GEOSS Common Infrastructure and the Big Data challenges

S. Nativi (1), J. Van Bemmelen (2), M. Santoro (1), G. Colangeli (2)
O. Ochiai (3), P. De Salvo (3)

(1) Institute of Atmospheric Pollution Research, National Research Council of Italy
(2) European Space Agency
(3) GEO Secretariat
Group on Earth Observation and Global Earth Observation system of systems

GEO AND GEOSS
The Group on Earth Observation (GEO)

GEO is a partnership of more than 100 national governments and in excess of 100 Participating Organizations that envisions a future where decisions and actions for the benefit of humankind are informed by coordinated, comprehensive and sustained Earth observations. Ministers of the GEO member governments meet periodically to provide the political mandate and overall strategic direction for GEO. The Mexico City Ministerial Declaration from the GEO Ministerial Meeting in 2015 saw world leaders commit to support open Earth observation data for the next decade.
106 Participating Organizations
Global Earth Observation System of Systems (GEOSS)

Together, the GEO community is creating a **Global Earth Observation System of Systems (GEOSS)**.

Earth observations from diverse sources, including satellite, airborne, in-situ platforms, and citizen observatories, when integrated together, provide powerful tools for understanding the past and present conditions of Earth systems, as well as the interplay between them.

GEOSS aims to better integrate observing systems and share data by connecting existing infrastructures.

There are more than 200 million open data resources in GEOSS from more than 150 national and regional providers such as NASA and ESA; international organizations such as WMO and the commercial sector such as Digital Globe.
GEOSS Societal Benefit Areas

- Biodiversity and Ecosystem Sustainability
- Water Resources Management
- Disaster Resilience
- Sustainable Urban Development
- Energy and Mineral Resources Management
- Public Health Surveillance
- Food Security and Sustainable Agriculture
- Infrastructure and Transport Management
IMPLEMENTING GEOSS
GEOSS Applications

GEOSS Providers

GEOSS Application Developers (intermediate Users)

GEOSS end-Users

GEOSS Portal

GEOSS Common Infrastructure

APIs

Mediation modules

DOWNSTREAM

MIDSTREAM

UPSTREAM

GEOSS Supply Chain

SBA 1

Enterprise System 1

Enterprise System 2

Enterprise System 3

Enterprise System j

SBA 2

Enterprise System 4

Enterprise System 2

Enterprise System 3

Enterprise System 4

Enterprise System K

Enterprise System Z

Enterprise System 1

Enterprise System 2

Enterprise System 3

Enterprise System 4

GEOSS Providers

GEOSS Portal
GEOSS Common Infrastructure (GCI)

Data Providers

Societal Benefit Areas

GEOSS Portal

Registration

M2M

GUI

200 million data resources spanning all SBAs
Enhanced GEOSS Portal - Overview

• Enhanced during 2016
• Accessible from www.geoportal.org
• Coordinated with ESA, CNR-IIA, DG-RTD, DG-JRC and GeoSec
• Focus on engagement, delivery and advocating
• Structured in 3 phases
  • 1st phase – 2016: interface restyling: completed
  • 2nd phase – 2017/18: deployment of major upgrades
  • 3rd phase – 2019 onwards – operations and evolutions
GEO Discovery and Access Broker (DAB)

GEO DAB is a **brokering framework** that interconnects hundreds of heterogeneous and autonomous supply systems (the enterprise systems constituting the GEO metasystem) by providing mediation, harmonization and transformation capabilities.
BIG DATA IN GEOSS
Big Data Enabling Technologies

- Computing
- Storage
- Monitoring
- Auto Scaling
- Load Balancing
- Routing
- NoSQL Database Clustering
Big Data challenges for the GCI

VARIETY
Variety in GEOSS

- *Variety* is the most important V for GEOSS.
<table>
<thead>
<tr>
<th>Service</th>
<th>Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>OGC CSW 2.0.2 AP ISO 1.0</td>
<td>INPE</td>
</tr>
<tr>
<td>OGC CSW 2.0.2 ebRIM EO</td>
<td>CKAN</td>
</tr>
<tr>
<td>OGC CSW 2.0.2 ebRIM CIM</td>
<td>DCAT</td>
</tr>
<tr>
<td>ESRI GEOPORTAL 10</td>
<td>GI-cat</td>
</tr>
<tr>
<td>OAI-PMH 2.0</td>
<td>ESRI GEOPORTAL 10</td>
</tr>
<tr>
<td>OpenSearch 1.1</td>
<td>NCML-OD</td>
</tr>
<tr>
<td>OpenSearch 1.1 ESIP</td>
<td>BCODMO</td>
</tr>
<tr>
<td>OpenSearch GENESI DR</td>
<td>NCML-CF</td>
</tr>
<tr>
<td>CKAN</td>
<td>NetCDF-CF 1.4</td>
</tr>
<tr>
<td>CUAHSI HIS-Central</td>
<td>FTP populated with supported metadata types</td>
</tr>
<tr>
<td>ESRI REST API 10.3</td>
<td>WAF Web Accessible Folders</td>
</tr>
<tr>
<td>OGC WCS</td>
<td>GeoNetwork (2.2.0 or greater)</td>
</tr>
<tr>
<td>OGC WMS</td>
<td>Ecological Markup Language 2.1.1</td>
</tr>
<tr>
<td>OGC WFS 1.0.0, 1.1.0, 2.0.0</td>
<td>NERRS (National Estuarine Research Reserve System)</td>
</tr>
<tr>
<td>OGC WMTS</td>
<td>HMA CSW 2.0.2 ebRIM/CIM</td>
</tr>
<tr>
<td>OGC SOS 1.0.0, 2.0.0, 2.0.0 Hydro Profile</td>
<td>HDF</td>
</tr>
<tr>
<td>OGC WPS 1.0.0</td>
<td>IADC DB (MySQL)</td>
</tr>
<tr>
<td>OGC CSW 2.0.0 Core</td>
<td>GrADS-DS</td>
</tr>
<tr>
<td>OGC CSW 2.0.2 AP ISO 1.0</td>
<td>FedEO</td>
</tr>
<tr>
<td>OGC CSW 2.0.2 ebRIM/EO AP</td>
<td>ARPA DB (based on Microsoft SQL)</td>
</tr>
<tr>
<td>OGC CSW 2.0.2 ebRIM/CIM AP</td>
<td>ESRI Map Server</td>
</tr>
<tr>
<td>IRIS Station</td>
<td>SHAPE files (FTP)</td>
</tr>
<tr>
<td>IRIS Event</td>
<td>KISTERS Web - Environment of Canada</td>
</tr>
<tr>
<td>HYRAX THREDDS SERVER 1.9</td>
<td>Environment Canada Hydrometric data (FTP)</td>
</tr>
<tr>
<td>OAI-PMH 2.0 - Harvesting</td>
<td>OpenSearch 1.1</td>
</tr>
<tr>
<td>GBIF</td>
<td>Earth Engine</td>
</tr>
<tr>
<td>DIF</td>
<td>RASAQM</td>
</tr>
<tr>
<td>HYDRO</td>
<td>EGASKRO</td>
</tr>
<tr>
<td>UNAVCO</td>
<td>SITAD (Sistema Informativo Territoriale Ambientale Diffuso)</td>
</tr>
<tr>
<td>CDI 1.04, 1.3, 1.4</td>
<td>File System</td>
</tr>
<tr>
<td>ISO19115-2</td>
<td>GDACS</td>
</tr>
<tr>
<td>THREDDS 1.0.1, 1.0.2</td>
<td>GeoRSS 2.0</td>
</tr>
<tr>
<td>THREDDS-NCISO 1.0.1, 1.0.2</td>
<td>Degree catalog service 2.2</td>
</tr>
<tr>
<td>THREDDS-NCISO-PLUS 1.0.1, 1.0.2</td>
<td>OpenSearch GENESI DR</td>
</tr>
</tbody>
</table>
Adopted Solutions – GEOSS Portal

- **User-centric, considering various user communities:**
  - GEO Flagships and Global initiatives
  - ESA Thematic Exploitation Platforms

- **SBA/Thematic Customization:**
  - **Satellite:** includes smart filters for imagery (Landsat, Sentinel 2) and SAR-type (Sentinel 1) satellite data;
  - **Disater Resilience SBA:** Earthquake events filters
Big Data challenges for the GCI

VOLUME
Volume in GEOSS

• GEOSS has to deal with the large amount of datasets provided by the end systems, e.g. millions of discoverable (small to medium size) products, and long EO time/space series.

• While GEOSS does not store the datasets, it has to collect metadata (at least for harvested catalogs) and provide effective discoverability.
Adopted Solutions

- Dealing with such numbers, normally constrained queries commonly match a large number of datasets.
- GCI addresses this challenge by returning a smaller and/or an ordered result sets.

Ranking and Paging
Ranking and Paging

No-DB SQL

- Good performances on large stores
- No preliminary constraint on data structure
- Need to preliminarily index queryable elements

GEO DAB Internal Metadata Model

Pre-calculated in batch, based on:
- Metadata Quality
- Accessibility
- Etc.

Calculated on the fly, based on:
- Query Constraints

Applied to scores (configurable)
GEOSS View

• Definition:
  – Subset of the whole GEOSS resources defined by applying, via the DAB, a set of clauses
    • *Discovery clauses* (e.g. spatial envelope, keywords, sources, etc.)
    • *Access clauses* (e.g. data format, access protocol, CRS, etc.)
  – Defined “View” exposed on the GEOSS Portal

**Consumer-defined View** – i.e. Client-side These views are available only for the client application which defined the view.

**Provider-defined View** – i.e. Server-side These views are available for all client applications.
Big Data challenges for the GCI

VELOCITY
Velocity in GEOSS

- In GEOSS, *Velocity* related challenges include:
  - Processing rate to transform and preview data
  - Asynchronous approach for data access
  - Real-time (or near real-time) data access
Adopted Solutions – Fast Preview

- GEO DAB provides a fast preview service allowing to get data preview:
  - Metadata record is augmented by adding a reference to data preview; preview tiles at different zoom levels are generated in a batch mode.
  - To store and retrieve single tiles in an efficient way, GEO DAB utilizes a NoSQL key-value DB.
  - When available, GEO DAB utilizes data provider fast preview services by implementing the required mediation.

- GEOSS Portal uses allows Users to quickly evaluate discovered data before deciding the download.
Adopted Solutions – Asynchronous Approach

• In an environment such as GEOSS, no matter which technique is implemented there will always be cases in which the required processing is taking too much time for a click-and-get pattern.

• The DAB + GEOSS Portal transformation allows to deliver discovered datasets according to a common grid format, Coordinate Reference System, spatial and temporal extent and resolution.

• Where the transformation workflow requires a long processing time, Users are allowed to opt for an asynchronous version of the same services.
Adopted Solutions – Real-time (or near real-time)

- GEOSS must support near real-time data discovery and access (i.e. GEOSS must be able to broker near real-time systems)
- Two strategies have been pursued to broker these systems:
  - Provides Users with the most updated content
  - Lower performance
  - Non-consistent ranking
  - GEOSS must support near real-time data discovery and access (i.e. GEOSS must be able to broker near real-time systems)
  - Does not provide Users with the most updated content
  - Good performance
  - Consistent ranking

Two strategies have been pursued to broker these systems:

- Global Biodiversity Facility (GBIF)
- INPE Steallite Imagery
- ESRI ArcGIS Online
Big Data challenges for the GCI

VISUALIZATION
Visualization in GEOSS

• In GEOSS, challenges related to Visualization stem from datasets heterogeneity and volume.
• In addition, GEOSS needs to address the requirement to support diverse (cross-)disciplinary applications targeting different Communities and User categories which have different needs, as for data visualization in an informative and significant way.
Adopted solutions

• GEOSS Portal customization:
  – In addition to what was described in Variety challenge, GEOSS Portal is focusing on providing *resuable Portlets* (for integration in external Community Applications) and *custom visualization of results* (e.g. display seismic events according to magnitude)

• A set of high-level APIs (Application Program Interfaces) have been designed and developed along with documentation and usage examples (the GEO DAB APIs) to allow the development of ad-hoc applications exploiting GEOSS content.
Different APIs for serving diverse Application development use cases (environments)

A set of standard **Web service interfaces**:  
- e.g. OGC service interfaces, CKAN, OAI-PMH, FTP, etc.

A set of APIs for software developers:
- **Client side APIs:**
  - (high-level) **JavaScript library**
  - … . (Python)
- **Server side APIs:**
  - REST/JSON APIs
  - OpenSearch APIs
  - … .
Big Data challenges for the GCI

VERACITY AND VALUE
Veracity and Value in GEOSS

• Giving access to a huge amount of datasets coming from different systems with their own mandate and governance, GEOSS has to consider the veracity and value of the published information.

• Particularly true if considering that GEOSS targets not only research communities, but also decision and policy makers, and therefore the veracity and value of the published information may affect relevant decisions.
Adopted Solutions

• GEOSS Data Management Working Group provides a set of Data Management Principles, including quality-related aspects;

• Essential Variables:
  – EVs can be defined as those parameters required for study, reporting, and management of problems in a specific scientific or societal domains.
  – This effort is particularly important for an infrastructure such as the GCI: the formalization and use of the EVs concept, and related instances, allows extracting the most valuable data matching User's request.
Conclusions

• In the past 10 years GEOSS has developed a truly Global and multidisciplinary System-of Systems
• A valuable framework to experiment and learn how to face Big Data challenges – in particular Variaty and Volume ones.
• The new GEOSS Portal + DAB platform significantly improved the discoverability and accessibility of shared GEOSS resources, addressing more and more User requirements.
Thank you
Backup