EnviroInfo 2005 (Brno)
Informatics for Environmental Protection - Networking Environmental Information
 ISBN: 80-210-3780-6

Access to Distributed Earth Science Data Supported by Emerging Technologies

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

Abstract

Earth scientists use large amounts of electronicly available data, information and associated knowledge from different satellite sensors, aerial campaigns and in-situ measurements in combination with related models, tools and documents to perform their research. Accessing the right data and requested associated information to understand and use them isn't straightforward however. Several reasons are mentioned in this paper. Literature shows (e.g., Moore 2003 and Fusco/Van Bemmelen 2004) that scientist may benefit in their struggle for data by adapting more e-collaborative ways of doing science also known as e-Science (eventhough not always mentioned as such explicitely). This paper provides an outline of how different, as we call them, e-collaboration technologies, including Grid and web-services, digital library and semantic technologies, can help in improving Earth science data access by describing progress in the European Space Agency (ESA) Grid on-Demand, the ESA-GSP (General Studies Programme) study The Voice and the European Commission (EC) Sixth Framework Programme project Diligent.

1. Problems in accessing Earth science data

Earth scientists are often hindered in conducting their research because of problems with locating and accessing the right data, products and other information they require to turn the data into knowledge, i.e. to be able to interprete the data. Data provission services are far from optimal for reasons related both to science and infrastructure capabilities. The process of identifying and accessing data typically takes up the most time and money. Different causes are at the base to this of which the more frequent ones relate to:

- *The scatterness of data*. Data are often dispersed over different data centres and local archives distributed all over Europe and abroad, and, inherent to this, the different policies (access, costs) applied, the variety in interoperability, confidentiallity and search protocols as well as the diversity in data storage formats. To access this multitude of data storages, users need to know how or where to find them and possess a lot of technical/system background to interface the individual systems. Furthermore, often only the catalogues can be accessed online, while the data themselves are still to be retrieved off-line.
- *The diversity in (meta) data formats.* New data formats are being introduced every day, not only because of the multitude of data centres, but also, e.g., as a result of advances in science and (satellite and sensor) instrumentation that create entirely new types of data for research.
- *The total mass of data.* The total quantity of information produced, exchanged and requested is enormous and is expected to grow during the next decades even faster than it did before in part as a result of the revolution in computational capacity and connectivity and advances in hardware and software that together are expanding the quality and quantity of research data and are providing scientists with a greatly increased capacity for data gathering, analysis and dissemination (International Council for Science 2004). For example, the early 2002 launched ESA Envisat satellite (http://envisat.esa.int), with 10

[1] Intecs SpA c/o ESA-ESRIN (Italy), Via Galileo Galilei. 00044 Frascati (RM), Email: Joost.van.Bemmelen@esa.int
[2] ESA ESRIN (Italy), Via Galileo Galilei. 00044 Frascati (RM), Email:luigi.fusco@esa.int
[3] Kelly Services c/o ESA-ESRIN (Italy), Via Galileo Galilei. 00044 Frascati (RM), Email: veronica.guidetti@esa.int

sensors aboard, increases the total quantity of data available with some 500 Terabytes of data per year while the ESA ERS satellites produced roughly 5-10 times less data per year. Moreover, large volume data access is a continuous challenge for the Earth science community. E.g., the validation of Earth remote sensing satellite instrument data and the development of algorithms for performing the necessary calibration and geophysical parameters extraction often require a large amount of processing resources and highly interactive access to large amounts of data to improve the statistical significance of the process. The same is true when users need to perform data mining or fusion for specific applications. As alternative to the traditional approach to deliver data products from the acquisition/storage facilities to the user site, "ad-hoc" user specified data processing modules could be moved there where data are available in real time to let the end-to-end EO data exploitation process benefit in performance.

- *The unavailability of historic data*. Scientists do not work with 'fresh' data only but use historic data as well. For example, for global change research based on the availability from multiple time periods. Here, different problems can be distinguished. First, it shows that often, there is no metadata defined (or no common metadata standards are being used) and relevant knowledge (associated support information, e.g., in science and technical reports) is missing in order for the scientists to understand and use the data. Eventhough this problem is valid for fresh data as well, it yields even more for historic data. It is the metadata that will be at the heart of every effort to preserve digital data in the next few decades; they will be used to create maintenance and migration programs and will provide information on collections for the purpose of orienting preservation strategies and systems (Gauthier 2003). And second, there are no sufficient preservation policies in place that allow for accessing the data after longer periods of time during which new technologies may have been introduced, hardware and software been upgraded, formats may have changed and systems replaced. E.g., it is almost impossible today to read files off of the 8-inch floppy disks that were popular just 25 years ago. Vast amounts of digital information from just 25 years ago is, for all practical purposes, lost (Besser 1999).
- *The quantity of actors involved.* Science is becoming more international and more interdisciplinary resulting in an increased total number of actors (not only human) involved. Just to give a rough number: currently, ESA is serving some sixthousand users in the Earth observation domain and many of them need to exchange data, information and knowledge.

At a global level, e.g., the International Council for Science deals with data access issues (International Council for Science 2004). In Europe, different initiatives are supported by the EC, e.g., as part of their specific action on research infrastructures (part of the 6th Framework Programme) that promotes the development of a fabric of research infrastructures of highest quality and performance and their optimum use on a European scale to ensure that researchers have access to data, tools, models etc, they need.

ESRIN, the ESA establishment in Italy, is participating in different initiatives as well, focussing, in particular, on the use of emerging technologies for data access. This paper provides an overview of these initiatives focussing on the use of selected emerging technologies referred to in the following section.

2. Emerging technologies for data access

Fusco/Van Bemmelen (2004) provides an overview of the use of Grid (also called the *next generation Internet* or the *new World Wide Web*), Web Services and Digital Library technology for long-term data preservation. The same technologies can be used for accessing data in general. Moreover, emerging technologies can support data access, e.g., via infrastructures based on high-speed networks that could drastically speed-up the transfer of the enormous quantities of data, the use of Grid for managing distributed heterogeneous resources including storage, processing power and communication offering the possibility to significantly improve data access and processing times and digital libraries that can help users locating data via advanced data mining techniques and user profiling.

It is reasonable to consider that a shared infrastructure supported by emerging technologies, integrating data dissemination with generic processing facilities, accessible by science (but also industrial) users can be seen as a very valuable and cost-effective approach to support Earth science data access. The following section discusses different projects in-line with this approach.

3. Earth science applications

3.1 Grid on-demand

The ESA Grid on-Demand web-portal (http://eoGrid.esrin.esa.int) 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 (Fusco/Guidetti/Van Bemmelen 2005).

3.2 Thematic vertical organisations and implementation of collaborative environments

The two-phase, early 2004 started ESA General Studies Programme financed study The Voice, short for Thematic Vertical Organisations and Implementation of Collaborative Environments, (http://www.esa-thevoice.org) is looking at 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 ar-

chitecture and derive a so-called 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. 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 (Betti/Camporealle/Charvat/Fusco/Van Bemmelen 2005).

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.

Essential Service	Tool
Job Management	Grid: DataGrid/EGEE gLite
Workflow Management	GridAssist
Data Catalogue	MUIS/INFEO/eoportal
Data Provission	Prototype dependent
Data Configuration Managm.	Grid: DataGrid/EGEE gLite
VO Management	VOMS
AuthN & AuthZ	MyProxy
E-mail	Based on SMTP
Contact Management	VOMS
Problem Tracking	Bugzilla
Notification	JMS
Visualisation	Prototype dependent
Publish & Subscribe	WSRF: WS Notification/JMS

Essential Service	Tool
Action Item Tracking	BSCW
Discussion Forum	P2P: AOL
Calendar	Calendars net
Application Management	Prototype dependent
Earth science related	Prototype dependent
Resource & Data Discovery	Grid: DataGrid/EGEE gLite
AuthN & AuthZ Mapping	Based on MySQL
Instant messaging	P2P: AOL
Video/Audio Conferencing	ISABEL/AccessGrid
File Archive Management	BSCW

Table 1.

Essential and Additional services and tools as defined in the ESA project THE VOICE.

3.3 A digital library infrastructure on grid enabled technology

ESA-ESRIN is leading the Implementation of Environmental Conventions (ImpECt) scenario as part of the EC project Diligent, short for A Digital Library Infrastructure on Grid Enabled Technology (http://www.diligentproject.org) that focuses on integrating Grid and digital library technologies towards building a powerful infrastructure that allows globally spread researchers to collaborate by publishing and accessing data and knowledge in a secure, coordinated dynamic and cost-effective manner.

Main ImpECt requirements concern retrieval of Earth sciences related information based on spatial, topic and temporal selection criteria and the accessibility of services and applications able to process this information. Existing Earth sciences related digital library systems can not handle such queries in a sufficient manner and do not host any similar services as those required by the ImpECt scenario.

A first ImpECt implemenation uses well-known data sources and services, including Envisat and other satellite products as well as services capable to generate and elaborate them. Grid on-demand has a strategic role as service and data provider. The core feature is the automatic interaction between separated entities as the test digital library (DL) and external services able to accept queries from ImpECt users, process the information on the ESA Grid and publish the results on the DL. The test DL is based on OpenDLib

DLMS (http://www.opendlib.com) while the Grid infrastructure relies on the gLite v. 0.1 middleware (http://glite.web.cern.ch/glite/default.asp). The specific information provided in the test DL concerns *ocean chemistry*, in particular ocean colour, being it an Earth sciences consolidated topic with many and different types of information (e.g., environmental reports, images, videos, specific applications, data sets, scientific publications, thesaurus).

The 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 and to maintain heavy documents as environmental conventions reports by an automatic refresh of the information they hold.

Future work will allow virtual organisations to create on-demand ad-hoc defined DLs, to get newly generated information processed on the Grid in a totally trasparent way, and to navigate the information with the support of domain specific and top level ontologies.

4. Conclusions

Emerging technologies can help in easing the accessibility of data and related information needed to interprete the data as is being demonstrated in different initiatives at ESA. There are still a lot of technological challenges, however, that need to be explored further and ESA intends to follow these closely because of their interest to guarantee users transparent access to the ever growing amounts of Earth science data, products and related information from their and third party missions. In particular, the use of Grid, Web-services and Digital Library technology has high priority.

Bibliography

Besser, H. (1999): Digital Longevity, Chapter in Maxine Sitts (ed.) Handbook for Digital Projects: A Management Tool for Preservation and Access, Andover MA: Northeast Document Conservation Center, 2000, pages 155-166.

Betti, P., Camporeale, C., Charvat, K., Fusco, L., Van Bemmelen, J. (2005): 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.

Fusco, L., Van Bemmelen, J., (2004): Earth Observation Archives in Digital Library and Grid Infrastructures, Data Science Journal, Volume 3, pages 222-226, 30 December 2004.

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.

Gauthier, P. (et al) (2003): Ensuring the sustainability of online Cultural and Heritage Content: From an Economic Model to an Adapted Strategy, M2W FIAM - 2003.

International Council for Science (2004): Report of the CSPR Assessment Panel on Scientific Data and Information, December 2004, ISBN 0-930357-60-4

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