Long term preservation, discovery, access and exploitation of Earth Science data: the CASPAR and GENESI-DR combined approach

Sergio Albani [(1)], Roberto Cossu [(2)], Eliana Li Santi [(3)]

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

[(2)] ESA-ESRIN
Via Galileo Galilei,00044 Frascati, Rome, Italy
EMail: Roberto.Cossu@esa.int

[(3)] INTECS c/o ESA-ESRIN
Via Galileo Galilei,00044 Frascati, Rome, Italy
EMail: Eliana.Li.Santi@esa.int

ABSTRACT

Long Term Preservation of Earth Science (ES) data and of the capability to discover, access, process and use them is essential for scientists needing broad series of data for several types of investigations (e.g. Global Change); to better satisfy these requirements, ESA is (in complement to other internal initiatives) participating to several EU-funded projects.

Recently a Working Group led by ESA has been created to establish a framework of collaboration between the CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) and GENESI-DR (Ground European Network for Earth Science Interoperations - Digital Repositories) projects.

On the one hand CASPAR is building a framework to support the end-to-end preservation lifecycle for scientific, artistic and cultural information, based on existing and emerging standards; on the other hand, GENESI-DR has the challenge of establishing open and seamless access to data, information, products and knowledge originating from space, airborne and in-situ sensors from several dispersed Earth Science digital repositories.

From this cooperation it is expected that: 1) CASPAR will benefit from the GENESI-DR services to validate in a more complete form its data preservation framework in the Earth Science domain and 2) GENESI-DR Research Infrastructure will demonstrate its ability to adopt data preservation and curation mechanisms defined in CASPAR.

This paper will focus on these integration activities.

Long Term Data Preservation, Earth Science, GENESI-DR, CASPAR, ESA

INTRODUCTION

Planet Earth is increasingly in danger due to the strong presence and impact of the human life. In response to this, the United Nations have supported environmental conventions attempt to define internationally agreed protocols (e.g., Kyoto, Biodiversity and Montreal) to limit and monitor status of our global environment. The implementation and the systematic monitoring of international Environmental conventions need data, tools and world-wide infrastructures to gather and share the data. Data play indeed a central role in this context and are the basic input to any ES application, directly or indirectly as they validate the results. Two major aspects are of relevance in the context of this paper: 1) data discovery and access and 2) data curation.

Currently, data (which come from sensors on different platforms, such as satellites, planes, boats, balloons, or located at ground on the land), information about the state of the Earth, relevant services, analysis results, applications and tools are accessible in a very scattered and uncoordinated way. ES community would significantly benefit from a shared global ES infrastructure able to give simple access to historical data holdings and networks of sensors, broadband communications via ground and space, efficient, effective and distributed computing and storage resources, etc. In the context, ESA is leading the GENESI-DR project, which will develop an ES dedicated infrastructure providing reliable, easy, long-term access to Earth Science data and services via the Internet.

With regards to data curation, 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. 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, such as CASPAR.

It should be clear from the previous analysis that there is a strong need of easily discovering and accessing data as well as adopting proper and effective data curation and preservation strategies. In this paper we report the results of the collaboration between CASPAR and GENESI-DR projects, addressing data curation and data discovery, access, and processing respectively. This collaboration has the overall objectives of:

1) evaluating the effectiveness of the CASPAR data preservation framework as well as the impact on users and data providers while increasing its visibility in the Earth Science community by taking advantage of the GENESI-DR audience;
2) demonstrating the ability of GENESI-DR Research Infrastructure to provide easy access to heterogeneous and disperse data which are preserved according to the mechanisms defined in CASPAR.

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. 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 (Ingest, Archival Storage, Data Management, Access, Administration, 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 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, Information Access, Designated Community and Knowledge Management, Communication Management and Security Management.

ESA participation to CASPAR 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 CASPAR, ESA plays the role of both user and infrastructure provider for the scientific data 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.

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 GENESI-DR PROJECT

Ground European Network for Earth Science Interoperations - Digital Repositories (GENESI-DR), an ESA-led, European Commission (EC)-funded two-year project, kicked-off early 2008 and is taking the lead in providing reliable, easy, long-term access to Earth Science data via the Internet. As previously discusses, Petabytes of data about our planet are available but distributed at different locations. Currently, information about the state of the Earth, relevant services, analysis results, applications and tools are accessible in a very scattered and uncoordinated way, often through individual initiatives from Earth Observation mission operators, scientific institutes dealing with ground measurements, service companies, data catalogues, etc. Data access is a major logistic problem. The EC has funded GENESI-DR as a flagship project in Europe to help meet the challenge of facilitating life of scientists from different Earth Science disciplines located across Europe in discovery, access and use (combining, integrating, processing, …) of historical and fresh Earth-related data from space, airborne and in-situ sensors archived in large distributed repositories.

GENESI-DR is a response to the need for science users to be provided with data and tools to access, combine, and integrate the Earth-related data for performing their analyses.

This need has lead to the identification of the basic GENESI-DR infrastructure requirements:

- Capability, for Earth Science users, to discover data from different European Earth Science Digital Repositories through the same interface in a transparent and homogeneous way;
- Easiness and speed of access to large volumes of coherently maintained distributed data in an effective and timely way;
- Capability, for DR owners, to easily make available their data to a significantly increased audience with no need to duplicate them in a different storage system.

The first requirement is reflected in the GENESI-DR architecture on the Central Discovery Service, which allows users and applications to query information about data collections and products existing in heterogeneous catalogues, at federated DR sites.

This service can be accessed by users via web interface, the GENESI-DR Web Portal, or by external applications via open standardized interfaces exposed by the system.

More in detail, the Central Discovery Service identifies the DRs providing products complying with the user search criteria and returns the corresponding access points to the requester. This latter can refine its search towards the DRs, so that products complying with refined search criteria are identified and the corresponding metadata are returned to the client. These include the product access URL allowing product retrieval.

On this purpose, to meet the second requirement, flexibility and performance are taken into consideration by making use of different and efficient data transfer technologies such as HTTPS, GridFTP and BitTorrent. The cope with the third requirement, the Architecture of GENESI-DR provides the DR owners with a mechanism (Catalogue Generator) to produce a metadata catalogue by simply harvesting metadata from their storage systems.

The Central Discovery Service communicates with the different DR catalogues through a web service gateway.

GENESI-DR platform is in line with the platform visioned by DEGREE and with the expected achievements of the major ES communities programmes/initiatives, like GEOSS and INSPIRE. GENESI-DR has formal relation with GEO in several tasks such as: the management of large volumes and diverse types of Earth observation data; the implementation of the GEOSS architecture; the harmonisation of data, metadata and products; the use of satellites for risk management (inherited from the GPOD Fast Access to Imagery for Rapid Exploitation service).

GENESI-DR is also analysing common approaches to preserve the historical archives and the ability to access the derived user information as both software and hardware transformations occur. Ensuring access to Earth Science data for future generations is of utmost importance because it allows for the continuity of knowledge generation improvement.

THE CASPAR AND GENESI-DR COMBINED APPROACH

Recently a Working Group led by ESA has been created to establish a framework of collaboration between CASPAR and GENESI-DR in order 1) to adopt data preservation and curation mechanisms defined in CASPAR within GENESI-DR Research Infrastructure and 2) to provide CASPAR with a further way for validating the data preservation and curation mechanism in the domain of Earth Science, benefiting from GENESI-DR feedback and integrating services shared with other DRs to enlarge its capabilities to meet the ES community requirements.

To this end, different validation scenarios have been identified. The first step of this collaboration has been to define in detail the integration scenario between GENESI-DR and CASPAR. Two subsequent integration phases have been identified.

In the first phase the following actions have been performed:

- GENESI-fication of a CASPAR based DR. A DR, in order to make its data discoverable and accessible through GENESI-DR, needs to perform the so called GENESI-fication procedure (described in detail in the GENESI-fication Guide). This action allows users to discover and access data that are preserved according the mechanisms defined in CASPAR.
- Development of a validation service that estimates vertical profiles of Ozone starting from GOME L1 data stored in a CASPAR based DR. In this scenario, the user can discover and select, via GENESI-DR WebPortal, GOME L1 data (retrieved from the CASPAR based DR and preserved according to the mechanisms defined by CASPAR) and demand their processing, using GENESI-DR processors and computing resources, to obtain an estimate of vertical profiles of ozone.
- Development of a validation service that generates GOME L1C data starting from GOME L1 data.
- In this scenario, the user can discover and select, via GENESI-DR WebPortal, GOME L1C data. These data are generate on-the-fly using related GOME L1 and processors stored (and so preserved) in a CASPAR based DR.

These scenarios demonstrate:

- how CASPAR can benefit from GENESI-DR in using shared services for discovering, accessing and processing preserved data;

- how GENESI-DR can benefit from CASPAR in providing the user with preserved data and processors;
- the possibility for CASPAR users to discover and access data generated on-the-fly and in a transparent way by using services provided by GENESI-DR and source data and processors specified by the curation mechanism.

In Figure 1, illustrating one of the validation scenarios previously described, the user obtains ozone profile information using the GENESI-DR processing software and the GOME L1 data preserved in CASPAR.

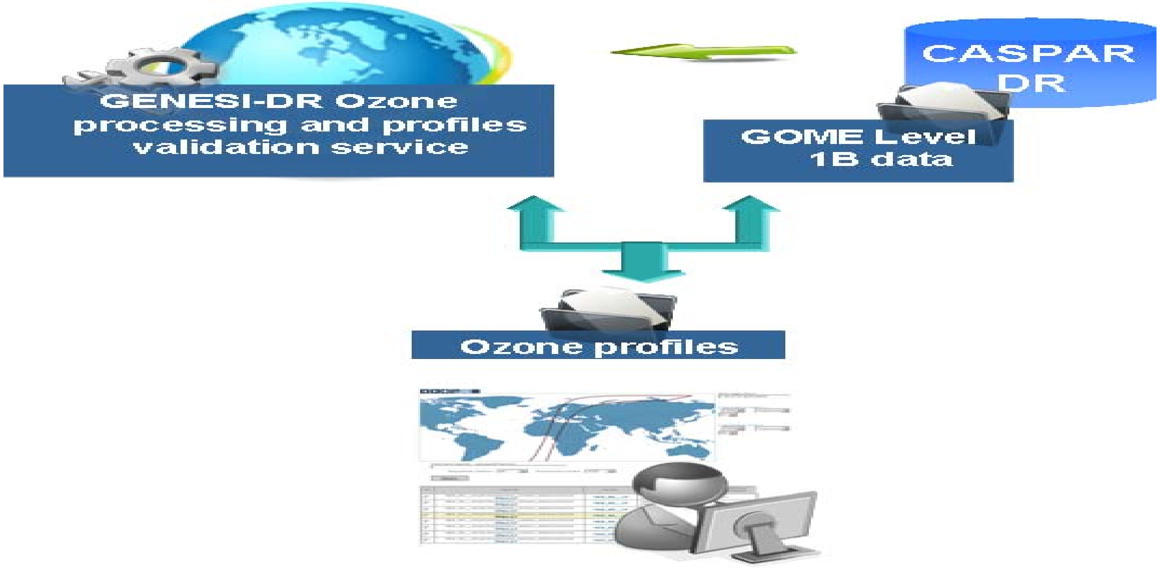

Figure 1 – Access and exploitation of preserved data

In Figure 2, illustrating another validation scenarios previously described, the user derives GOME L1C products from GOME L1 using processing software and data both preserved in CASPAR.

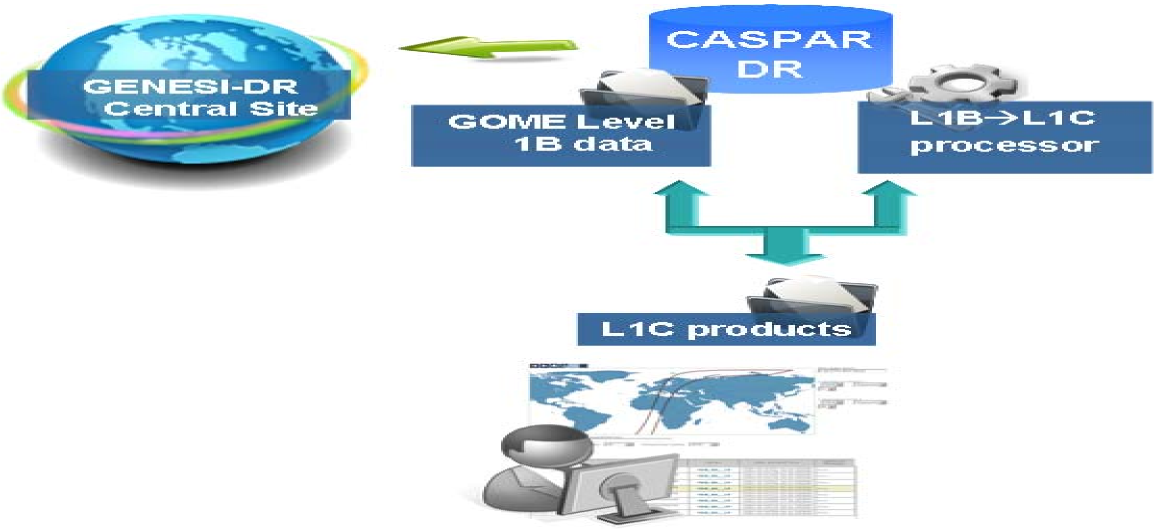

Figure 2 Access and exploitation of preserved data and preserved processors

In a second phase of the cooperation, the actions needed for the following tasks will be analysed and performed:

- to make the previously defined services accessible from outside GENESI-DR webportal (e.g. webservices) ;
- to store the processing results in CASPAR-DR;
- to return profile-based RepInfo to GENESI users (see Figure 3: a less expert user asking for L1 products gets the products, processing software and related documentation, while an expert user performing the same query only gets products and related processing software);
- to define a strategy for propagating CASPAR features to other relevant GENESI DRs.

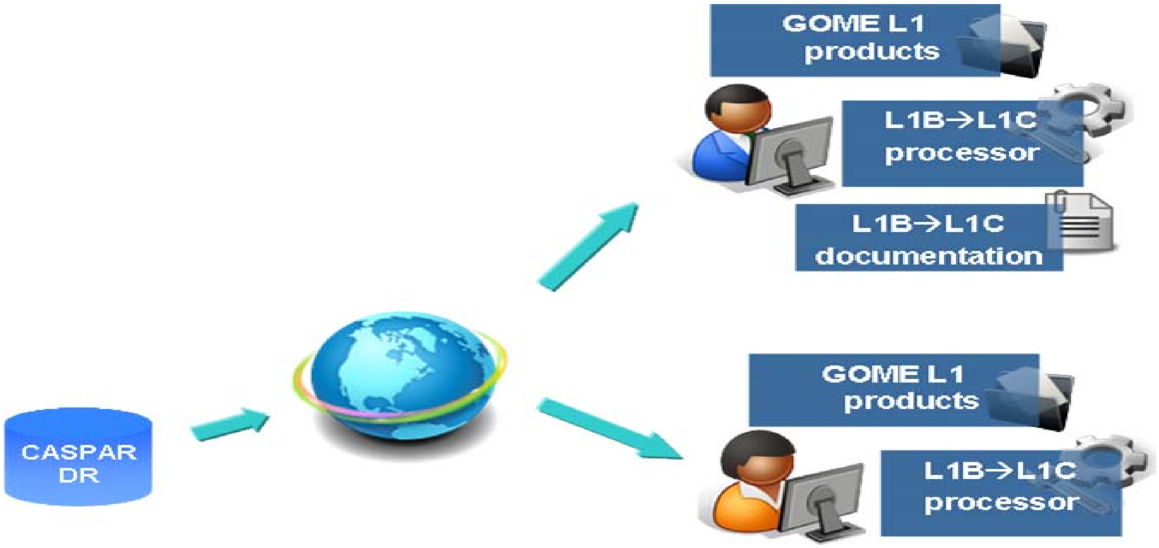

Figure 3 – Different Representation Information are returned to different users

CONCLUSION

Data play a central role in the implementation and the systematic monitoring of international Environmental conventions. Therefore, there is a strong need to easily discover and access data as well as to adopt proper and effective curation and preservation strategies.

In this paper we have described the results of a collaboration between two EC funded and ESA participated projects, namely CASPAR and GENESI-DR, addressing data curation and data discovery and access respectively. Different scenarios have been developed to demonstrate how the ES dedicated platform implemented by GENESI-DR can provide access to data that are preserved using the approach proposed by CASPAR.

REFERENCES

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

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

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

[website] http://www.genesi-dr.eu/

ACKNOWLEDGEMENTS

Many thanks to Vincenzo Beruti (ESA-ESRIN), Luigi Fusco (ESA-ESRIN) and David Giaretta (STFC) 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.

Roberto Cossu

Roberto Cossu received the "Laurea" (M.S.) degree in electronic engineering (summa cum laude) from the University of Genoa, Italy, in 1999 and the PhD in Image Processing and Pattern Recognition from the University of Trento, Italy, in 2002. Since 2000, he has collaborated with the Remote Sensing Laboratory of the University of Trento. In 2002 and 2003, he has been working for INRIA, Sophia Antipolis, France. Since 2005, he is working for the European Space Agency (ESA), where he is now in charge of the ESA contribution to several ESA and EC-funded RTD projects. His main research activity is in the use of innovative technology for accelerating the access and exploitation of EO data. In particular, his interests include the utilisation of GRID and emerging Web-based technologies for EO and environmental applications, such as flood mapping, land-cover map generation, partially unsupervised updating of land-cover maps.

Eliana Li Santi

Eliana Li Santi received a full marks degree in Computer Engineering (with a specialization in Robotics and Industrial Automation) from the University of Pisa (Italy) in 2004. Her professional career began in Milan, where she worked as a consultant for ENI S.p.A in the EAI (Enterprise Application Integration) functional area. Since 2006 she has worked as a functional analyst on several projects in the Defence and Space fields (such as Vessel Traffic System and Cosmo SkyMed); she presently works at the European Space Agency (Earth Observation Applications Strategy Office, ESRIN establishment) in the framework of the GENESI-DR project.