Emerging Technologies in support of Long-Term Data and Knowledge Preservation for the Earth Science Community

- Experiences with Digital Libraries and Grid at ESA –

Luigi Fusco[1], Joost van Bemmelen[2], and Veronica Guidetti[3]

[1] European Space Agency - ESRIN, Via Galileo Galilei. 00044 Frascati (RM), Italy
luigi.fusco@esa.int
[2] Intecs c/o European Space Agency - ESRIN, Via Galileo Galilei. 00044 Frascati (RM), Italy
joost.van.bemmelen@esa.int
[3]Kelly Services c/o European Space Agency - ESRIN, Via Galileo Galilei. 00044 Frascati (RM), Italy
veronica.guidetti@esa.int

Nowadays, accessing all Earth observation data (real time and historical) is still a non-trivial issue and requires loads of effort from Earth Scientists; being a specialist in their own research area is not enough. There are several difficulties to overcome, some of them due to the fact that data are dispersed over many different geographically dispersed data acquisition sites, archive sites, and other locations, other to the lack of descriptive information, i.e. "associated support information" that helps other users to interpret the data, even in the far future. Emerging technologies, like Grid and digital libraries, are considered with high interest for improving issues related to data access and long-term data and knowledge preservation. This paper provides an update on ESA-ESRIN experiences and status on related initiatives including the EC-FP6 project DILIGENT, the ESA-GSP study THE VOICE, and on ESA Grid on-Demand.

1 Introduction

One of the key responsibilities of ESRIN, the Frascati establishment of the European Space Agency (ESA) in Italy [1], is to operate as the European reference centre for Earth Observation (EO) payload data exploitation acting as the central hub of a network of over thirty globally distributed National and private entities owned archiving and ground stations which receive, process and archive satellite Earth observation (EO) data (the ESA data holdings are estimated in 2 PBytes, with some 4-500 TBytes added per year), distribute them in near real-time and provide on-line access to EO missions related meta-data (over 10 million references), data and derived information products to several thousands of users worldwide, e.g., through its User Data Segments (UDS).

These activities, which are part of ESA's Directorate of Earth Observation Programme, focus on the development and operations of an extensive trans-European and international Earth Observation payload data handling infrastructure, aimed to provide data and services to the European and international science, operational institutional and commercial user communities.

ESRIN started to operate EO missions in 1975, and, since then, has handled data from many EO satellites and missions, such as SeaSAT, HCMM, Nimbus-7, Landsat MSS and TM, AVHRR, MOS, SPOT, JERS, SeaWiFS, etc. [2]. Though the major EO operation activities today are related to the exploitation of the ERS missions (ERS-1 launched in 1991 and retired from operation in early 2000, and ERS-2 launched in 1996) and Envisat (launched in 2002), access to data and related knowledge from historical missions is very much requested as well since they have high continuing research value for which appropriate measures shall be taken to guarantee long-term usefulness of such data and correlated derived information and services. That is, it shall be ensured that the knowledge embedded in the complex relations between the different 'digital objects' being accumulated in the large and distributed archives can be accessed, transferred to and/or retrieved by users in a convenient manner even in a far future.

As described in Van Bemmelen et al. [3], there are various interrelated problems in accessing Earth science data. Some of them due to the fact that data are dispersed over many different geographically distributed data acquisition and archive sites, some of them relate to the multitude of different data and meta data formats, others to the total number of actors involved and, again, others relate to the need to access historic data for which there is the lack of descriptive information (metadata and knowledge, e.g., in science and technical reports), i.e. "associated support information" that helps turning data into "accessible knowledge" for other users. In this process, it is fundamental to include all the scientific work and results immediately related to the same given domain (e.g. instrument calibration, validation, data analysis, models, applications…).

New developments and application of high-bandwidth networks (e.g., Géant), standards (e.g., the Open Archiving Information System [OAIS] Reference Model [4] that, for example, is used in the ESA RTD activity study for an Advanced Data ARchiving System for Earth Observation (ADAR) [5] and the eXtended Mark-up Language) and advanced technologies, like Digital Libraries (DL), Grid, Web-Services, workflow management and persistent archives are providing promising results for dealing with problems related to data access and long-term data and knowledge preservation issues, some of which are described by Fusco et al. [6, 7]. ESA, recognizing their importance for preservation activities for the Earth science community, is applying them in various initiatives it is involved in. This paper provides an overview of some of these initiatives, in particular, it provides an update on experiences and project status of: (1) Diligent, an EC-FP6 project that focuses on integrating Grid and DL technologies towards building a powerful infrastructure to allow geographically distributed researchers to share knowledge and collaborate in a secure, coordinated dynamic and cost-effective manner. (2) The Voice, an ESA study aimed at building an infrastructure based on emerging technologies, including Grid and Web-services to support e-collaboration between researchers in the Earth Observation field. (3) The ESA Grid on-Demand initiative, that supports scientific

applications enabling large ENVISAT data set access. It provides quick accessibility to data, computing resources and results.

2 Preservation and Relevant Technology Developments

Preserving digital objects means building an archival form of them characterising their context and managing their storage. To define the archival form for digital objects, beyond standard descriptive and administrative metadata, the addition of a context and of information on object's integrity and authenticity are necessary. Moreover, actual research is addressing ontologies as a means to complete the description of digital objects: they could be characterised by ontologies written using a standard relationship descriptive syntax. In time, the ontology can be migrated onto new relationship encoding standards, continuing in representing the structure of the digital objects [8].

When reasoning about digital objects preservation it is essential to differentiate between *data*, *information* and *knowledge*. According to Moore [9], data are digital objects, i.e. bit streams, information is any tagged data treated as an attribute, knowledge is the relationships between attributes.

The attributes may be within the digital object or associated with it and give the digital object a context; the relationships between attributes can be procedural/temporal (workflow systems), structural/spatial (GIS systems), logical/semantic (digital libraries cross-walks), functional (scientific feature analysis, feature extraction) [10].

The *storage management* can be intended as logical representation for storage systems (store data), for information repositories (collections, store attributes about data), for knowledge repositories (store relationships between attributes) [11, 12]. Since the Earth science community deals with remote and distributed data, the actual technologies for storage management on the web have to be considered:

- Grids
- Data grids and Data webs
- Digital libraries
- Persistent archives
- Knowledge-based grids and Semantic web

In the following, these technologies are briefly presented. We start with Grids and end with knowledge-based grids, the latter being considered the most complete one of the ones described. Knowledge-based grids have a particular research focus in the project Diligent, described later on in this paper. The information presented in the following is derived from [13].

Grids provide access to distributed resources (computing, storage, sensors etc) and examples of middleware services are remote job execution, remote file access, authentication across administration domains.

Data grids provide mechanisms for managing distributed data extending a grid to include data management: they enable large scale resource sharing of computational and data resources. Digital objects can rely on any generic data management

infrastructure but to be discovered they need to be 'contextualised', that's why they are organised in collections. Data grids work with remote file-based data: data is stored in files and transported via GridFTP.

Data webs are web based infrastructure for directly accessing and browsing remote attribute-based data, designed just as the web browses remote documents.

Digital libraries can be implemented on top of data grids adding support to collections creation, browsing, discovery and user defined metadata definition. Essentially they provide services to discover, access, manipulate information organised in collections. If grids focus on execution of access services, digital libraries focus on the management of the results.

Virtual data grids link multiple data collections and can execute processes to recreate derived data: they integrate data grid and digital library technology to manage processes. In addition to data grids, virtual data grids support the interoperability among services for manipulate, present and discover digital objects.

A *persistent archive* describes archived data as collections, the processes used to create collections and manages the technology evolution: in addition to a virtual data grid it provides mechanisms to manage relationships over time, i.e. interoperability services to migrate collections from old technologies to new technologies. The migration must be possible across media, storage systems, collections, information markup languages standards. Persistent archives can be implemented on data grids by adding integrity and authenticity metadata, needed to assert the invariance of the stored digital objects.

Knowledge-based grids deal with data, information and knowledge throughout services for ingesting, managing and accessing them. 'Knowledge-based' to underline they provide *concept spaces* for discovering relevant digital objects: actually grids have evolved from a file-based access (digital objects identified by path name) and collection-based access (digital objects identified by collection attributes) to a knowledge-based access where digital objects are identified by domain specific concepts (the mapping from concepts used by the specific domain to collection attributes to local file name is then necessary).

Another key of knowledge-based grids is that they allow migrating collections to new database technology forward in time, exploiting the ability to manage collections independently of the information repository or database. By adding new drivers for the new versions of the storage and information repositories grids makes it possible to both store and discover digital objects on new technology and transparently manage replicas between the old and new software systems.

A knowledge-based grid provides a way to integrate digital library and grid technologies, moreover infrastructure independence mechanisms along with authenticity and integrity mechanisms are critical components of a preservation environment.

The *Semantic web* extends the web's HTML infrastructure including semantic information defined by XML and RDF standards: it provides knowledge-based access to data by using ontologies, W3C's standards and agent based architectures.

3 Projects at ESA

3.1ESA's Grid On Demand Initiative

The ESA Grid on-Demand web-portal [14] is a demonstration of a generic, flexible, secure, re-usable, distributed component architecture using Grid & Web-services to manage distributed data and computing resources. Specific data handling and application services can be seamlessly plugged into the system. Coupled with the high-performance data 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.

The main functionality offered by the Grid on-Demand environment can be summarised as follows:

- It supports science users with a common accessible platform for focused e-collaborations, e.g., as needed for calibration and validation, development of new algorithms or generation of high-level and global products.
- It acts as a unique and single access-point to various metadata and data holdings for data discovery, access and sharing.
- It provides the reference environment for the generation of systematic application products coupled with direct archives and near real-time data access.

In particular, the by ESA developed Grid on-Demand Service Infrastructure allows for autonomous discovery and retrieval of information about datasets for any area of interest, exchange of large amounts of EO data products, and triggering concurrent processes to carry out data processing and analysis on-the-fly.

Access to Grid computing resources is handled transparently by the EO Grid interfaces that are based on Web Services technology (HTTP-S and SOAP/XML), and developed by ESA within the DataGrid project (EC Grant IST-2000-25182E). This project, completed recently, has lately demonstrated the potential of Grid systems for providing a suitable infrastructure to ESA's EO scientific users to support their activities related to data and algorithm validation.

The collocation of a Grid on-Demand node with the EO facilities performing data acquisition or data archiving (e.g. ESA PACs) can minimise and optimise the need and availability of high speed networks.

As a typical application, the generation of 10-day composite (e.g., NDVI) over Europe derived from Envisat-MERIS data, involves the reading of some 10-20 Gbytes of Level 2 MERIS data for generation of a final Level 3 product of some 10-20 Mbytes, with a great saving of data circulation and network bandwidth consumption.

Grid on-Demand is used in the projects documented below [15].

3.2 ESA-GSP's Study THE VOICE

The two-phase, early 2004 started ESA General Studies Programme financed study THE VOICE, short for Thematic Vertical Organisations and Implementation of Collaborative Environments [16], analyses how e-collaboration technologies can support the Earth science community. During its first phase a survey of e-collaboration technologies was performed that was matched with results of an analysis of Earth science e-collaboration service requirements to define a service oriented architecture and derive a generic collaborative environment node (GCEN) that serve as a basis for the implementation of selected prototypes, including atmospheric instruments calibration & validation, agricultural production support and decision planning, forest management, ocean monitoring and urban area monitoring during the second phase of the study that started December 2004.

The first phase has demonstrated that most principle needs relate to seamless (and getting the delivery in a relatively short time) access to and/or use of data, information and knowledge without having to worry about where they are, their format, their size, security issues, multiple logins, etc., all essential requirements for long term data preservation. After a careful analysis of prototype requirements, essential and additional services have been derived, and technologies and tools have been selected for implementation as given in the tables below. Besides mentioned technologies, also wireless technologies are used [17].

The study has already implemented the essential services as part of the GCEN and will complete the prototypes before the end of 2005. At the end of the project it will demonstrate near real-life scenario's with distributed actors, resources, data and other relevant items. Next to mentioned technologies and tools, it is also looking into the use of standards like the once defined by OGC and W3C to facilitate data access.

3.3 The EC-FP6 IST Project Diligent

The Diligent [18] envisaged infrastructure perfectly adapts to the actual research in integrating data grids and digital libraries technologies towards semantically advanced and preservation aware environments. The power of a grid infrastructure will help in processing and managing large amounts of satellite data, putting the basis for their long term preservation, while digital libraries services and third party applications will allow the users to build on-demand, retrieve, analyse complex digital objects. Complex and time-consuming algorithms such as services for feature extraction, summarisation, automatic content source description on video, images and sound will become viable with acceptable performance.

The Diligent test-bed knows two complementary real-life application scenarios of which the one named ImpECt, led by ESA, regards the *implementation of environmental conventions*. ImpECt users require the retrieval of Earth Sciences related information by submitting queries based on spatial, topic and time criteria and the accessibility of services and applications able to process this information. Currently existing Earth Sciences digital library systems cannot handle such queries in a sufficient manner and do not host any similar services as those required by the ImpECt scenario.

A first project prototype will use well-known data sources and services, including Envisat and other satellite products as well as services capable to generate and elaborate them. The core feature of the prototype will be the automatic interaction between separated entities as the test digital library and external services able to accept queries from ImpECt users, process the information on the ESA Grid and publish back the results on the digital library.
This activity is intended to allow the users to annotate available contents and services, to arrange contents in user-defined collections, to submit advanced search queries for retrieving georeferenced information, to build user-defined compound services to run specific processing, to keep digital objects like environmental conventions reports alive by an automatic refresh of the information they hold.
The test digital library is built by using the OpenDLib DLMS (http://www.opendlib.com) while the grid infrastructure will rely on the gLite1.1 middleware (http://glite.web.cern.ch/glite/default.asp).
Next steps in the future will allow virtual organisations to create on-demand ad-hoc defined digital libraries, to get newly generated information processed on the grid in a totally transparent way, to navigate the information with the support of domain specific and top level ontologies (Diligent as a knowledge-based grid).

5. Conclusions

Like for many scientific domains, also for the Earth science domain, the capacity both to produce and consume digital information has advanced steadily. These advancements in capacity of information production and consuming require adequate preservation policies and technologies to ensure that the information can be used as well in the future. The paper has given an overview of different ESA initiatives analysing and integrating emerging technologies to improve (digital) Earth science data and related knowledge access in the short and the long-term. First of all, these activities confirm the relevance of preservation activities for the Earth science community and, secondly, they proof that much more work is needed in this area of research. ESA is well aware of the importance of such activities and foresees as well future activities not only considering technologies of interest but also taking into account standards, frameworks and methodologies.

References

1. European Space Agency: Homepage of European Space Agency. Available from http://www.esa.int
2. European Space Agency: Homepage of (Earth Observation) Earthnet online: http://earth.esa.int/
3. Van Bemmelen, J., Fusco, L., Guidetti V.: Access to Distributed Earth Science Data Supported by Emerging Technologies, The 19th international conference EnviroInfo 2005 – Informatics for Environmental Protection, Brno, Czech Republic, September 7-9, (2005).

4. Lavoie, B.F.: The Open Archival Information System Reference Model: Introductory Guide, Office of Research OCLC Online Computer Library Center, Inc. January (2004)
5. ADAR: Abstract for PVDST: ADAR, an Advanced Data Archive for Earth Observation earth.esa.int/rtd/Articles/ADAR_PVDST_Abstract.doc
6. Fusco, L., Van Bemmelen, J., Guidetti, V., Castelli, D., Digital library and Grid technologies as infrastructure for Earth observation data archives exploitation and long-term preservation, *PV2004*, Frascati, October 5-7, (2004).
7. Fusco, L., Van Bemmelen, J.,: Earth Observation Archives in Digital Library and Grid Infrastructures, Data Science Journal, Volume 3, pages 222-226, 30 December (2004).
8. Cannataro, M.: Knowledge Discovery and Ontology-based services on the Grid, The First GGF Semantic Grid Workshop held at the Ninth Global Grid Forum on 5 October, (2003) Chicago IL, USA
9. Moore, R. W.: Preservation of Data, Information and Knowledge R. Moore, Proceedings of the World Library Summit, Singapore, April (2002).
10. Moore, R.W.: Preservation and Long Term Access to Data and Records in a Knowledge-based Society, Singapore Archives; Singapore; April 23,(2002).
11. Moore, R.W., Rajasekar, A., Wan, M.: Data Grids, Digital Libraries and Persistent Archives: An Integrated Approach to Publishing, Sharing and Archiving Data, Proceedings of the IEEE, Volume: 93, Issue: 3, page(s): 578- 588, March (2005)
12. Moore, R.W.: Data Management Services, San Diego: San Diego Supercomputer Center, University of California, (2001)
13. Moore, R.W.: Knowledge-based Grids, Proceedings of the 18th IEEE Symposium on Mass Storage Systems and Ninth Goddard Conference on Mass Storage Systems and Technologies, San Diego, April (2001).
14. European Space Agency, Grid on-Demand Homepage of ESA: Grid on-Demand. Available from http://eoGrid.esrin.esa.int
15. Fusco, L., Guidetti, V., Van Bemmelen, J. (2005): e-Collaboration and Grid on-Demand computing for Earth Science @ ESA, ERCIM News, No. 61, April (2005), pages 12-13.
16. THE VOICE, Phase 1 Executive summary: TVO-SYS-DAT-TN-025-1.0.PDF, ESRIN Contract No. 18104/04/I-OL, April 2005, see also http://www.esa-thevoice.org
17. Betti, P., Camporeale, C., Charvat, K., Fusco, L., Van Bemmelen, J.: Use of Wireless and Multimodality in a Collaborative Environment for the Provision of Open Agricultural Services. XIth year of international conference - Information systems in agriculture and forestry - on the topic e-Collaboration, Prague, Czech Republic, 17-18 May (2005).
18. A DIgital Library Infrastructure on Grid Enabled Technology: Homepage of A DIgital Library Infrastructure on Grid Enabled Technology. Available from http://www.diligentproject.org