# EARTH OBSERVATION ARCHIVES IN DIGITAL LIBRARY AND GRID INFRASTRUCTURES

***Luigi Fusco*[1*]*, Joost van Bemmelen*[2]**

[*1] *European Space Agency – ESRIN, Via Galileo Galilei, 00044 Frascati - Italy*
*Email: luigi.fusco@esa.int*
[2] *Intecs c/o European Space Agency – ESRIN, Via Galileo Galilei, 00044 Frascati - Italy*
*Email: Joost.van.Bemmelen@esa.int*

## *ABSTRACT*

*Earth Observation Missions provide continuous surveillance of the Earth regardless of atmospheric conditions producing huge amounts of data every year that need to be processed, elaborated, appraised and archived by dedicated systems. Emerging institutional and international environmental initiatives, like the ESA and EC Global Monitoring for Environment and Security (GMES), require access to full historical data collections, including the performed data elaborations, scientific analysis, models and results. The historical ESA Earth Observation archives account for Petabytes data holding, which is augmented, since the launch of Envisat in 2002, by some 500 Terabytes per year. The access and utilisation of these archives is an important measurement for long-term data preservation; improving it is a continuous challenge at programmatic, technological and operational level. This article describes how Digital Library and Grid technology can support the underlying infrastructure for long-term data preservation.*

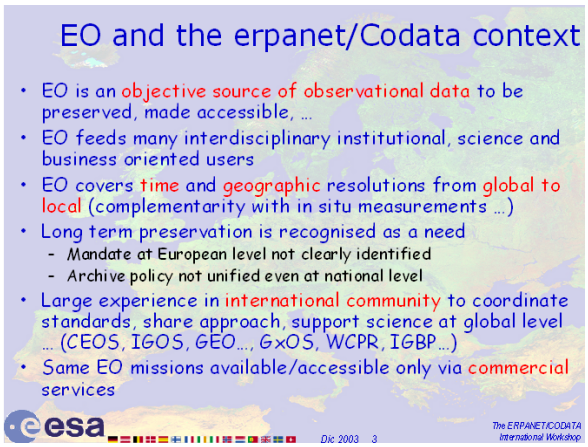
**Keywords:** Earth observation, Infrastructures, Long-term data preservation, Grid, Digital Libraries

## 1 INTRODUCTION

Much more digital content is available and worth preserving; researchers increasingly depend on digital resources and assume that they will be preserved (Hedstrom, 2003, August). Long-term data preservation (LTDP) is a challenging task that requires a common data policy, storage strategy and compatible technology on a worldwide level. LTDP was recently discussed at the ERPANET/CODATA workshop on "The Selection, Appraisal and Retention of Digital Scientific Data" with the clear vision that the issue is common to all sciences (ERPANET, 2004). It was universally agreed that context is of crucial importance in enabling reuse of digital data and that this could only be guaranteed through the application of quality metadata. Metadata and interoperability of both data and metadata have a major impact on accessibility of data and could be used as a means of appraising the long-term value of data. Communication and collaboration are considered important elements, e.g., for exchanging information, deciding about policies, cooperation in using standards and strategies and gaining social benefits. These types of activities can benefit from an adequate underlined infrastructure, supported by emerging information and communications technology, such as Digital Libraries (DL) and Grid. In fact, these networked infrastructures can handle large and distributed data collections and user communities, and facilitate the handling of multiple and partial copies of distributed data sets. This paper describes LTDP for the Earth science community and technologies and initiatives at the European Space Agency (n.d.) to support LTDP. Finally some conclusions and references are given. The figures included in this article are extracted from the ESA-presentation on LTDP presented by Fusco at the ERPANET/CODATA Workshop in Lisbon (ERPANET, 2004).

## 2 LTDP FOR THE EARTH SCIENCE COMMUNITY

Envisat (n.d.), the advanced European polar-orbiting Earth observation satellite launched early 2002, carrying a payload of 10 instruments, is responsible for nearly 500 Terabytes of Earth Observation (EO) data generated and archived every year (i.e. more than a terabyte per day). It provides huge amounts of measurements of the atmosphere, ocean, land, and ice. These data, as well as data from other Earth observation missions, are used in various kinds of processes in Earth science where they are elaborated in line with objectives from international initiatives like the EC/ESA GMES - Global Monitoring for Environment and Security (n.d.). The need of providing a unifying strategy for current and future EO activities in Europe, in view of presenting the user communities an "open" and "operational" environment for data access and utilisation, across the multiple data acquisition, archiving and processing facilities is recognised in the ESA Oxygen (Achache, 2003, June) initiative. It is important to preserve these data, their elaborated products as well as the ancillary and auxiliary data, science algorithms, models and other relevant information, i.e. the knowledge about how they were generated, the reason(s) why certain choices were done, etc. Thus, the future Earth science community generations will be allowed to exploit the acquired knowledge, i.e., to understand how and the circumstances under which products were produced, to reproduce the same results and for long-term analysis studies.




But, in Earth science, accessing historical data, information and related knowledge may be quite complex and sometimes difficult, both, due to the lack of descriptive information (metadata) that could provide the context in which they fit, but also because of the lack of the information and knowledge themselves. There is no clear mandate at European level to preserve EO mission data, relevant information and knowledge. The responsibility falls under the remit of the individual mission owners and/or national archive holders. Coordinating efforts on standards, approaches and use of emerging technology to preserve the most valuable European EO data will be required to guarantee access and reusability of these, often distributed, data. Within the EO-community, it is considered an important issue, e.g., within the Committee of Earth Observation Satellites (CEOS) that is looking at the use of XML for Science Data Access (Suresh & McDonald, 2002) and within ESA (e.g., see PV-2004 Workshop on Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data, 2004).

## 3 TECHNOLOGIES AND EXPERIENCES AT ESA TO SUPPORT LTDP

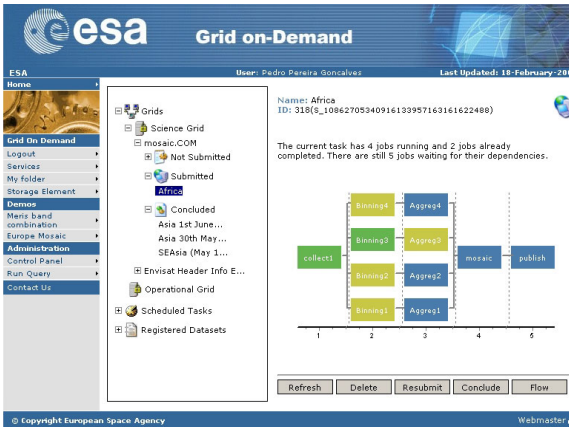
The last few years, different communities have tackled the LTDP problem experimenting different technologies. These include semantic Web focussing on semantic-Web access, data Grid technology focussing on management of distributed data, digital library technology focussing on publication, and persistent archive technology focussing on management of technology evolution (Moore, 2003). This paragraph gives a short overview of some of the emerging technologies of interest for LTDP that have been or are being experimented in different initiatives at ESA.

### 3.1 Grid Technology and Web Services

Foster, (2002, July 22) characterises a Grid (also referred to as the new World Wide Web) as a system that:

- Coordinates resources that are not subject to centralized control;
- Uses standard, open, general-purpose protocols and interfaces;
- Delivers nontrivial qualities of service.

The Grid offers the capability to assign processing power, storage, and server capacity on an as-needed basis and is infrastructure independent. Being a connectivity technology, i.e. a technology that connects & controls distributed resources, including storage, processing power and communication, Grid offers the possibility to improve significantly data access and processing times. Grid marries well with Web services. Moreover, recently announced new Web services specifications (IBM, 2004, January 20) will integrate Grid and Web services standards into a Web services notification and Web Services Resource Framework (WSRF). Grid infrastructures have been experimented and implemented throughout the world in the last few years. An example is the successful European DataGrid (n.d.) project that was the first large-scale international Grid project and the first aiming to deliver a Grid infrastructure to several different Virtual Organisations or dynamic collections of individuals, institutions, and resources (Foster, Kesselman & Tuecke, 2001) for High Energy Physics, Biology and EO. Some Grids have been demonstrated in an operational environment and on-going work is now aimed at creating a reliable and dependable European Grid infrastructure (e.g., in EGEE: Enabling Grids for E-science in Europe, n.d.).



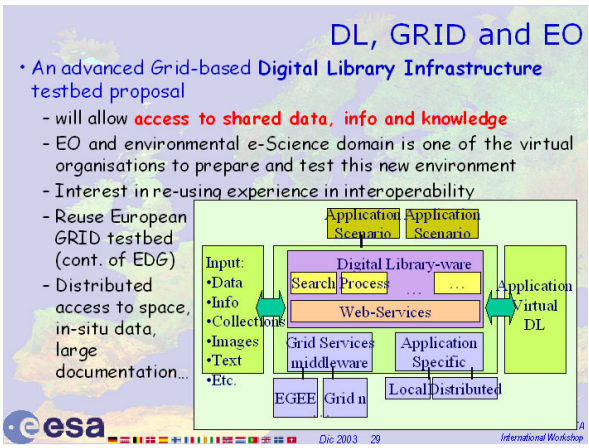

The ESA experience in EDG (having the responsibility of the Earth observation application demonstration) has permitted the development and the deployment of the ESA (n.d.) Grid On-Demand portal. It provides Near Real Time (NRT) access to different level products of various sensors of given ESA Earth observation satellites. It defines a generic infrastructure where specific data handling and application services are seamlessly plugged in. Together with the high-performance data handling processing capability of the Grid, it provides the necessary flexibility for building an application virtual community with quick accessibility to data, computing resources and results. It integrates access to the ESA catalogues and archive systems

using Web services technology. The use of this portal has been demonstrated accessing data from multiple sources, including satellite data, ground-based measurements and climate databases for use in typical EO applications (Ers/GOME, Envisat/GOMOS, AATSR, MERIS) carrying out production and validation of data products (ESA, n.d.).

## 3.2 Digital Library Technology and the DILIGENT project

The term Digital Library (DL) appeared for the first time in 1993, designating collections of electronic information maintained and possessed by the library itself. The concept of DL does however not point to the collections only. DLs may host a whole bunch of functions and services, including storage, discovery, retrieval, and conservation of data and related information. They are seen as an essential element for communication and collaboration among scientists and represent the meeting point of a large number of disciplines and fields including data management, information retrieval, library sciences, document management, information systems, the web, image processing, artificial intelligence, human-computer interaction etc. Even though there is no general consensus about a definition for DL, the one given by Leiner (1998) for the DLib Working Group provides a good picture of what can be done with a DL: "A digital library is the collection of services and the collection of information objects that support users in dealing with information objects and the organisation and presentation of those objects available directly or indirectly via electronic/digital means." Most of current DL systems run in a single organisation, mainly handling textual documents since the handling and preservation of large sets of multimedia documents and data require computational and storage resources that are rarely available in a single organisation. Automatic elaboration of multimedia data, e.g., automatic extraction of contents descriptions of multimedia documents, is often very expensive, and only recently a few experimental DL Management Systems based on distributed architectures have been proposed.


Grid can be considered a valuable complementary technology to the DL technology as it addresses the major DL architecture requirements, including openness, scalability, security and quality. Its data replication and security handling techniques contribute greatly to the definition of new DL preservation techniques, also because Grids provide the abstraction mechanism needed to deal with heterogeneous hardware and software environments (A DIgital Library Infrastructure on Grid Enabled Technology, n.d.). Grid and DL technologies could be used to build so-called Virtual Digital Libraries (VDL), i.e. transient DLs based on shared computational, multimedia and multi-type content and application resources. Functions like content feature extraction, summarization, automatic content source description, etc. on video images and sound, which are based on complex and time-consuming algorithms, could become viable with acceptable performance. In summary, Grid and DL technology could help in performing LTDP. They support a distributed environment capable to handle multiple copies of the same information and, as said above, the preservation task of migrating from old to new technology is really similar to managing distributed data access across multiple sites in Grid, while data and metadata organisation in information collections requires discovery and access techniques as provided within DLs. A combination with semantic technologies , a general area of research that is getting renewed attention now that there is considerable excitement in the vision of the semantic Web (World Wide Web Consortium, 1994), would add even more value to the LTDP process. That is, they allow to complete the data description in a structured but technology and infrastructure independent way




Digital libraries and Grid technology will both be integrated in Diligent, A DIgital Library Infrastructure on Grid Enabled Technology (n.d.). This EC project will create an advanced test-bed allowing members of dynamic virtual organisations to access shared knowledge and collaborate in a coordinated, secure, dynamic and cost-effective way. It will be able to serve different research as well as industrial applications. A test-bed is planned that will be demonstrated by two complementary real-life application scenarios of which one from the environmental e-Science domain led by ESA. The project's kick-off was September 2004 and it will last three years.

## 4 CONCLUSIONS

We have discussed a few issues related to technologies for LTDP and have given some examples of initiatives that use these technologies. Mentioned technologies are not only relevant for LTDP, but they are as well fully relevant for enhancing the *exploitation* of the existing data. A convergence of the two data utilisation views is more and more necessary for the proper handling and preservation of the huge volumes of environmental data available and planned.

For the Earth science community it is important to continue with and invest in activities related to LTDP. Projects like Diligent and initiatives like ERPANET/CODATA need full attention. It does not need to be said that non-accessibility or usability of data, missing information and/or knowledge may be quite expensive.

## 5 REFERENCES

Achache, J. (2003, June) Oxygen, A new strategy for Earth Observation, *Earth Observation Quarterl*y, 71, Retrieved August 20, 2004 from the ESA website: http://esapub.esrin.esa.it/eoq/eoq71/chap1.pdf

A DIgital Library Infrastructure on Grid Enabled Technology (n.d.) Homepage of A DIgital Library Infrastructure on Grid Enabled Technology. Available from http://www.diligentproject.org

*Enabling Grids for E-Science in Europe* (n.d.) Homepage of Enabling Grids for E-Science in Europe. Available from http://www.eu-egee.org

Envisat (n.d.) Homepage of Envisat. Available from http://envisat.esa.int

ERPANET (2004) The Selection, Appraisal and Retention of Digital Science Data (ERPANET/CODATA Workshop). Retrieved August 1, 2004 from the ERPANET website: http://www.erpanet.org/events/2003/lisbon

ESA (n.d.) Grid on-Demand Homepage of ESA Grid on-Demand. Available from http://giserver.esrin.esa.int/

European Space Agency (n.d.) Homepage of European Space Agency. Available from http://www.esa.int

Foster, I. (2002, July 22) What is the Grid? A three-point checklist, GRID Today, 1(6).Retrieved August 20, 2004 from the Grid Today website: http://www.gridtoday.com/02/0722/100136.html

Foster I., Kesselman C. & Tuecke S. (2001) The Anatomy of the Grid, Enabling Scalable Virtual Organizations, International Journal of High Performance Computing Applications, 15(3)

*Global Monitoring for Environment and Security* (n.d.) Homepage of Global Monitoring for Environment and Security. Available from http://www.gmes.info

Hedstrom, M (2003, August) It's About Time: Final Report Workshop on Research Challenges in Digital Archiving and Long-Term Preservation April 12-13, 2002

IBM (2004, January 20) Grid and Web Services Standards to Converge, IBM - Press Release. Retrieved August 20, 2004 from the IBM website: http://www-1.ibm.com/grid/grid_press/pr_120.shtml

Leiner, B. M. (1998) The Scope of the Digital Library, Draft Prepared by for the DLib Working Group on Digital Library Metrics. Retrieved August 20, 2004 from the D-Lib working Group on Digital Library Metrics website: http://www.dlib.org/metrics/public/papers/dig-lib-scope.html

Moore, R. W. (2003) Preservation of Data, SDSC Technical Report 2003-06. San Diego: San Diego Supercomputer Center, University of California.

PV-2004 Workshop on Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data (2004) Homepage of PV-2004 Workshop on Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data. Available from http://www.congrex.nl/04a08/

Suresh, R. & McDonald, K. (2002) XML for Science Data Access, CEOS Joint Sub-Group Meeting, Frascati, Italy.

The DataGrid Project (n.d.) Homepage of The DataGrid Project. Available from http://www.eu-datagrid.org

World Wide Web Consortium (1994) Homepage of World Wide Web Consortium. Avialable from http://www.w3c.org