Long Term Data and Knowledge preservation to guarantee access and use of the Earth Science archive

Sergio Albani [(1)], David Giaretta [(2)]

[(1)] ACS c/o ESA-ESRIN
Via Galileo Galilei,00044 Frascati, Rome, Italy
EMail: Sergio.Albani@esa.int

[(2)] STFC Rutherford Appleton Lab
Didcot, Oxon, OX11 0QX, UK
EMail: david.giaretta@stfc.ac.uk

ABSTRACT

Earth Observation (EO) Space Missions provide global coverage of the Earth by generating on a continuous basis huge amounts of data from a variety of sensors. Locating and accessing these 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. Preserving 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 EU-funded CASPAR (Cultural, Artistic, and Scientific knowledge for Preservation, Access and Retrieval) project is just building a framework to support the end-to-end preservation lifecycle for digital information, based on the OAIS reference model, with a strong direction on the preservation of the knowledge associated to data. In the context of the current ESA overall strategies carried out in collaboration with European EO data owners/providers, entities and institutions, this paper will focus on the ESA participation and contribution to the CASPAR project with the main objective of guaranteeing long term preservation not only of EO data but also of the right knowledge needed to scientists to process and use them.

Long Term Data Knowledge Preservation, Earth Science, CASPAR, ESA

INTRODUCTION

Earth Observation (EO) data provide global coverage of the Earth across both a continuum of timescales (from historical measurement to real time assessment to short and long term predictions) and a variety of geographical scales (from global scale to very small scale). EO data acquired from space constitute therefore a powerful scientific tool to enable better understanding and management of the Earth and its resources. 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 the long term science and the long term environmental monitoring; moreover the requirements for accessing historical archives have been strongly increased over the last years and the trend is likely to increase and increase. Therefore, the prospect of losing the digital records of science (and with the specific unique data, information and publications managed by ESA) is very alarming.
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 formed a Long Term Data Preservation (LTDP) Working Group (Jan 2008) within the GSCB (Ground Segment Coordination Body) to define 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.
Among these guidelines we should highlight at least the following ones:

- "Archived data shall contain all the elements necessary to be accessed, used, understood and processed to obtain mission products to be delivered to users";
- "Adoption of ISO 14721 - OAIS standard as the reference model and adoption of common archive data formats for AIPs (e.g. SAFE, Standard Archive Format for Europe)".

ESA member states, as part of ESA's mandatory activities, have currently approved a three year initial LTDP programme with the aim to establish 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.

In addition ESA-ESRIN is participating to 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, outlined in the next sections, 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 of the scientific domain;
- to extend the results/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) 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.

As digital information is becoming more ubiquitous and indispensable and at the same time extremely fragile, CASPAR intends to provide tools and techniques for secure, reliable and cost-effective preservation of digitally encoded information for the indefinite future. CASPAR is 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 and facilitate the sharing of the effort needed to do this. To achieve these challenges, CASPAR has assembled a consortium composed of international professionals and organisations, such as scientific, cultural and creative professionals and experts, commercial partners and information preservation leaders. CASPAR also collaborates with other relevant international digital preservation initiatives, such as Digital Preservation Europe, PLANETS and the UK Digital Curation Centre.

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 we will show that we 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), which is a conceptual framework for archival systems dedicated to preserving the understandability and usability of, and maintaining access to, digitally encoded information over the long term. The CASPAR framework, handling the preservation of digital resources of diverse user communities, will enhance state of the art technology in digital preservation and will develop the technological solutions required.

Another goal of CASPAR is to be user oriented. CASPAR is an open system able to interoperate with as many different systems as possible, to be operated in the framework of existing preservation solutions and be re-implemented as systems evolve. The active participation of the CASPAR Preservation User Community, which is a growing, worldwide aggregation of institutions and individuals interested in digital preservation at all levels, will facilitate a wide adoption of CASPAR and guarantee that the system can evolve with the requirements for which it has been designed.

Based on the OAIS Standard, which defines a Functional Model for a digital archive identifying 6 macro functional components (see Figure 1), 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).

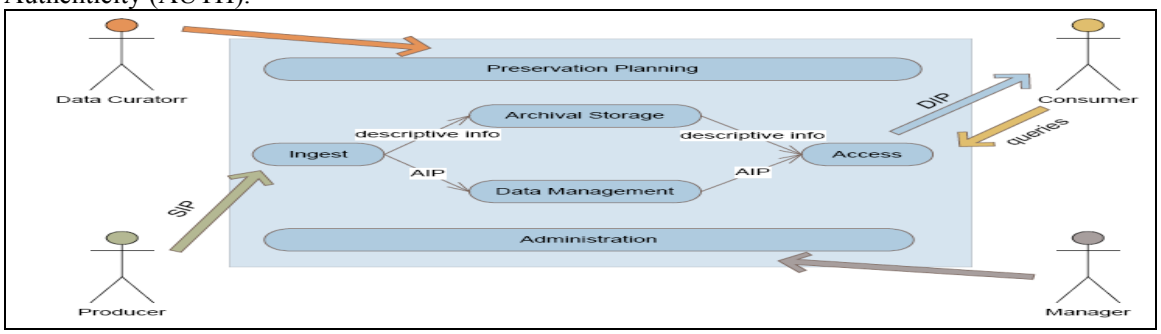

Figure 1: OAIS Functional Model

These CASPAR key components can be seen as part of the 6 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 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) and Security Management (the main key CASPAR component responsible for these activities is DAMS, supported by DRM and AUTH).
The following diagram (see Figure 2) shows an overview of the CASPAR Preservation Workflow.

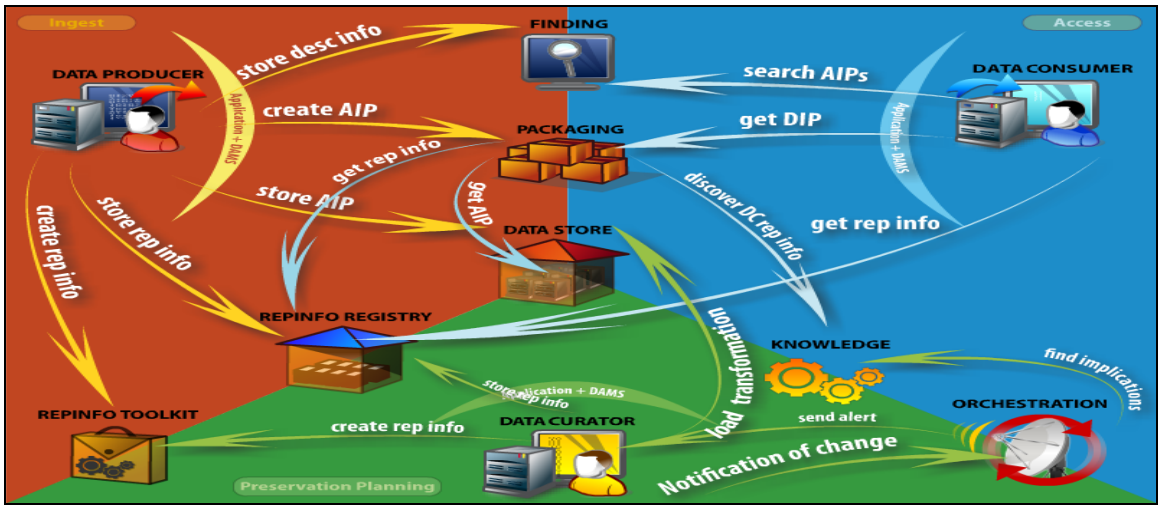

Figure 2: The CASPAR Preservation Workflow Diagram

We have to point out that CASPAR aims to put knowledge at the heart of preservation. This includes techniques to preserve the information and knowledge that is encoded in digital objects, not just "the bits", and in particular via Semantic Representation Information. In addition, we use similar techniques to:

- support defining Designated Communities, a fundamental requirement of OAIS;
- facilitate the development of new Representation Information and identify missing Representation Information needed as the Knowledge Bases of the Designated Communities change;
- assist in searching for relevant data objects;
- capture information such as provenance.

We have designed and developed a number of high-level knowledge management services for digital information preservation systems, all based on Semantic Web technologies.

THE ESA SCIENTIFIC TESTBED IN CASPAR

The ESA role in CASPAR

In CASPAR, ESA plays the role of both user and infrastructure provider for the scientific data testbed.
ESA participation to CASPAR (coherently with the above guidelines of 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, an archiving format developed by ESA in the framework of its Earth Observation ground segment activities);
- developing preservation techniques/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 selected dataset for the scientific testbed

The selected ESA scientific dataset consists of data from GOME (Global Ozone Monitoring Experiment), a sensor on board 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 information distributed with a high level of complexity; is unique because it provides more than 11 years global coverage; is very important for the scientific community and the Principal Investigators (PI) 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 involved in many other Earth Observation instrument datasets.
The GOME dataset includes different data products, processing levels and associated information. The commonly used names and descriptions of these types of data 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. These can be combined as:
- Level 3 - consisting of a mosaic composed by several level 2 data products with interpolation of data values to fill the satellite gaps.

In addition there are a number of types associated pieces of information which must be available: level processors (needed to process data from one level to another), auxiliary data (the ancillary info needed to process data), documents, methods, data viewers, format converters, examples of GOME science applications, etc..
Figure 3 illustrates the processing chain to derive GOME Level 3 data from Level 0.

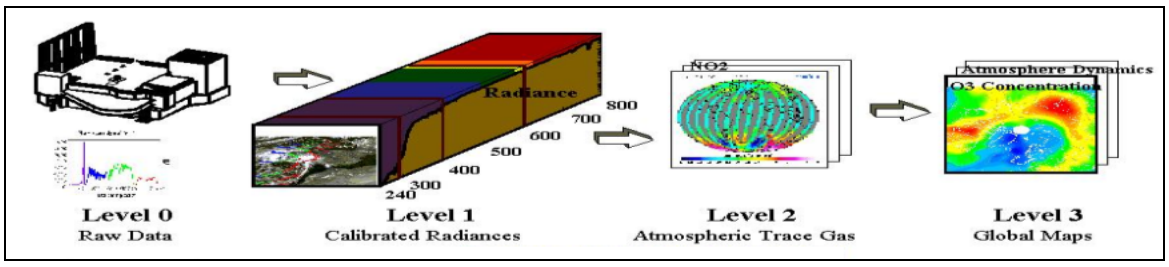

Figure 3: The steps of GOME data processing

Moreover a particular processing allows also to generate GOME Level 1C data (fully calibrated data) starting from Level 1 data (raw signals plus calibration data, also called 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 returned of L1 data and of the processor needed to generate Level 1C data.

The ESA testbed scenario

The core of the CASPAR dedicated 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 enables the operational processing for generating products at higher levels.
As first demonstration case, it has been decided to preserve the ability to produce GOME Level 1C data starting from Level 1 data; at this moment the ESA testbed is able to demonstrate the preservation of this 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 the following: after the ingestion in the CASPAR system of a complete (see Figure 4) and OAIS-compliant GOME L1 processing dataset, something (e.g. OS or gLib version) changes and a new L1->L1C processor has to be developed/ingested to preserve the ability to process data from L1 to L1C.
So we must cope with changes related to the processing by managing a correct information flow through the system, the system administrators and the users, using a framework developed using only the CASPAR components.

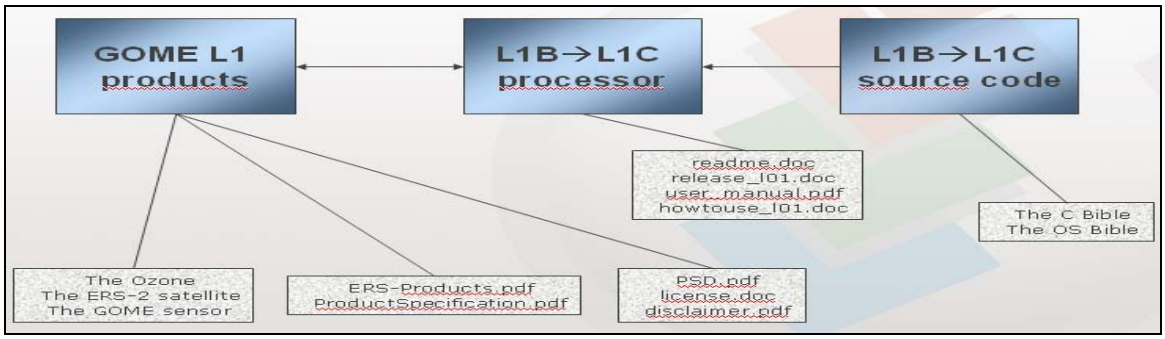

Figure 4: The GOME L1B->L1C processing dataset to be preserved

Moreover we want to be able to return to a user asking for L1C data not only the related L1 data plus the processor needed to generate them but also all the Representation Information needed to perform this process depending on the user needs and knowledge.
To reach this goal, we have developed an ISO 21127:2006 CIDOC-CRM based ontology that describes the L1→L1C processing event linking it to the EO products generation/archiving chain and to the technical elements related to the processing chain (e.g. compiler, OS and programming language); the Representation Information relationships and dependencies have been stored on the Knowledge Manager component. In Figure 5 you can see an example of how different Representation Information are returned to different Designated Communities according to their Knowledge Base. After the ingestion of 1) Knowledge Modules defining what is needed to understand the meaning of data (i.e. EO expertise, EO data archiving expertise or GOME data expertise) and 2) Designated Communities Knowledge Profiles (i.e. Generic User, GOME Expert, Archiving Expert or System Administrator), the Knowledge Manager component is able to understand that Scientist 1 does not need anything to use the L1 data while Scientist 2 (who is performing the same query) has to be returned with some documents in order to be able to understand the meaning of the data and to use them.

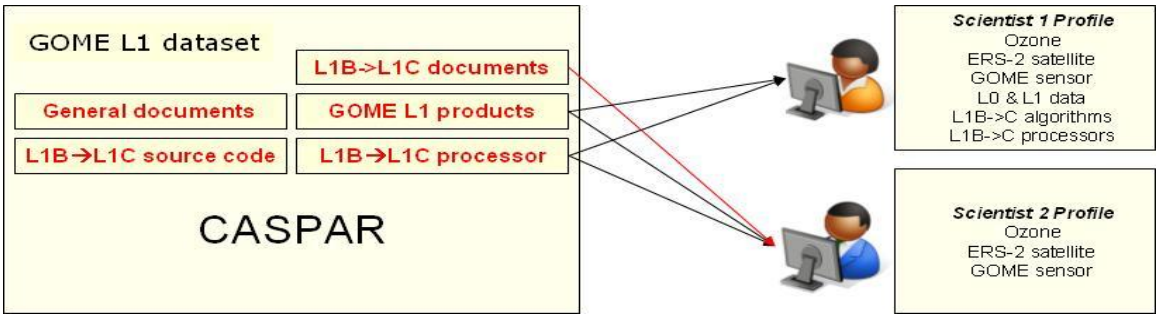

Figure 5: Different RepInfo are returned to different users

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 about changes are POM dedicated topic subscribers
The system administrator retrieve and access the source code of the processor	FIND DAMS PDS REG	The system administrator is one of the POM dedicated topic subscriber and has the responsibility to take appropriate corrective actions
The system administrator recompiles/upgrades the processor executable and reingest 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 have also to be updated
By a notification system all the interested users communities are correctly notified of this change	POM	People interested about changes are POM dedicated topic subscribers

The scenario above has been implemented in ESA-ESRIN by ESA and ACS (Advanced Computer Systems SpA, technical partner for the testbed implementation) through a web-based interface which allows users to perform and visualize the scenario step by step by rich graphical components.

CONCLUSION

The current ESA strategy for long term EO data preservation is based on the assumption that there is a fundamental requirement to guarantee to the scientific and operational user communities access and use for as long as possible to long time series of EO data for long term scientific research and environmental monitoring.

This contribution has provided an overview on some ESA-ESRIN initiatives carried out in collaboration with European data owners/providers, entities and institutions, with the objective to guarantee long term data preservation; in particular the paper focus on the ESA participation and contribution to the CASPAR project with a detailed description of the scientific testbed implemented in ESA-ESRIN that

provides a good proof for the effectiveness of the CASPAR preservation framework in the Earth Observation domain focusing on methodologies and techniques to guarantee the preservation of the knowledge associated to data.

REFERENCES

[website] http://www.esa.int

[website] http://earth.esa.int/gscb/

[website] http://www.casparpreserves.eu

ACKNOWLEDGEMENTS

Many thanks to Vincenzo Beruti and Luigi Fusco (ESA-ESRIN) for their hard work on the topics covered by this article.

AUTHORS' BIOGRAPHY

Sergio Albani

Sergio Albani is a Physicist (Astrophysics and Space Physics branch) with several years of experience in the space field and particularly in Earth Observation (EO). He was awarded a Master in "Journalism and Scientific & Institutional Communication" by the University of Ferrara (Italy) in 2005. After four years work in industry (Advanced Computer Systems SpA, Rome) as Software Engineer for the processing of EO satellite data, since 2006 he is employed as contractor at the European Space Agency (EO Applications Strategy Office, ESRIN establishment) to manage the CASPAR project and to contribute to other ESA initiatives in the area of Long Term Preservation of EO data.

David Giaretta

Dr Giaretta has had extensive experience in planning, developing and running scientific archives and providing and managing a variety of services to large numbers of users. In 2003 he was awarded an MBE for services to Space Science. As chair of CCSDS Panel 2 he made fundamental contributions to the OAIS Reference Model (ISO 14721) which is 'now adopted as the "de facto" standard for building digital archives' according to the NSF report: Cyberinfrastructure Vision for 21st Century Discovery, and continues to contribute to developing the follow on standards. He is now Associate Director for Development in the DCC and has played an active role in all aspects of the project; he also leads the CASPAR project which seeks to address fundamental issues of digital preservation. In addition he leads the PARSE.Insight EU project.