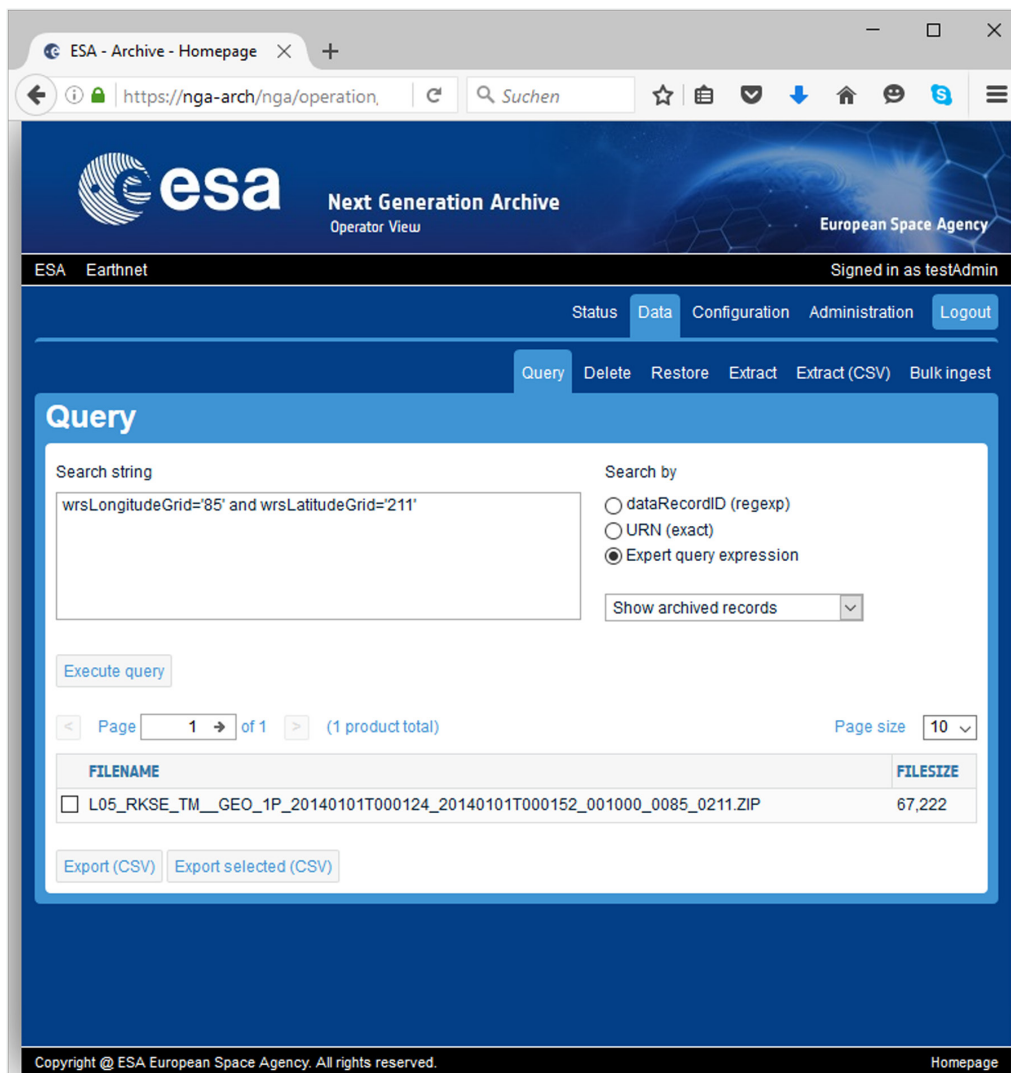
The CBA archive can handle the following format of data:

- ESA EO Submission Information Package (EO-SIP)
- Standard Archive Format for Europe (SAFE)
- Native format data
- Documentation data

Data is handled by an ingestion software specifically developed for the matter; the Next Generation Archive (ngA) which validates the data and extract metadata when available. Metadata is used to populate its Database to be used for advanced queries. The Next Generation Archive (ngA) (**Figure 11**) is the software we use to ingest EO-SIP packages into the Cold Backup Archive. The ingestion software inspects the package and populate its database with data extracted from the EO-SIP included metadata xml files.



**Figure 10:** CBA Volumes, January 2019.



**Figure 11:** CBA Next Generation Archive.



The ngA replaces the formed ingestion and archiving software that had data compatibility constrains.

The infrastructure of the CBA consists of two Robotic Libraries, capable of storing 24 PB of online data each. The Main Library is a 3000 slot STK SL8500, equipped with 8 T10000D, 4 robotic arms and 10 T10000B drives, the Disaster Recovery Library is a 3000 slot STK SL3000 with 4 LTO-7 drives (due to be upgraded to LTO-8 in 2019).

Data is initially written on an SSD based cache and then moved to the different tiers of the storage by the ORACLE Hierarchical Storage Manager (Oracle HSM) where the final tier is the tapes of the Robotic Archives.

Synchronization between the two libraries is performed through a dedicated 16 Gb/s fibre optic connection.

#### 4. Data Information System

The Data Information System (DIS) is a management support system based on EO data products. DIS is in charge of managing metadata information for all products generated, used and distributed by any activity (project or service) for all the ESA and Third Party Missions. DIS has been developed within the same Contract, which implements the service for data consolidation and reprocessing (DSI) as an extension of the internal inventory system. DSI is managed by Serco Italy.

The metadata available in DIS is extracted by Master Archive from the products managed during the ingestion phase, as part of the archiving process aimed to populate the local inventory. The format mainly used for products distribution by ESA is the EO-SIP format, where metadata is explicitly provided in the package by dedicated xml files, but metadata is extracted as well from all products using common formats. All the information available is stored, including key attributes for data access and quality information when available.

The original core of the system has been upgraded to manage the information about products available at many different systems and the relationship among them and the history of data. DIS supports awareness and control of the data and the value added to the data, as the system is keeping trace of any metadata for all products stored at different processing and archiving sites. DIS is consisting of a database repository, an Extraction, Transformation and Loading (ETL) module and a Business presentation layer. DIS facilitates improved control of the EO data owned by ESA, concentrating information spread over many different external services in a single place, and supporting verification that all data is aligned at all different sites and distribution of the right data for the activities requesting it. DIS is fed from different services at ESA; each service is providing all the available information about its data, how it is organized and the changes applied to it. DIS collects all the information, integrating what has been received to build up the full set of metadata for each product. A major requirement for DIS is to trace history of data, the changes applied in the past, what are the datasets including this product, what the different versions of the datasets consolidated over the time, what is the version accessible for users requesting it (Figure 12).

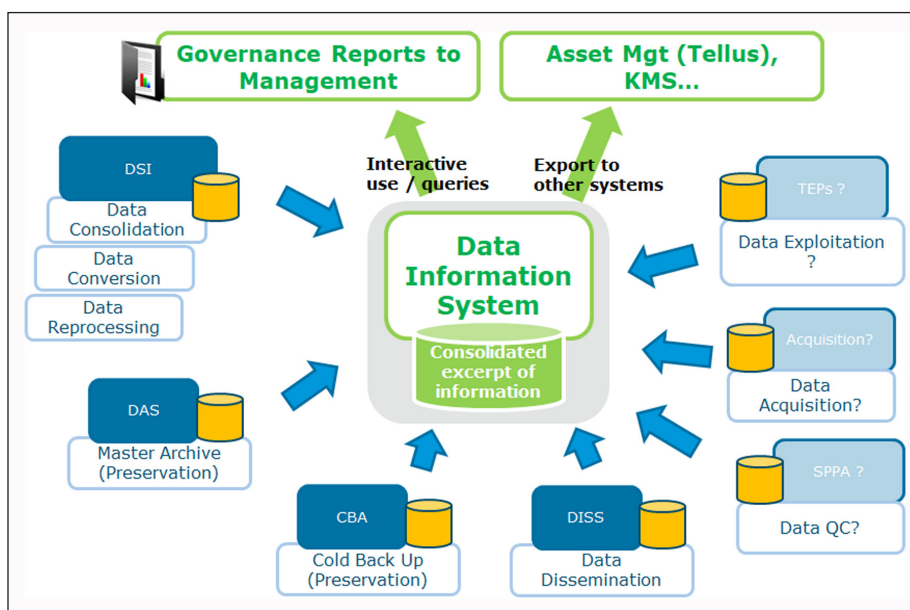


Figure 12: Data Information System.

DIS has initially started collecting data from the DSI projects, and now it is integrating with DAS. In a near future additional service will be added to complement the scenario as repository of data, specifically Cold Backup and Dissemination Services. Further extensions of DIS coverage will be analysed later on. The information available in DIS is accessible through the Business Intelligence layer on top of it, allowing to design and publish on web the reports defined by users to monitor and control their data. A set of predefined reports supports direct access to the most common views over the data, and dedicated reports are being developed for specific tasks. The presentation layer of DIS is supporting drill-in on the charts presented, allowing to select the specific data of interest to retrieve high level information up to the full detail.

For each product DIS is storing and referencing not only metadata, but cross-references to other products and processes related to the product, as

- The datasets to which the product is associated/disassociated, and the versions involved
- The changes applied to the product, and history of different versions available
- The auxiliary files used to generate the product, and the versions of each auxiliary file used
- The lower level products used to generate it
- The products generated from the product during any further processing activity

Most of the above information is already available on the presentation layer, or it will be added in the next future. Analysing large sets of data at product level cannot easily managed, and therefore DIS provides full support for the dataset level, where products can be freely aggregated to highlight common characteristics. For each dataset, DIS stores:

- Information available at each archiving site
- List of products associated to each dataset
- History of all versions created for the dataset, and the changes applied a dataset version for migrating to the next version
- How the dataset was generated, using which datasets and auxiliary datasets, and by means of which transformations

DIS has been designed to support requirements from different user categories:

- Management can monitor the data holdings and respect of archiving policies for all data in the organization and the value added by the organization
- Data engineers can verify products replication over different sites, alignment of repositories and availability of data requested by users
- Projects can retrieve information about data requested with the full history of how this data was generated and archived
- Operators can follow data ingestion activity, checking errors and issues to ensure feeding of DIS for smooth operations

A campaign of interviews with all user profiles of the system allowed the design team to filter driving requirements and usage scenarios for the implementation of a stable core architecture of DIS. The reporting interface is subject to continuous improvements, as new requests for dedicated views on the data result in additional reports taking advantage of the powerful BI tool selected.

Access to the DIS reporting interface is web-based, supporting user profiles for accessing dedicated or restricted resources. Reports are grouped to allow a direct access to the reports of interest for each user category.

The provenance view on the DIS website shows graphically the relationship between a dataset and the others, and the processes applied to it, using a high-level view, easy but very helpful for the understanding of the data involved (**Figure 13**).

Verification of the Archiving Policy applied to products is critical, as it ensures correct preservation of data but avoiding the costs of an uncontrolled replication (**Figure 14**). DIS allows a monitoring of how many copies are being stored for each product (**Figure 15**).

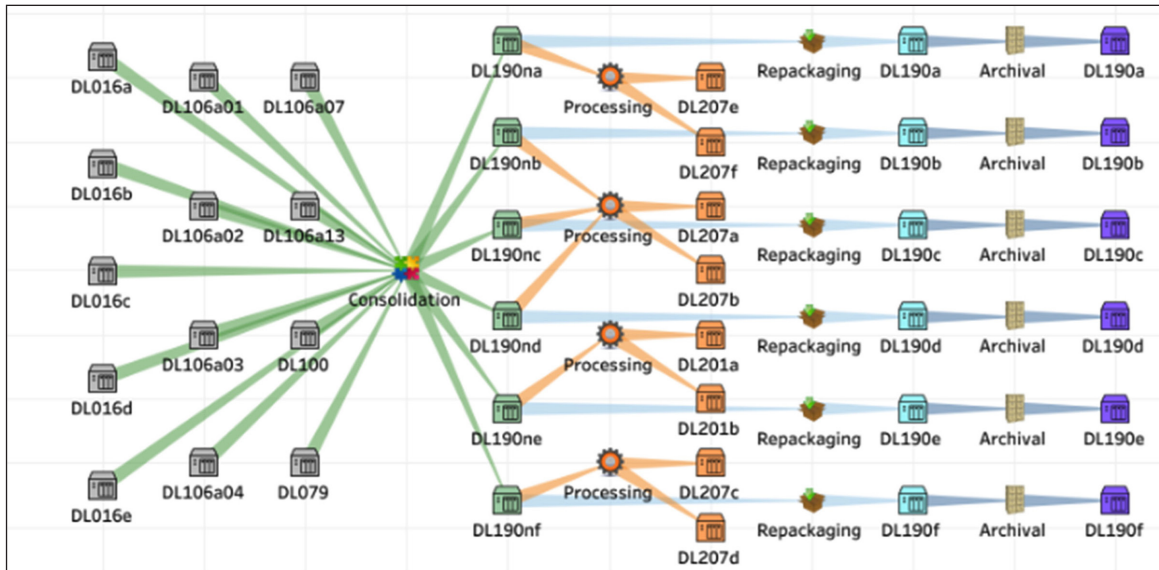


Figure 13: DIS Data Provenance.

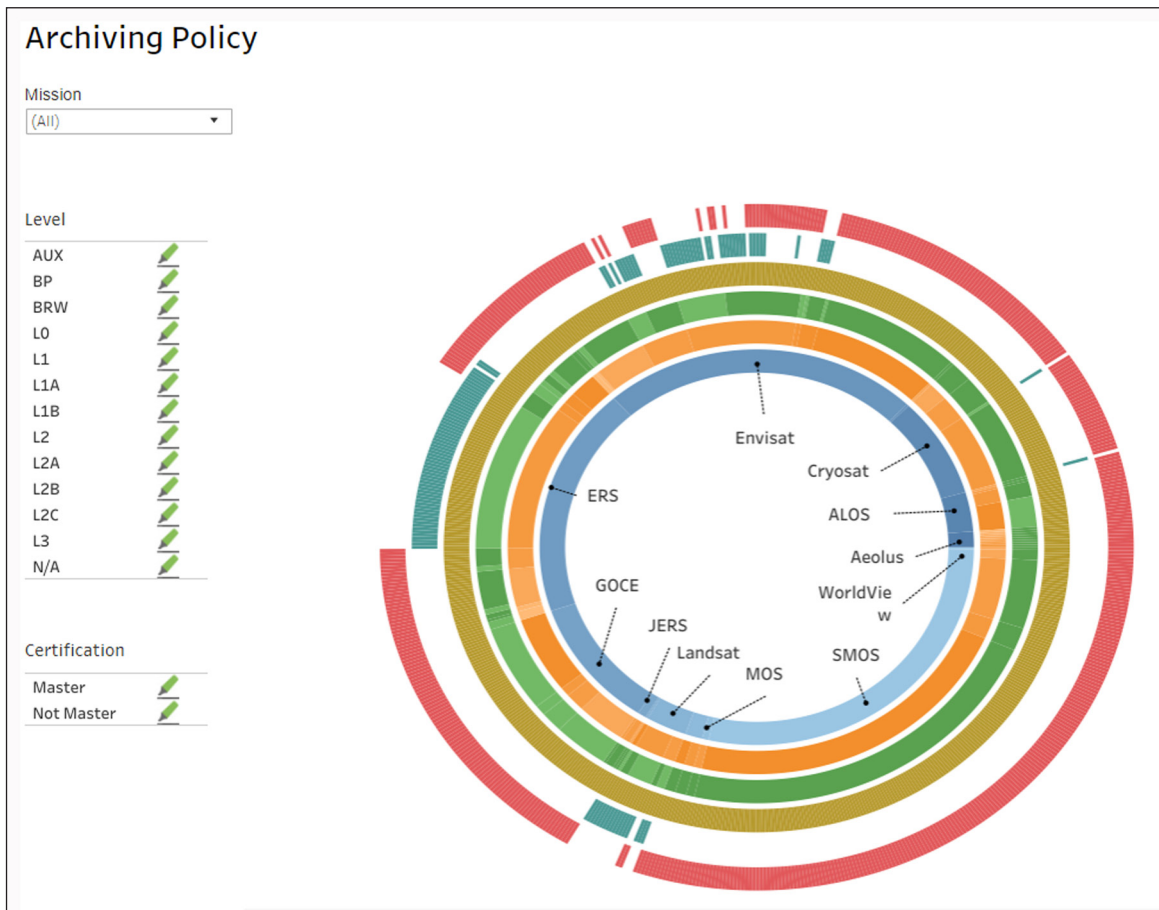


Figure 14: Archiving Policy.

Events in the lifecycle of each product can be monitored, assessing “activity” for each product type in terms of number of projects working on it and resources allocated (Figure 16).

The DIS repository maintaining the knowledge of data available, their attributes and relationships is a fundamental tool for the full organisation, and it will gear new reporting features in the future, allowing

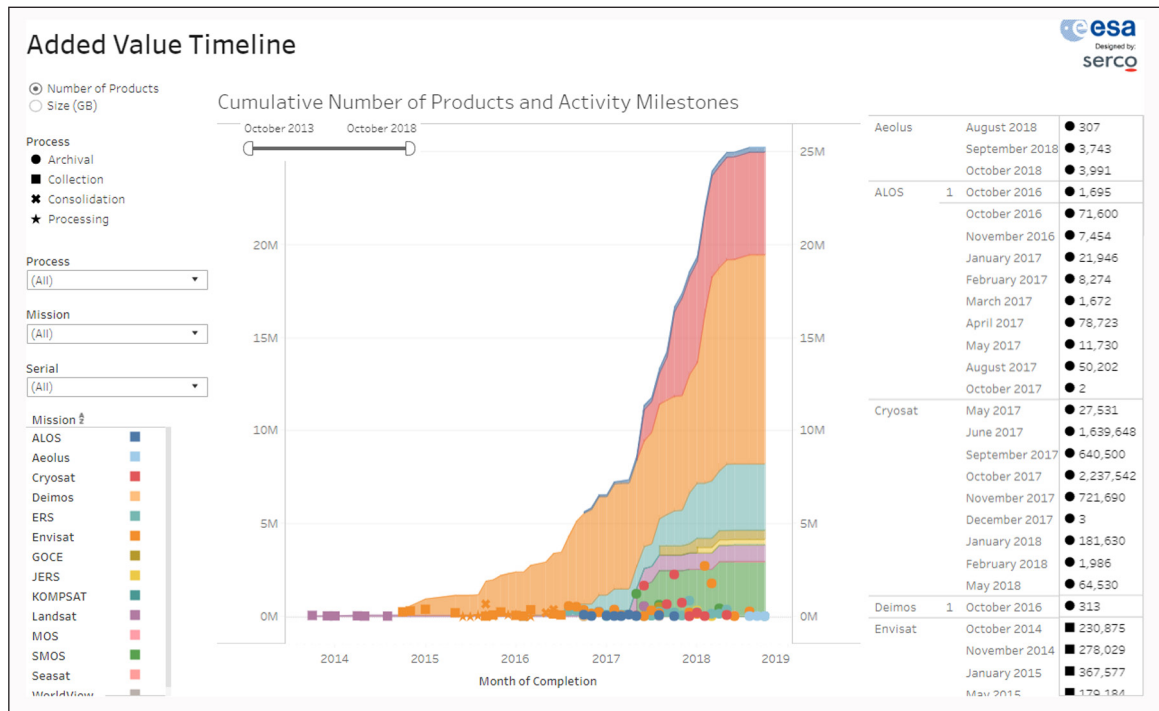


Figure 15: Data Monitoring.

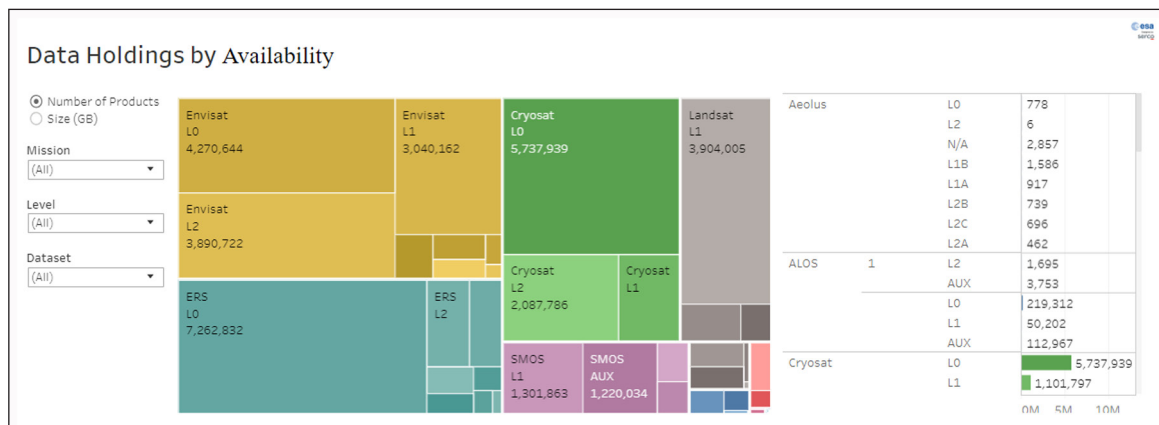


Figure 16: Data Holdings by Availability.

to exploit and analyse new unexpected and exciting relations among the data in way that we can't yet imagine.

### 5. Conclusion

Since its birth, the EO Data Preservation System has allowed ESA and its Heritage Data Program (LTDP+) to preserve both the data and the knowledge, and being compatible with the Open Archival Information System (OAIS) of Consultative Committee for Space Data System (CCSDS) and relevant preservation standards and guidelines. Both Archives manage the Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP) in line with the OAIS functional Model.

The Data Preservation System will be continuously supported and enhanced in terms of both scope and functions, in order to ensure the preservation and valorisation of the ESA Earth Observation Assets.


### Competing Interests

The authors have no competing interests to declare.

**How to cite this article:** Albani, M, Douzal, M, Castrovillari, D, Boezi, P, Iozzino, D and Maggio, I. 2020. ESA EO Data Preservation System. *Data Science Journal*, 19: 20, pp.1–13. DOI: <https://doi.org/10.5334/dsj-2020-020>

**Submitted:** 26 January 2019    **Accepted:** 07 January 2020    **Published:** 07 May 2020

**Copyright:** © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

**OPEN ACCESS** 