Earth Science Community view on Digital Repositories

Luigi FUSCO - ESA
Global sea level rise of ~3mm/year and sea surface temperature increase of ~0.1 deg. C since 1992 (Envisat + ERS).
Satellite data to deliver ‘state-of-the-art’ air quality information

The ‘Integrated Air Quality Platform for Europe’ service was developed to provide end-users information about air quality and is currently providing forecasts for up to 72 hours at a resolution of 50 km. The service includes data on ozone, nitrogen dioxide and particulate matter (the sum of all particles suspended in air, including dust, smoke, pollen,…).

Atmospheric Carbon Dioxide

Published by European Environmental Agency

Contract no. 212073

http://www.esa.int/esaEO/SEMDC5PR4CF_environment_1.html


Mauna Loa, Hawaii, USA (3397 m), NOAA/ESRL
Mace Head, Ireland (25 m), NOAA/ESRL

Michael.Buchwitz@iup.physik.uni-bremen.de
Satellites witness lowest Arctic ice coverage in history

The area covered by sea ice in the Arctic has shrunk to its lowest level this week since satellite measurements began nearly 30 years ago, opening up the Northwest Passage – a long-sought short cut between Europe and Asia that has been historically impassable.

Each mosaic contains approximately 200 Envisat images processed by the Earth Observation G-POD (Grid Processing On Demand) operated at ESA/ESRIN. G-POD is a powerful GRID-based environment coupled with large online archives of Earth Observation data products.

http://www.esa.int/esaEO/SEMYTC13J6F_planet_0.html

World-wide published news / picture!
Earth Science Community expectations

- **Community**
  - Researchers and stakeholders operating over a widespread geographic scale to provide political and technological solutions to global environmental issues

- **Main activities**
  - Sophisticated analysis of sensors data; integration and correlation of different data sources; reasoning, information and knowledge management

- **Requirements**
  - Secure “user friendly data handling environments” where access to huge amount of very different type and sources of information, added-value applications and services, definition of workflows and on-demand processing of data are all seamless tasks

Contract no. 212073
The expected data deluge...

The evolution of all national archives follows a similar trend.

Earth Observation Sat

ESA new EO missions indicates 5-10 times more data to be handled in next 10-15 years.
What is a repository?
(after EC e-SciDR study)

- Not easy to define!
- Some basic (non-defining) architectural components:
  - Storage for content and metadata
  - Input / output and access controls
  - Connectivity
- Many characteristics – no one defining subset:
GENESI-DR will:
- Demonstrate a single way to access space (from ESA and other mission operators) and “non space” Earth Science data products
- Experiment new technologies (e.g. fast data access by users via P2P)
- Provide data products (already planned to be accessible online) in GRID based Storage Elements for further user processing under user responsibility
- …

GENESI-DR is an example of DATA INFRASTRUCTURE
GENESI-DR projects goals

- To provide a base for (establishing) a world-wide e-infrastructure for Earth Science repositories
- To provide reliable, easy, effective, and operational access to a variety of data sources (space and ground)
- To harmonise operations at key Earth Science data repositories
- To demonstrate effective curation and prepare the frame for long term preservation
- To validate capabilities to access distributed repositories for involving new communities, including education…
- To integrate new scientific and technological derived paradigms

Contract no. 212073
Special features in GENESI-DR

- Exploit and strengthen best practice in distributed data archiving, **discovery**, **access** and processing

- A set of services for **indexing**, **searching**, sharing and storing very large spatial data sets.

- Publication of **metadata** and syndication of data between peers automatically

- **Discovery** of data relevant to an application, establishment of **data usage rights** and access

- Access and processing on user demand in open Grid environment
Interfaces for ES data access

• Simple interfaces to complex standards
  – Import/export support for ISO (ISO 19115, 19139) and OGC standards for geospatial metadata and processing (WPS)
  – Conformance with basis of emerging INSPIRE Implementing Rules for Metadata and Discovery

• Specific actions in GENESI-DR
  – Collaboration between OpenGeospatial Consortium and OGF
  – New experimentation based on Atom+GeoRSS+Dublin Core for wider use and OpenSearch protocol with Geo extensions
Data Curation

- **Meta-data curation**
  - including creation, extraction, augmentation (automatic semantic analysis, using machine learning)

- **Data format evolution towards data preservation:**
  - considering variability in Earth Science data
  - Build data format description languages for Earth Science

- **Development of a co-operative model for preservation**
  - Standardise registry/repository for representation information to facilitate sharing

- **Use of emerging technologies for data curation**
  - Semantic approach, knowledge virtualisation etc.
**Digital Repositories Fast Access**

- Focus on guaranteeing fast and reliable access in a real-time manner by
  - Online access
  - Virtualisation of resources, via a single global digital virtual repository, with standard interface
  - Exploitation of the P2P technology like BitTorrent
Aspects in “preservation”

• Digital preservation is hard
  – No organisation/project can guarantee its own longevity
  – Reduction of risk of losing information
  – In the end it depends on money and interest

• Need to document “Preservation Description Information”
  – Response to system changes (in hardware/storage, operating system, …)

• Semantics
  – Documentation of meanings and inter-relationships

• Need to be able to share the load

• Exploit OAIS (Open Archive Information System) concepts to the fullest extent

• Need to preserve information & knowledge – not just “the bits”
  – Data and Documents...
  – more Representation Information

The main issue is metadata

http://www.casparpreserves.eu
Data and Resources Access Policies

• Relation with European and National Policy Bodies
  – Important to establish a working relation with relevant national bodies to address topics such as: long term geoscience archives, new relevant EO mission...

• Development of Data Archiving and Dissemination Policy, e.g.
  – EOPOLE study (data pricing),
  – ESA principles for science data access (Cat-1),
  – US approach to ES data policies

• Development of a European Archiving Sustainability Infrastructure Concept
  – Assuring coherent approach in generalising the archive base infrastructure services for the Earth Science user community.
  – Based on national and European inputs
  – Developments in GMES and INSPIRE shall be special input
GENESI-DR architecture view

Contract no. 212073
• Roadmap and New Challenges
  – Review DEGREE technical results and lessons learned
  – Consider value added Industry development, FP7 strategic objectives
  – Develop concept and specify the technology base for Earth Sciences community dedicated Grid Services Platform
  – Identify new technical challenges and make recommendations for future actions
  – Develop the roadmap for building and promoting the ES Grid services platform

• ES Community Outreach
  – Generate new contacts through existing ES research and industry bodies
  – Promote and advance complementary Grid technologies for providing new / better applications services

DEGREE (support action) is finalising the generation of the ES Grid Roadmap, which is in line with the ambitions of GENESI-DR
ES requires data, tools and worldwide infrastructures to gather and share the data.

Grid can form this infrastructure providing easy and seamless access to large volumes of dispersed data.

ES community has a need for gathering, storing and exchanging data and information.

Grid can be used for building up an e-collaboration platform for sharing data, tools, and information.

The requirements for intensive calculation due to large numbers of files have been increased over the last years.

Grid can provide computational resources otherwise difficult to be collected or managed in one single institute or organization. Furthermore it provides largescale facilities that would not be otherwise accessible to individual researchers or small organizations.

- In addition, Grid technology:
  - Makes possible to apply certain methods (for example inversion algorithm, computations with the Green function, Monte Carlo simulations, optimisation models, 3D cellular automaton with next neighbour interaction and others) otherwise almost unfeasible;
  - Introduces new business opportunities; (it reduces the processing time, reduce the whole project turnaround time, and therefore reduce the costs)
Roadmap highlights

- Community building
  - Maximise approach to support large science community
  - Standard approach to distributed data and metadata
- Support porting new applications to the Grid
  - Data handling tools, Grid storage...
- Propose Earth Sciences community dedicated Grid Services Platforms

GRID as commodity for the ES community to enlarge collaboration / creativity / productivity

form the ES data e-infrastructure
Thank you!

http://www.genesi-dr.eu
http://www.eu-degree.eu

For scientists to access and share seamlessly observations of the Earth System, derived information and knowledge