The International Journal of Digital Curation

Issue 3, Volume 4 | 2009

Long-term Preservation of Earth Observation Data and Knowledge in ESA through CASPAR

Sergio Albani,

ACS c/o ESA-ESRIN, Italy

David Giaretta,

STFC Rutherford Appleton Lab, UK

Abstract

ESA-ESRIN, the European Space Agency Centre for Earth Observation (EO), is the largest European EO data provider and operates as the reference European centre for EO payload data exploitation. EO Space Missions provide global coverage of the Earth across both space and time generating on a routine continuous basis huge amounts of data (from a variety of sensors) that need to be acquired, processed, elaborated, appraised and archived by dedicated systems. Long-term Preservation of these data and of the ability to discover, access and process them is a fundamental issue and a major challenge at programmatic, technological and operational levels.

Moreover these data are essential for scientists needing broad series of data covering long time periods and from many sources. They are used for many types of investigations including ones of international importance such as the study of the Global Change and the Global Monitoring for Environment and Security (GMES) Program. Therefore it is of primary importance not only to guarantee easy accessibility of historical data but also to ensure users are able to understand and use them; in fact data interpretation can be even more complicated given the fact that scientists may not have (or may not have access to) the right knowledge to interpret these data correctly.

To satisfy these requirements, the European Space Agency (ESA), in addition to other internal initiatives, is participating in several EU-funded projects such as CASPAR (Cultural, Artistic, and Scientific knowledge for Preservation, Access and Retrieval), which is building a framework to support the end-to-end preservation lifecycle for digital information, based on the OAIS reference model, with a strong focus on the preservation of the knowledge associated with data.

In the CASPAR Project ESA plays the role of both user and infrastructure provider for one of the scientific testbeds, putting into effect dedicated scenarios with the aim of validating the CASPAR solutions in the Earth Science domain. The other testbeds are in the domains of Cultural Heritage and of Contemporary Performing Arts; together they provide a severe test of preservation tools and techniques.

In the context of the current ESA overall strategies carried out in collaboration with European EO data owners/providers, entities and institutions which have the objective of guaranteeing long-term preservation of EO data and knowledge, this paper will focus on the ESA participation and contribution to the CASPAR Project, describing in detail the implementation of the ESA scientific testbed.[1]

[1] This paper is based on the paper given by the authors at the 5th International Digital Curation Conference, December 2009; received November 2009, published December 2009.

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. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

ESA-ESRIN, the European Space Agency[2] establishment in Frascati (Italy), is the largest European Earth Observation (EO) data provider and operates as the reference European centre for EO payload data exploitation.

EO data provide global coverage of the Earth across both a continuum of timescales and a variety of geographical scales. EO data are generated by many different instruments producing multi-sensor data in long time series with a variable geographical coverage, variable geometrical resolution and variable temporal resolution. EO data acquired from space therefore constitute a powerful scientific tool to support better understanding and management of the Earth and its resources. More specifically, large international initiatives such as ESA-EU GMES (Global Monitoring for Environment and Security) and the intergovernmental GEO (Group on Earth Observations) focus on coordinating international efforts in environmental monitoring, that is, to provide political and technical solutions to global issues, such as climate change, global environment monitoring, management of natural resources and humanitarian response.

At present several thousand ESA users worldwide (Earth scientists, researchers, environmentalists, climatologists, etc.) have online access to EO missions' metadata (millions of references), data (in the range of 3 to 5 PB) and derived information for long-term science and long-term environmental monitoring; moreover the requirements for accessing historical archives have greatly increased over the last years and the trend is likely to keep increasing. Therefore, the prospect of losing the digital records of science (and in particular the unique data, information and publications managed by ESA) is very alarming. Issues for the near future include:

- the identification of the type and amount of data to be preserved;
- the location of archives and their replication for security reasons;
- the detailed technical choices (e.g. formats, media);
- the availability of adequate funds.

Of course decisions should be taken in coordination with other data owners together with the support and advice of the user community.

ESA Overall Strategies

Currently the major constraints are that the data volumes are increasing dramatically (ESA plans of new missions indicate 5-10 times more data to be archived in the next 10-15 years). The available financial budgets are inadequate (preservation and access to data of each ESA mission are covered only until 10 years after the end of the mission). Futhermore, data preservation/access policies are different for each EO mission and each operator or Agency.

To respond to the urgent need for a coordinated and coherent approach for the long-term preservation of the existing European EO space data, ESA started consultations with its member States in 2006 in order to develop the European LTDP (Long Term Data Preservation) strategy, presented at DOSTAG (Data, Operations, Scientific and Technical Advisory Group) in 2007, and also formed a LTDP Working

[2] European Space Agency (ESA) Portal http://www.esa.int/

Group (January 2008) within the GSCB (Ground Segment Coordination Body)[3] to define European LTDP Common Guidelines (in cooperation with the European EO data stakeholders) and to promote them in CEOS (Committee on Earth Observation Satellites) and GEO (Group on Earth Observations). This group is defining an overall strategy for the long-term preservation of all European EO data, ensuring accessibility and usability for an unlimited time-span, through a cooperative and harmonized collective approach among the EO data owners (European LTDP Framework) by the application of European LTDP Common Guidelines.

ESA member states, as part of ESA's mandatory activities, have currently approved a three-year initial LTDP programme with the aim of establishing a full long-term data preservation concept and programme by 2011. ESA is now starting the application of the European LTDP Common Guidelines to its own missions. High-priority ESA LTDP activities for the next 3 years are focused on issues such as security improvement, migration to new technologies, increase in the number of datasets to be preserved and enhancement of data access. In addition ESA-ESRIN is participating in a number of international projects partially funded by the European Commission and concerned with technology development and integration in the areas of long-term data preservation and distributed data processing and archiving. The scope of ESA participation to such LTDP related projects is:

- to evaluate new technical solutions and procedures to maintain leadership in using emerging services in EO;
- to share knowledge with other entities, also outside the scientific domain;
- to extend the results and outputs of these cooperative projects in other EO (and ESA) communities.

The CASPAR Project

CASPAR (Cultural, Artistic, and Scientific knowledge for Preservation, Access, and Retrieval)[4] is an Integrated Project co-financed by the European Union within the Sixth Framework Programme (Priority IST-2005-2.5.10, Access to and preservation of cultural and scientific resources) that started on 1 April 2006.

CASPAR intends to provide tools and techniques for secure, reliable and cost-effective preservation of digitally encoded information for the indefinite future defining the methodology and infrastructure to deal with the impacts of changing technologies, including support for new media and data formats with evolving user communities. The CASPAR mission is to specify and build components for a framework which will apply to all types of digitally encoded information; to test this framework, CASPAR shows that it can preserve a heterogeneous spectrum of data that is subdivided into three broad interdisciplinary user communities: Cultural, Contemporary Performing Arts and Scientific Data testbeds.

The CASPAR framework is based on the OAIS Reference Model (Open Archival Information System, ISO:14721:2002) and, handling the preservation of digital resources of diverse user communities, enhances state-of-the-art technology in digital preservation and develops the technological solutions required.

[3] Ground Segment Coordination Body (GSCB) Home Page http://earth.esa.int/gscb/

[4] CASPAR Preservation User Community http://www.casparpreserves.eu/

Moreover, CASPAR is an open system able to interoperate with as many different systems as possible. It is to be operated in the framework of existing preservation solutions and be re-implemented as systems evolve.

According to the OAIS Standard, which defines a Functional Model for a digital archive identifying six macro functional components (Ingest, Archival Storage, Data Management, Administration, Access and Preservation Planning), the CASPAR Architecture Team has defined the "CASPAR Overall Component Architecture and Component Model", identifying 11 CASPAR Key Components: Registry (REG), Knowledge Manager (KM), Preservation Orchestration Manager (POM), Representation Information (REPINF), Preservation Datastore (PDS), Data Access and Security (DAMS), Digital Rights (DRM), Finding Aids (FIND), Virtualisation (VIRT), Packaging (PACK) and Authenticity (AUTH). These CASPAR key components can be seen as part of the six OAIS macro functional components and working together fulfil all the OAIS responsibilities of an archive.

In particular five main functional blocks have been identified:

- Information Package Management (the main CASPAR key component responsible for these activities is PACK, supported by REPINF, REG, PDS, FIND and VIRT);
- Information Access (the main CASPAR key component responsible for these activities is FIND, supported by KM, PACK and PDS);
- Designated Community (DC) and Knowledge Management (the main CASPAR key component responsible for these activities is KM, supported by REG and POM);
- Communication Management (the main CASPAR key component responsible for these activities is POM, supported by KM, REG and AUTH);
- Security Management (the main key CASPAR component responsible for these activities is DAMS, supported by DRM and AUTH).

The following diagram (Figure 1) shows an overview of the CASPAR Preservation Workflow.

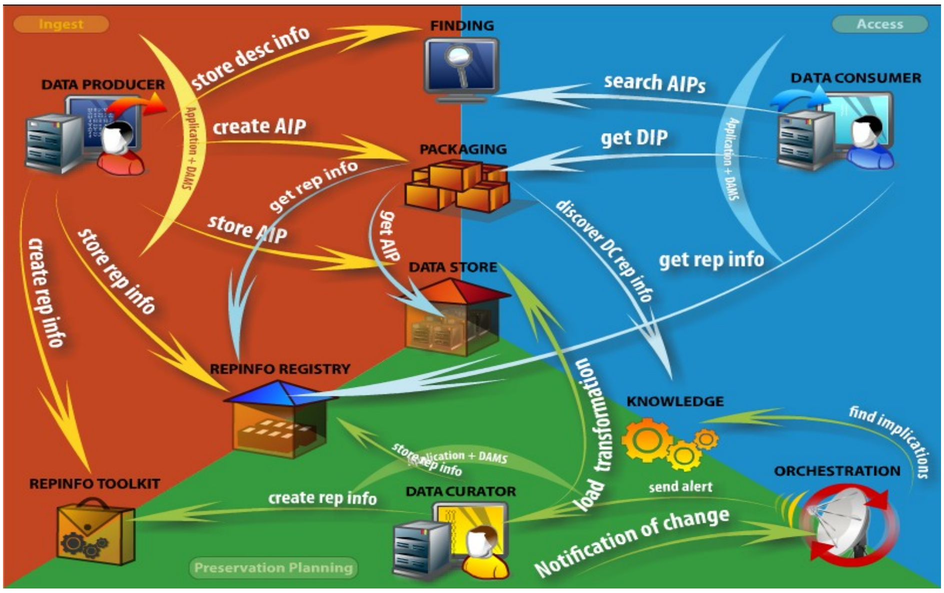


Figure 1. The CASPAR preservation workflow diagram.

The ESA Role in CASPAR

In CASPAR, ESA plays the role of both user and infrastructure provider for the scientific data testbed. ESA participation in CASPAR (consistent with the guidelines defined by the LTDP Working Group) is mainly driven by the interest in:

- consolidating and extending the validity of the OAIS reference model, already adopted in several internal initiatives (e.g. SAFE, Standard Archive Format for Europe, an archiving format developed by ESA in the framework of its Earth Observation ground segment activities);
- developing preservation techniques and tools covering not only the data but also the knowledge associated with them.

In fact locating and accessing historical data is a difficult process and their interpretation can be even more complicated given the fact that scientists may not have (or may not have access to) the right knowledge to interpret these data. Storing such information together with the data and ensuring all remain accessible over time would allow not only for a better interpretation but would also support the process of data discovery, now and in the future.

The ESA Scientific Testbed in CASPAR

The ESA Dataset for the Scientific Testbed

The selected ESA scientific dataset consists of data from GOME (Global Ozone Monitoring Experiment), a sensor on board the ESA ERS-2 (European Remote Sensing) satellite, which has been in operation for more than a decade. In particular, the GOME dataset:

- has a large total amount of distributed information with a high level of complexity;
- is unique because it provides more than 14 years global coverage;
- is very important for the scientific community and the Principal Investigators (PIs) that on a routine basis receive GOME data (e.g. KNMI and DLR) for their research projects (e.g. concerning ozone depletion or climate change).

Note that GOME is just a demonstration case because similar issues are applicable to many other Earth Observation instrument datasets.

The GOME dataset includes different data products, processing levels and a number of types of associated pieces of information which must be available as level processors (needed to process data from one level – see below - to another), format converters, auxiliary data (the ancillary information needed to process data), documents, methods, data viewers, examples of GOME science applications, and so on.

The commonly used names and descriptions of data products are as follows:

- Level 0 - raw data as acquired from the satellite, which is processed to:
- Level 1 - providing measures of radiances/reflectances. Further processing of this gives:
- Level 2 - providing geophysical data as trace gas amounts. They can be combined as:
- Level 3 - consisting of a mosaic composed of several level 2 data products

with interpolation of data values to fill the satellite gaps.

Figure 2 illustrates the processing chain to derive GOME Level 3 data from Level 0.

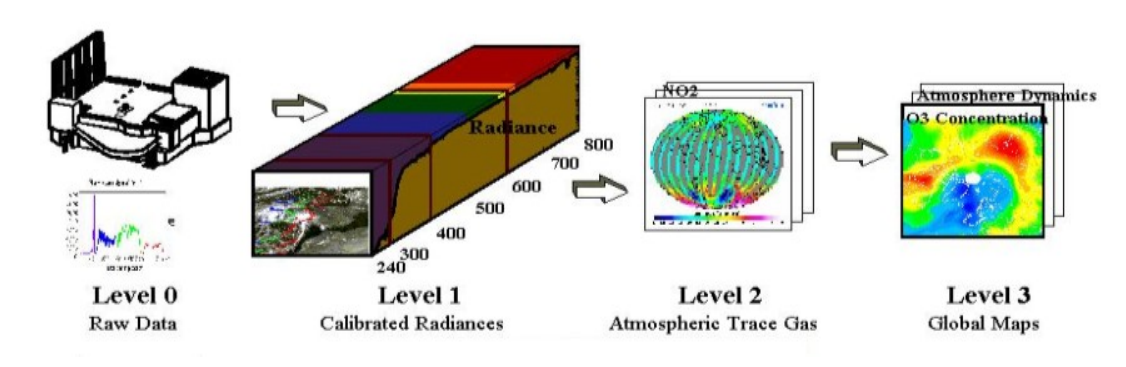


Figure 2. The steps of GOME data processing.

A particular processing scheme allows one to generate GOME Level 1C data (fully calibrated data) starting from Level 1 data (raw signals plus calibration data, also called L1B or L1b data); we have to point out that a single Level 1 data can generate (applying different calibration parameters) several Level 1C products and so a user asking for GOME Level 1C data will be usually sent the L1 data and the processor needed to generate Level 1C data. Figure 3 illustrates in more detail the processing chains to derive GOME Level 2 data from Level 0 and GOME Level 1C data from Level 1B.

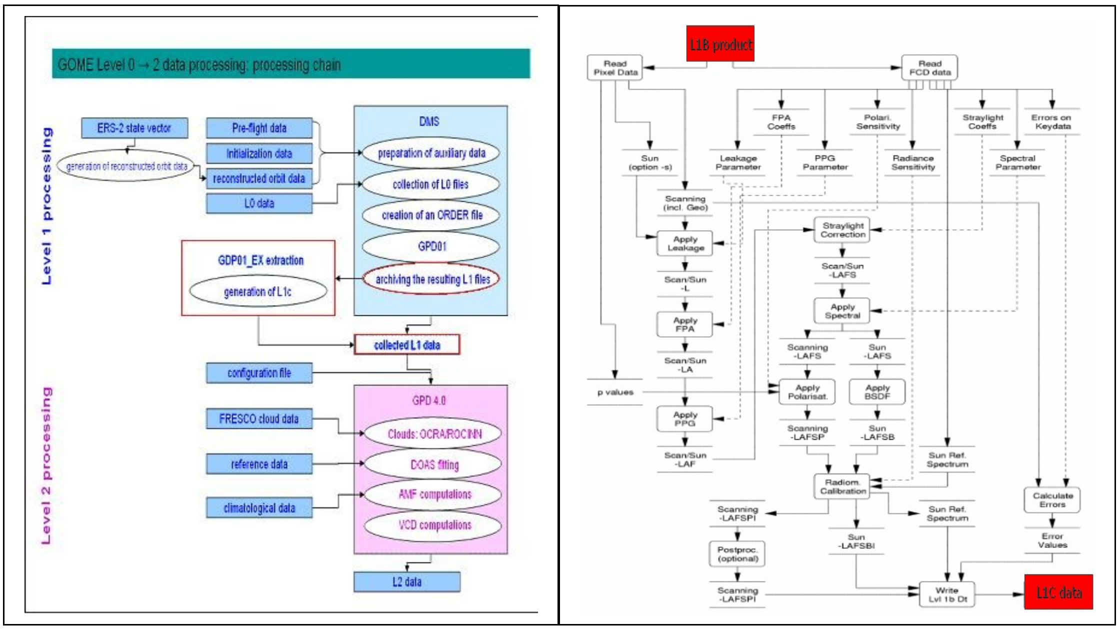


Figure 3. The GOME L0->L2 and L1B->L1C processing chains.

The ESA Testbed Scenario

The core of the CASPAR-dedicated ESA testbed is the preservation of the ability to process data from one level to another, that is the preservation of GOME data and of all components that support the operational processing which can generate products at higher levels. As a first demonstration case, it has been decided to preserve the ability to produce GOME Level 1C data starting from Level 1 data; at this time the ESA testbed is able to demonstrate the preservation of this part of the GOME processing chain at least against changes of operating system or compilers/libraries/drivers affecting the ability to run the GOME Data Processor.

The Preservation Scenario is as follows: after the ingestion in the CASPAR system of a complete and OAIS-compliant GOME L1 processing dataset, something (e.g., OS or gLib version) changes and a new L1B->L1C processor has to be developed/ingested to preserve the ability to process data from L1B to L1C. So we must cope with changes related to the processing by ensuring the correct information is passed through the system, for the benefit of system administrators and the users, by emphasizing a framework developed with only the CASPAR components.

The ESA testbed is divided into four logical phases:

1. CASPAR System Setup (configuration, knowledge modules and profiles creation);
2. Data and related Representation Information Ingestion;
3. Data Access, Browsing, Search and Retrieval;
4. Software Processing Preservation (Upgrade).

The last phase (Software Processing Preservation) is the testbed focus and needs a specific validation methodology while the other phases are validated by performing and then analyzing (evaluating) the correct implementation of CASPAR functionality.

CASPAR System Setup

ESA and ACS (Advanced Computer Systems SpA, technical partner for the testbed implementation)[5] have developed in cooperation with the Foundation for Research and Technology Hellas (FORTH), a basic EO ontology (based on a specialisation of the ISO 21127:2006 CIDOC-CRM, a formal ontology that fulfils the requirements of being the conceptual backbone for representing descriptive metadata) representing relationships and dependencies of the GOME Representation Information stored on the Knowledge Manager module used for the Testbed. The ontology is divided into two logic modules which are connected through the “L1B L1C processing” event:

- The first module links the processing event to the management of EO products and it is used to retrieve the Designated Community profile with the adequate knowledge about the searched data;
- The second module links the processing to those elements (e.g., compiler, OS, programming language) that are needed to develop and run a processor. Software-related ontologies are used by the System Administrator when the Upgrade is needed.

The first part of the schema is shown in Figure 4:

[5] Advanced Computer Systems A.C.S. S.p.A. (ACS) http://www.acsys.it/

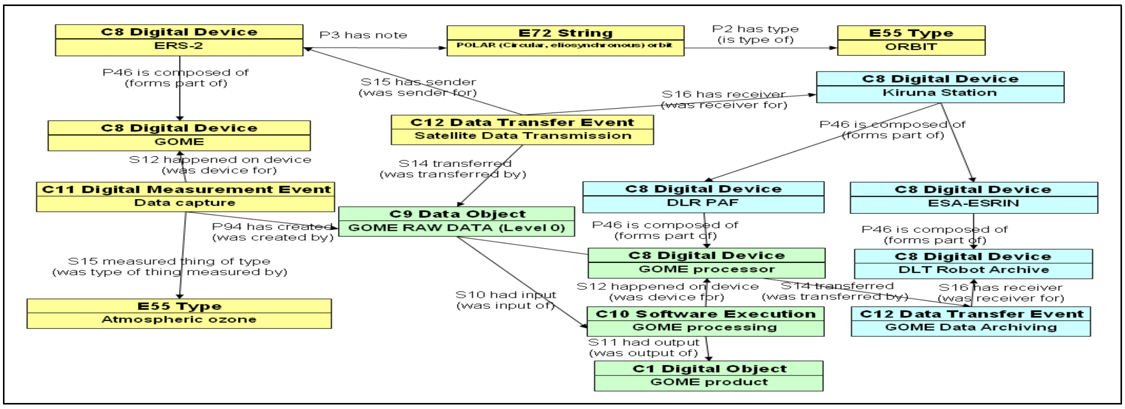


Figure 4. EO-based ontology.

The colours used in this ontology summarize different knowledge profiles: Earth Observation Expertise (Yellow), EO archives Expertise (Blue) and GOME Expertise (green). On this basis the testbed foresees four different DC profiles which are linked to knowledge profiles:

1. **GOME User** - user with no particular expertise about Earth Observation, GOME and related EO archives;
2. **GOME Expert** - expert in Earth Observation and GOME data but not in archiving techniques;
3. **Archive Expert** – expert in archiving techniques but not in EO;
4. **System Administrator** - the archive curator having knowledge of all modules. The System Administrator is the only DC Profile admitted to use the second ontology (see Figure 5):

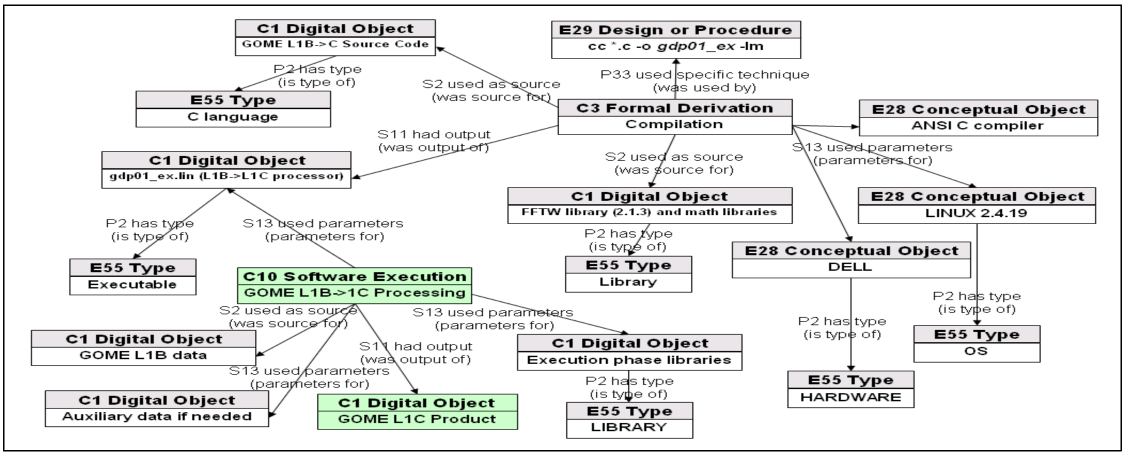


Figure 5. Software-based ontology.

Data Ingestion

The ingestion process allows the Data Producer to ingest into the CASPAR system GOME Level 1B data, the L1B L1C Processor and the Representation Information including all knowledge related to the GOME and processor data.

While GOME data and Processor are stored (through PACK) on the PDS component and searched or retrieved by FIND, all the related RepInfo are stored on the Knowledge Manager and browsed through the Registry. The scenario is represented in Figure 6.

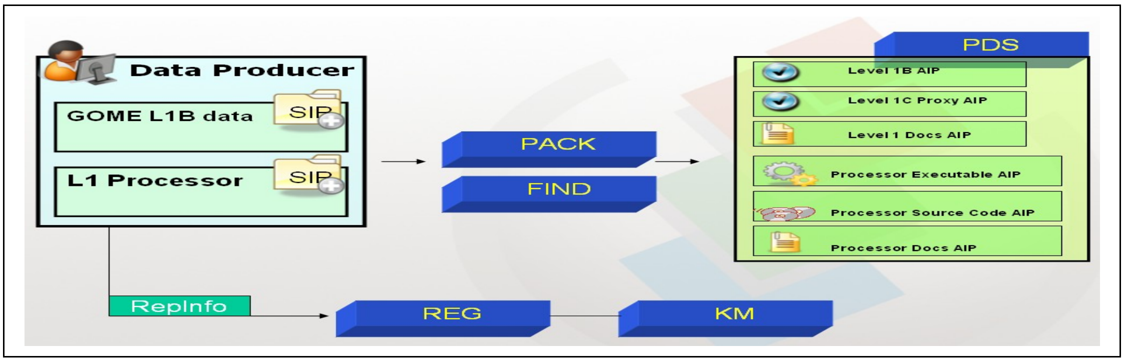


Figure 6. Data Ingestion phase.

To go into more detail, the dataset used in the ESA Scientific testbed is composed of the items listed in Table 1.

Dataset Item to be Preserved	Associated Representation Information (pdf files)
GOME L1B products (*.lv1b)	Technical Info (ERS Products Specification and L1B Product Specification), EO knowledge (ERS2 Satellite and GOME Sensor Specification), Legal data (Disclaimer, License)
L1B→L1C processor	Readme, User Manual
L1B→L1C processor source code	C Language and Linux OS Specifications

Table 1. The GOME L1B->L1C processing dataset to be preserved.

Data Access, Browsing, Search and Retrieval

CASPAR is able to return to any user, who asks for L1C data, not only the related L1 data plus the processor needed to generate them, but also all the information needed to perform this process, depending on the user's needs and knowledge. According to the DC Profiles knowledge (see Figure 4), different knowledge means different RepInfo modules retrieved during the search and retrieve session. The scenario is summarized in the following diagram (Figure 7).

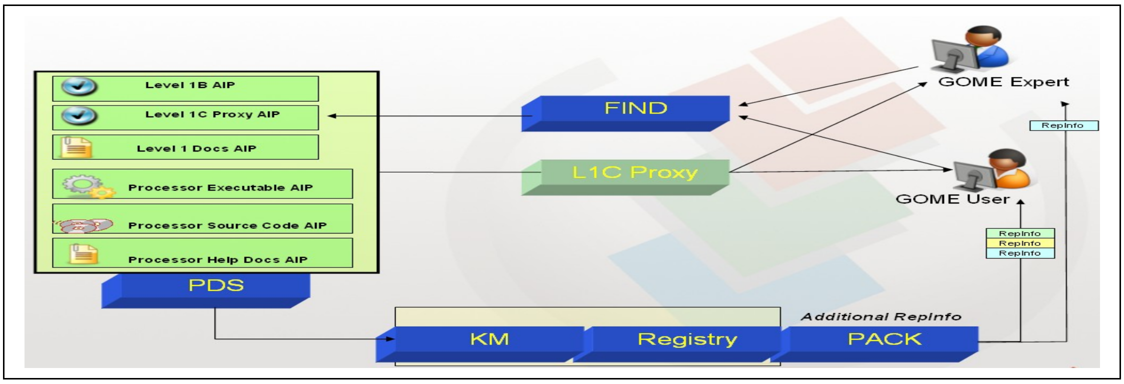


Figure 7. Different RepInfo is returned to different users.

To go into more detail, different Representation Information is returned to different users according to their Knowledge Base. After the creation of different profiles (i.e., different Knowledge Bases) for different users and the ingestion of appropriate Knowledge Modules (i.e., the competences that one should have to be able to understand the meaning of data) related to data, the KM component is able to understand that a GOME expert requires nothing to use the data while, a less expert generic GOME user (who is performing the same query) has to be returned the Representation Information needed for that user to understand the meaning of the data (see Figure 7).

Software Processing Preservation

The preservation phase (see Figure 8) can be summarized as follows:

- An external event affects the processor (e.g., a library or the Operating System changes) and an alert sent (through POM) by informed users is forwarded to the System Administrator;
- The System Administrator uses the Software Ontology to see which are the modules that need to be recompiled and updated (CIDOC-CRM defines the relationships between the modules);
- The System Administrator is able to retrieve, download and work on the source code of the processor to deliver a new version of the processor;
- The new version, with appropriate (updated) PDI and RepInfo, is re-ingested into the PDS;
- An alert mechanism notifies the users a new version is available;
- The new processor can be directly used to generate Level 1C products.

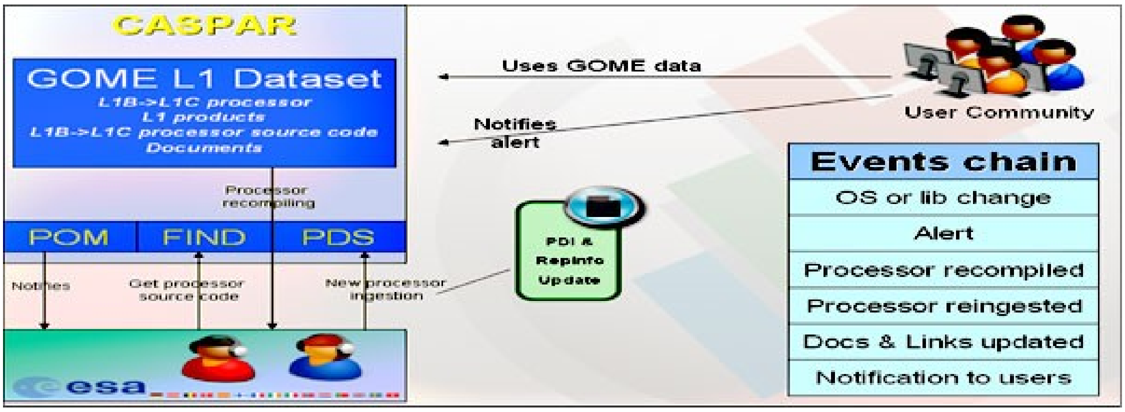


Figure 8. Preservation scenario.

Preservation procedure validation activities have been carried out in order to demonstrate two main scenarios:

- Library change: an object, external to the system, needed by the processor is out of date because of the release of a new version;
- OS Change: the processor needs to be run on a new Operating System.

In both cases the scenario purpose is to preserve the ability to process Level 1B products to generate Level 1C. In case of a change the following functionalities have to be carried out:

- allow the CASPAR user to notify an alert related to the processor;
- support the System Administrator to create and upload a new processor version with appropriate RepInfo, PDI and links to the previous versions;
- notify all users about the change.

Testbed Update Procedure Case 1: New Library Release.

The processor source code is in the C language and it can be compiled by an ANSI C compiler (it needs a FFTW library for Fast Fourier Transformation). In this scenario, a new library has been released by a simple redefinition of the fftw_one signature method; this does not affect the core business logic of the FFTW transformation, but does not allow the processor to be recompiled and run. The validation process tested that the correct alert was sent and the correct browsing, searching and retrieval of all those elements needed to rebuild the processor with the new library. By using the knowledge associated to the L1B->L1C processor, the System Administrator was able to access and download the processor source code, the GCC compiler and all related how-to's. Once all the necessary material and knowledge were downloaded, the new processor was recompiled, re-ingested and all associated RepInfo was updated to take into account the new version. The validation procedure has also demonstrated the correct process preservation by generating a new L1C product from an ESA-certified L1B and comparing the processing result with the corresponding ESA-validated L1C (obtained from running the original process using the same calibration parameters); L1C products were exactly equivalent.

Testbed Update Procedure Case : New Operating System.

To simulate a change in environment, it has been supposed that the LINUX operating system is becoming obsolete and so there is a need to migrate, for the purposes of this scenario, to the more popular SUN SOLARIS. After the notification of the need to switch to a SUN SOLARIS operating system, CASPAR has to allow the L1C creation on the new platform; the L1B->L1C processor creation and ingestion for SUN SOLARIS 5.7 was performed by using the same steps used in the previous example. The Representation Information knowledge tree has included an emulation system based on two open source OS emulators (VMWare and QEMU); for each emulator both the executable and the source code were available and browsable by means of CASPAR. Moreover, a second workstation (SOLARIS T1000 Sparc) was provided to perform the validation process on a real SOLARIS environment. The validation objective was to support the System Administrator in performing the processor update for all the above mentioned environments. By using the appropriate Representation Information, the System Administrator was able to compile the processor from Linux 5.3 to Linux Ubuntu (no major changes), from Linux 5.3 to Open Solaris (change in OS and libraries) and from Linux 5.3 to Solaris 5.10 (change in OS kernel and environment, libraries and CPU architecture) as shown in Figure 9.

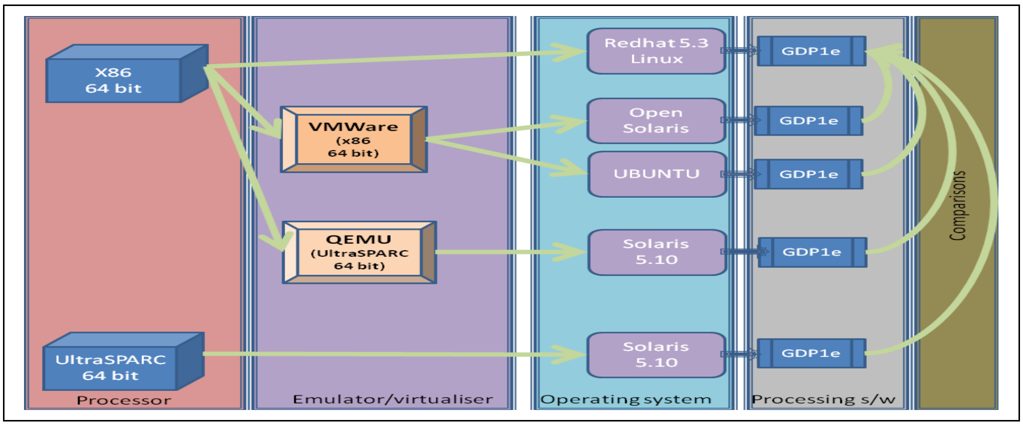


Figure 9. Combinations of hardware, emulator and software.

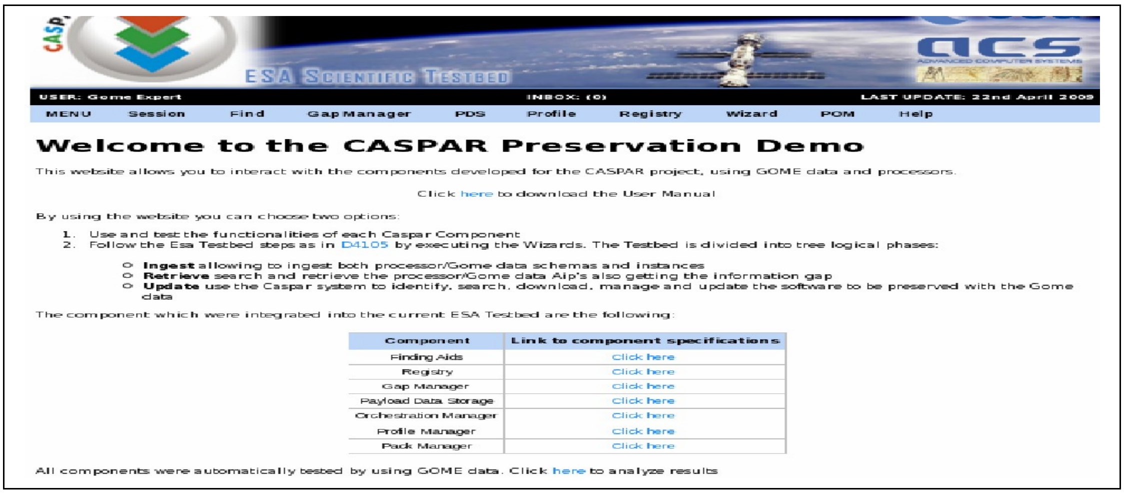
In both the update tests above, performed on the GOME data ingested into CASPAR and representative of more complex scenarios (e.g. changes in compilers, hardware, etc.), by browsing the RepInfo, the System Administrator was able to access to the source code, the compilers, the software environment, the emulators and all the related instructions to perform the critical steps needed to maintain the ability to process data. This has improved the ability of the System Administrator to guarantee the processing capability in more critical conditions.

The complete events chain for the scenario of the ESA scientific testbed is described in the following table:

Action	Main CASPAR components involved	Notes
L1 data and L1->L1C processor are ingested in the PDS of the CASPAR system	PACK REPINF KM REG PDS FIND	Data and processor are OAIS-compliant (SAFE-like format), with appropriate Representation Information and Descriptive Information
Data and appropriate Representation Information are returned to users according to their Knowledge Base	FIND DAMS KM REG	It is also possible to ingest as AIP an appropriate L1 to L1C Transformation Module into the PDS and access directly L1C data (with fixed user-decided calibration parameters) using a processor previously installed on the user machine
The OS or gLib version changes, and an alert is sent by informed users to appropriate people	POM	People interested in changes are POM-dedicated topic subscribers
The system administrator retrieves and accesses the source code of the processor	FIND DAMS PDS REG	The system administrator is one of the POM-dedicated topic subscribers and has the responsibility of taking appropriate corrective actions
The system administrator recompiles/upgrades the processor executable and reingests it into the CASPAR system	PACK KM PDS REG	An appropriate Administrator Panel showing the semantic dependencies between data will help the system administrator to identify what Representation (and Descriptive) Information has also to be updated
By a notification system all the interested user communities are correctly notified of this change	POM	Users are still able to process data from L1B to L1C: scientific capabilities remain intact.

Table 2. Events chain.

The scenario above has been implemented in ESA-ESRIN by ESA and ACS through a web-based interface which allows users to perform and visualize the scenario step by step using rich graphical components (see Figure 10).



Welcome to the CASPAR Preservation Demo

This website allows you to interact with the components developed for the CASPAR project, using GOME data and processors.

Click here to download the User Manual

By using the website you can choose two options:

1. Use and test the functionalities of each Caspar Component
2. Follow the Esa Testbed steps as in D4105 by executing the Wizards. The Testbed is divided into tree logical phases:
    - **Ingest** allowing to ingest both processor/Gome data schemas and instances
    - **Retrieve** search and retrieve the processor/Gome data Aip's also getting the information gap
    - **Update** use the Caspar system to identify, search, download, manage and update the software to be preserved with the Gome data

The component which were integrated into the current ESA Testbed are the following:

Component	Link to component specifications
Finding Aids	Click here
Registry	Click here
Gap Manager	Click here
Payload Data Storage	Click here
Orchestration Manager	Click here
Profile Manager	Click here
Pack Manager	Click here

All components were automatically tested by using GOME data. Click here to analyze results

Figure 10. CASPAR demo interface.

The application is now available and open to everyone for exploitation and further work.

Conclusions

The current ESA strategy for long-term EO data preservation is based on the assumption that there is a fundamental requirement to guarantee access to, and use of, long time series of EO data for long-term scientific research and environmental monitoring by the scientific and operational user communities for as long as possible. This contribution has provided an overview on some ESA-ESRIN initiatives carried out in collaboration with European data owners and providers, entities and institutions, with the objective of guaranteeing long-term data preservation; ESA participation in such initiatives will represent a major step towards a coordinated and coherent approach for a harmonized European archives management policy.

In particular the paper focused on the ESA participation and contribution to the CASPAR Project describing in detail the scientific testbed implemented in ESA-ESRIN, that provides convincing evidence of the effectiveness of the CASPAR preservation framework in the Earth Observation domain.

The overall impact of the CASPAR system and its future potential are quite clear to both the people that developed it and those who used it; the need to preserve and link EO tools and data has become more and more evident in past years. The ESA-ACS team is confident that the CASPAR solutions are going to be increasingly adopted in the years to come.

Acknowledgements

Many thanks to Luigi Fusco (ESA-ESRIN) and Fulvio Marelli (ACS) for their hard work on the topics covered by this paper.