‘Big data from Space’ : esperienze in GEOSS e altre iniziative internazionali

PAOLO MAZZETTI, STEFANO NATIVI
CNR-IIA – INSTITUTE OF ATMOSPHERIC POLLUTION RESEARCH, NATIONAL RESEARCH COUNCIL OF ITALY

“BIG DATA, NUOVE TECNOLOGIE E FUTURO DEI GROUND SEGMENT” - ASI, ROME - 26 NOVEMBER 2013
GEO (Group on Earth Observations)

- Launched in response to calls for action by the 2002 World Summit on Sustainable Development, Earth Observation Summits, and by the G8 (Group of Eight) leading industrialized countries

- Voluntary partnership of 88 governments and 61 international organizations

- It provides a framework within which these partners can develop new projects and coordinate their strategies and investments

- Charged with developing GEOSS (Global Earth Observation System of Systems)
GEOSS Common Infrastructure (GCI)
Big Data Challenges

- 3+2 Vs
  - Volume
  - Variety
  - Velocity
- Veracity
- Visualization
GCI & Big Data Challenge

GEOSS Common Infrastructure (GCI) has been addressing

- **Volume:**
  - Millions of discoverable (small & medium) products
  - Long EO time series

- **Variety:**
  - extreme heterogeneity characterizing different disciplines
  - Diverse product types: data, services, models, documents
An ecosystem of multiple Galaxies and contained Planets (data centers)

Realize a Community of Communities
Applying the Systems of Systems
(Network of Networks)
Principles
System of Systems Principles

- **Build on existing** (disciplinary) system/network infrastructures (i.e. Galaxies and Planets)
- **Supplement** but **not supplant** system/network (i.e. Community) mandates and governance arrangements
- Address heterogeneity to **lower entry barriers** avoiding to impose any common (federal) technology
- Be **flexible and extensible** to
  - Interconnect new system/network (Community) infrastructures
  - Sustain and advance the achieved interoperability
  - Allow each system/network (i.e. Community) infrastructure to evolve
Current Assets

Data Providers Brokered (capacities, systems, networks, etc.)

Data Providers successfully Tested (coming soon)
Current Assets

About 20 brokered data providers – capacities, systems, Communities

Publish

More than 7 Million (1.2 Million GEOSS Data Core) Discoverable and potentially Accessible aggregated resources (mix of data collections, datasets and individual images)

Contain [source: data providers]

More than 65 Million (50 Million GEOSS Data Core) Discoverable and potentially Accessible individual resources (e.g. satellite scenes, rain gauge records)
GCI & Big Data Challenge

GEOSS Common Infrastructure (GCI) has been addressing

- **Veracity:**
  - QA/QC
  - Fit for purpose
  - Essential (Community) Variables (EBVs, CEVs, CCEVs, ..)

- **Velocity:**
  - Needs to reach Developing Countries
    –with limited Web/Internet access
Other initiatives

- NSF EarthCube (A Community-Driven Data and Knowledge Environment for the Geosciences)
  - CNR-IIA is co-PI in the Bcube (A Broker Framework for Next Generation Geoscience) project

- RDA (Research Data Alliance)
  - Big Data Interest Group
  - Brokering Interest Group (led by S. Nativi – CNR-IIA)
Conclusions

- Highly distributed systems and multi-disciplinary frameworks need to address Big Data challenges.

- Big Data is not only handling of large datasets but also:
  - Large amount of datasets, not necessarily big (another aspect of **Volume**)
  - Heterogenous datasets in terms of metadata and data models, encodings, formats (**Variety**)

- These requirements will increase in future to integrate information from different observatories (including VGIs)

- GEO is dealing with these issues in building GEOSS
  - The brokered approach demonstrated successful in building a System of heterogeneous Systems

- The brokered approach to address Big Data Variety will be adopted in other international initiatives like NSF EarthCube and RDA.