Harmonizing digital preservation policies for Earth Science data

Fulvio Marelli - ESA

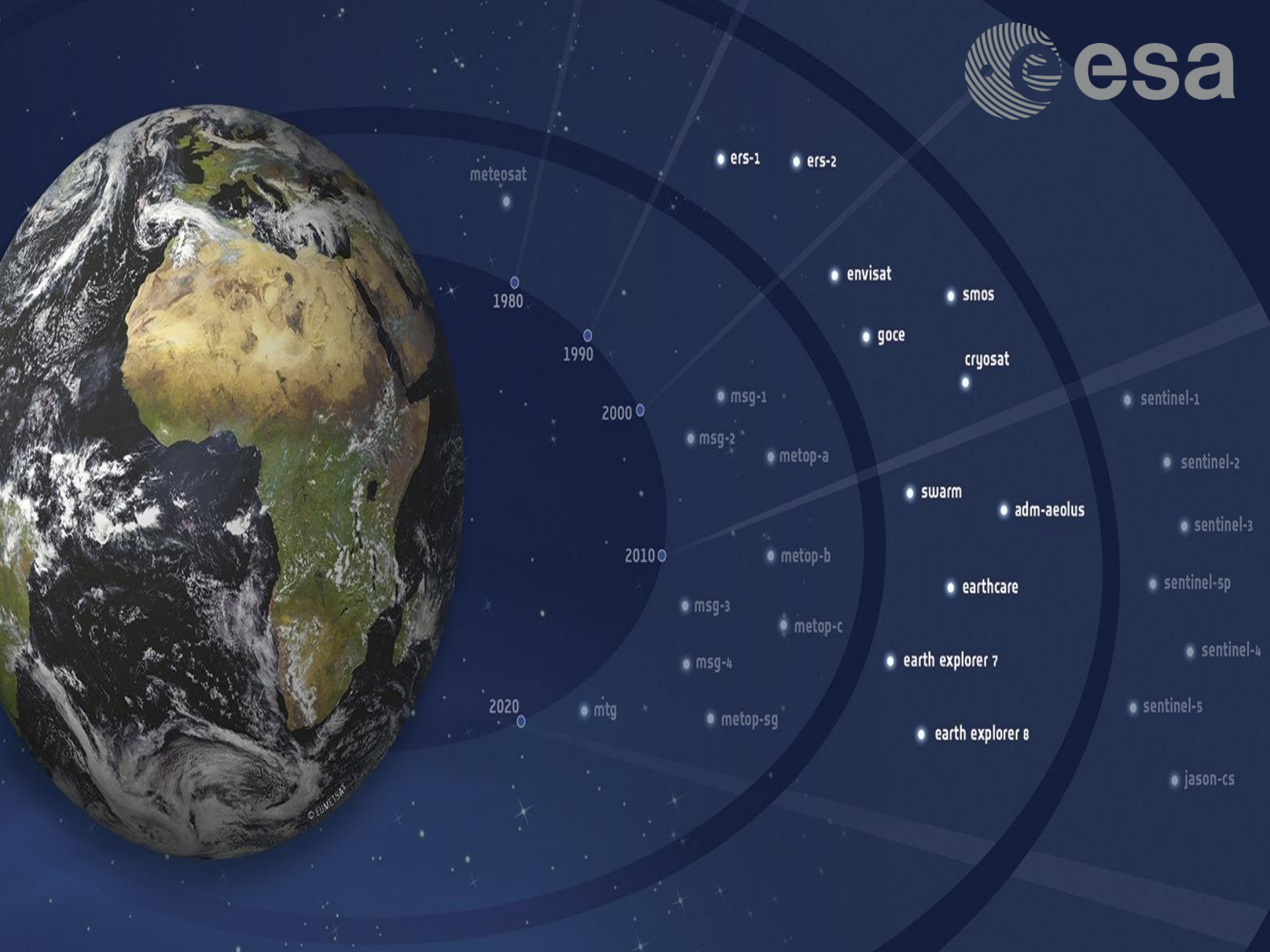
Data is the new gold. "We have a huge goldmine … Let's start mining it."

Neelie Kroes, Vice-President of the European Commission responsible for the Digital Agenda

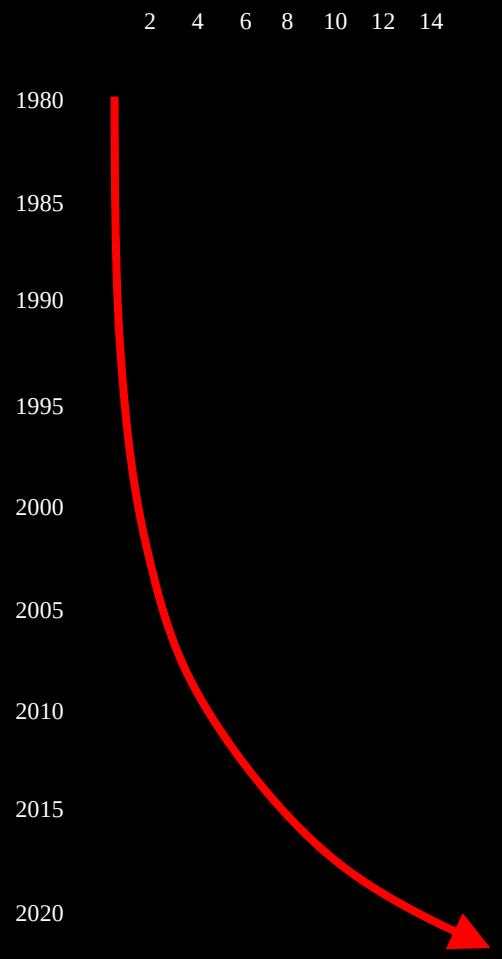
But…

- **Gold** is precious because
  - it is rare
  - it does not combine with other elements
  - it does not perish

- **Data** is valuable because
  - there is so much of it
  - it is more valuable when it is combined together
  - it is FAR from imperishable


esa
meteosat
ers-1
ers-2
envisat
smos
goce
cryosat
1980
1990
2000
2010
2020
msg-1
msg-2
msg-3
msg-4
metop-a
metop-b
metop-c
metop-sg
mtg
swarm
adm-aeolus
earthcare
earth explorer 7
earth explorer 8
sentinel-1
sentinel-2
sentinel-3
sentinel-5p
sentinel-4
sentinel-5
jason-cs
© EUMETSAT


2008-Aug-09 09:47:01 UTC
Lat : 0.0000
Lon : 0.0000
MLST : 09:47:01
SZA : 37.59 deg
Range : 13923 km
Altitude : 13923 km
Intersection Mode ON
Auto Steering ON
SPOT-4
RapidEye-1
ENVISAT
RapidEye-5
SPOT-5
RapidEye-4
SAVOIR - Swath Acquisition Viewer - © European Space Agency - Tattus Software


2 4 6 8 10 12 14
1980
1985
1990
1995
2000
2005
2010
2015
2020

Data & Knowledge Preservation

- ***The preservation of data (the "bytes") is useless without the preservation of the knowledge associated with the data (e.g. the "quality", the process to generate them).***

- We must:
  - Ensure, enhance and facilitate archived data accessibility.
    - Allowing to combine data from different sources and to perform more complex analyses (data unfamiliar become familiar)
  - Ensure coherency of approaches among different Earth Science providers.

- ESA is coordinating the LTDP cooperation activities in the Earth Observation domain with European partners through th LTDP WG formed within the Ground Segment Coordination Body (GSCB). Main goals:

  1. Preserve the European EO space data set for an unlimited time-span.

  2. Ensure and facilitate accessibility and usability of preserved data sets respecting the individual entities applicable data policies.

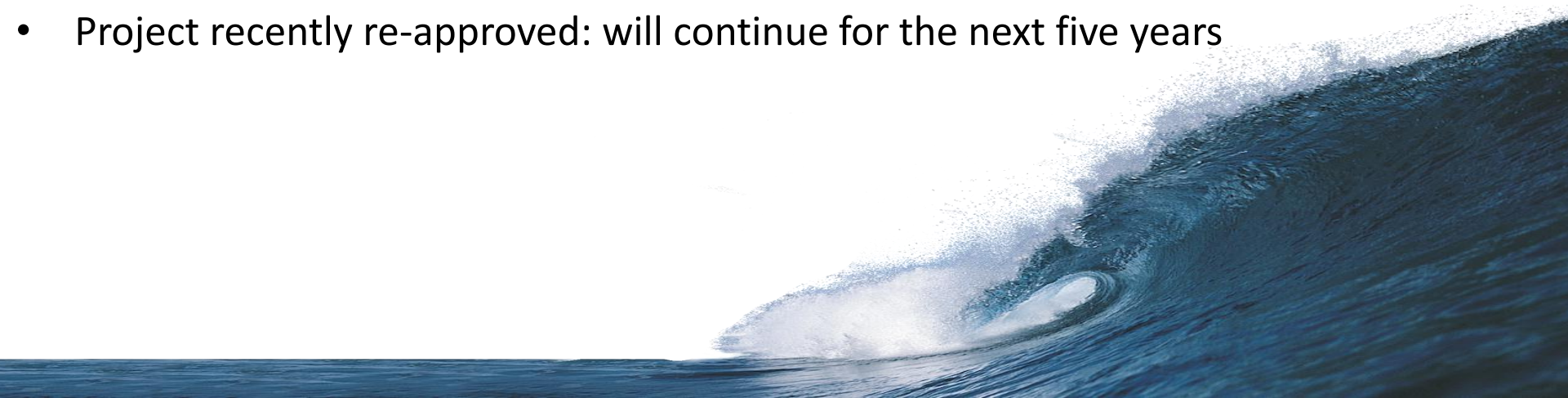
  3. Adopt a cooperative and harmonised collective approach among the data owners and archive holders, based on the application of European LTDP Common Guidelines and sustained through cooperative (multi-source) long term funding schemes.

  4. Ensure, to the maximum extent, the coherency with the preservation of other non-space based environmental data and international policies.

- Implemented the basic rules of the European LTDP Framework in Earth Observation: “LTDP Common Guidelines and Preserved Data Set Content”

  - ✓ Reflecting the consensus of the European EO Data providers.
  - ✓ Reviewed at GEO, CEOS and with NASA.
  - ✓ Being reviewed with QA4EO.
  - ✓ LTDP workshops every two years to disseminate results within the EO/LTDP community

- Project recently re-approved: will continue for the next five years



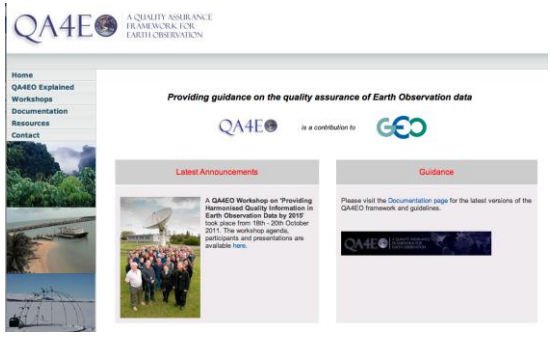
EO-LTDP Working group

- Defined the initial data set to be preserved, including the related glossary.

- Started several technical activities:

  - ✓ LTDP User Requirements Study (FIRST)
  - ✓ Archive Technology Study (LAST)
  - ✓ LTDP Initiatives and Standards Survey
  - ✓ LTDP/QA4EO Study
  - ✓ LTDP Architecture Definition Project

- Guaranteed the information flow through workshops, web sites, participation to conferences and LTDP related events.

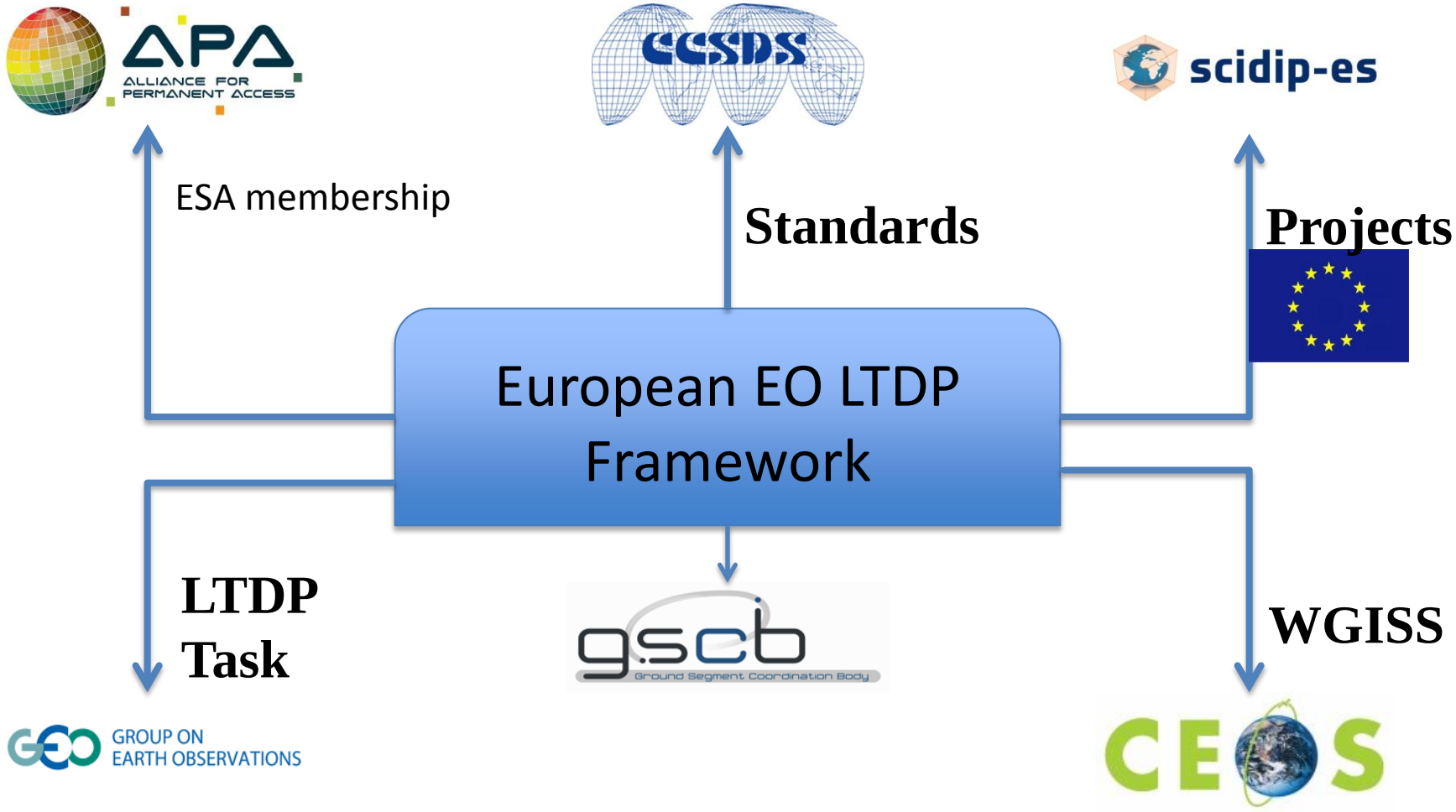


EO-LTDP Guidelines

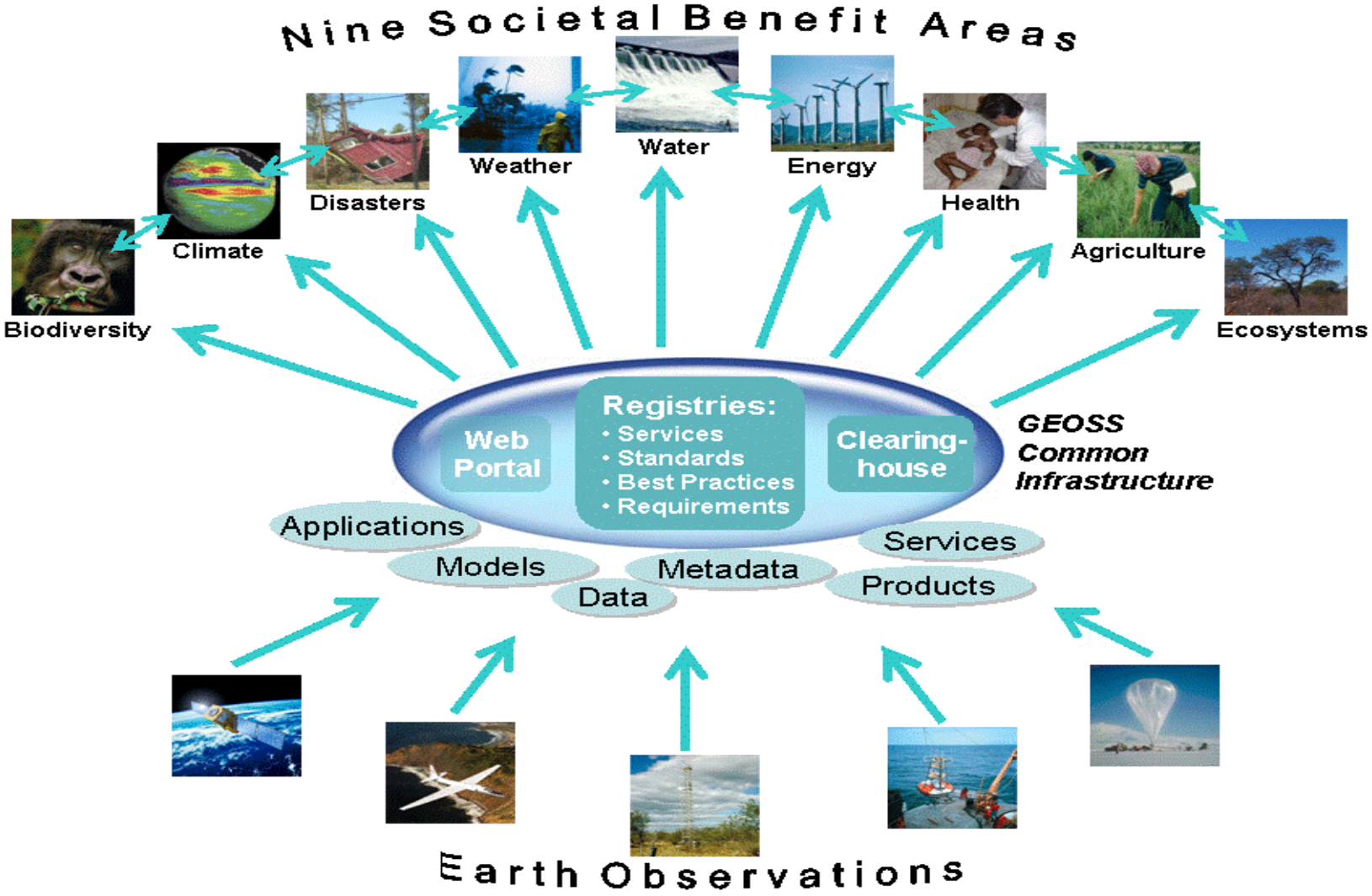
1. Preserved data set content
2. Archive operations and organization
3. Archive security
4. Data ingestion
5. Archive maintenance
6. Data access and interoperability
7. Data exploitation and re-processing
8. Data appraisal and purge prevention

- Institutional web site:
  - ✓ http://earth.esa.int/gscb/ltdp/

International EO LTDP Context



The Earth Science context



The Earth Science context

Earth Science can count 9 different data categories.

Each of them has its own:

- Data preservation policies and approaches (if available)
- Metadata and data formats
- Data description and semantic
- Different also inside each category depending on the owner

Surveys

- Survey of earth science users to assess understanding of issues and level of expertise with respect to long-term data preservation

- Identify current existing and utilised:
  - Data preservation policies and guidelines
  - Metadata, semantic and ontology models
  - Technologies for data discovery, access, management and visualisation

More in detail

- System infrastructure and architecture
- Data discovery
- Data access
- Processing, knowledge, extraction and management
- Data preservation, technologies, policies and guidelines
- Metadata, semantics and ontologies


Thank you

Survey conclusions

- System infrastructure and architecture
  - Archive systems:
    - EO community: proprietary systems/ Others: open source e.g PostGres
    - Often based on tape archives / disk storage for rapid access to frequently used datasets
  - Users generally have a better understanding of the discovery and retrieval system used to access the data than the underlying archive services

- Data discovery
  - Metadata standards ISO19115/19119 widely used in the ES domain
  - Metadata harvesting methods include OpenSearch 1.1 and Open Archives Protocol for Metadata Harvesting (OAI-PMH)
  - Earth Observation data also uses the Earth Observation Metadata profile of observations and measurements

Survey conclusions

- Data access
  - Web services are most commonly used for accessing an archive
  - Web based forms + FTP download or off-line ordering also used for very large data sets
  - Wide variety of portals currently in use for accessing different data types
  - Large cyber infrastructure projects also providing data access services e.g GEOSS, EarthCube

- Processing, knowledge extraction and management
  - Data analysis software separate from data discovery and access services
  - Data processing using a range of tools that are not domain specific
  - Limited number of file formats
  - Trend towards researchers using data beyond their own disciplines and from other geographic regions

Survey conclusions

- Data preservation, technologies, policies and guidelines
  - Archive service providers expect to retain data for up to 10 years
  - Data producers tend towards 5 to 10 years for retention of data
  - Stakeholders would like to retain data indefinitely but in reality this is between 10 and 20 years
  - Data preservation policies most common in EO domain and follow OAIS model and LTDP guidelines
- Metadata, semantics and ontologies
  - Main use is for querying and exchanging data
  - Many of the models are XML based
  - Majority are dealing with geoinformatics
- Findings and recommendations from WP15 and WP33 on data discovery and access will be made available to ES Community for further implementation.
  - Collaboration with other projects (e.g. GENESI-DEC, ENVRI).