1. Introduction

Satellites produce a wealth of data and information regarding the Earth sub-systems (land, atmosphere, oceans, solid Earth and biodiversity) and cross-cutting processes (climate change, sustainable development and security).
Space agencies cooperate together in order to optimize the usage of their data; this is done for example in the context of the CEOS (Committee on Earth Observation Satellites). CEOS is for example a forum of technical exchange on data services interoperability (discover, access, subset, visualization & process), Future Data Architecture (datacube, cloud, analysis ready data, exploitation platform…), …

Most of space data are open and free (with an exception for very high resolution imagery or for some countries or for cooperation with private entities).
The volume of space data is increasing exponentially with some programs like Copernicus, SWOT, NISAR or the last generation of weather forecast satellites (GOES, HIMAWARI, MTG). It represents dozens of PB and will reach several hundreds of PB in the next 3 three years (For example, for NASA alone, the growth rate of the archive will be around 50 PB from mid-2022.).

Given the explosion of data, it is now advisable not to download large volumes of data at home, but rather to move the processing where large data are hosted.

2. CNES organisation and services

CNES is the French space agency. It operates several satellites (JASON, CFOSAT, SARAL, Megha-Tropiques, Calipso, SMOS, IASI, PLEIADES, …) and develops new ones (MERLIN, MICROCARD, IASI-NG, CO3D, …).

Most of the missions are done in cooperation with other space agencies (ESA, EUMETSAT, ISRO, NASA, NOAA, …). In terms of processing, CNES is only responsible for the despatialization of data; that is to say up to a product level that does not require an expertise of the satellite or its instruments. In some cases, depending on the cooperation agreements, the treatments are carried out by our partners. Depending on the case, CNES may be responsible for the distribution of products and their long-term archiving or it may delegate these activities to its partners.
CNES is also very committed to promoting the use of its data and spatial data in general.

Some processing can be executed in the satellites themselves to reduce the amount of data to be transferred to Earth (Edge Computing); this is the case, for example, for the IASI instrument on METOP. But this is not a widespread practice now.

CNES also hosts a mirror of Copernicus data (French Copernicus Collaborative Ground-Segment) which represents about 10 PB (4PB on line and 14 PB capacity on tape).

When a data is processed by CNES, it is on its own computer center. A presentation (made in July 2019) of the CNES computer center can be found here. It allows numerical simulations (HPC), but the data processing is made in a specific HTC/HDA environment. For data reprocessing (very resource intensive), it is foreseen to use cloud-bursting solutions on commercial means.

The goal of CNES and other space agencies is to promote the use of data by the widest possible user communities. These categories of users can be classified macroscopically into two broad categories: research and the downstream sector.

For the downstream sector, the preferred solution to further exploit the data is commercial cloud computing; This for any type of treatments, including AI. For that, there are a large number of solutions:

- In Europe, 5 CDIAS (Copernicus Data and Information Access Services) have been initiated by the European Commission. They propose processing capabilities (cloud computing), services, and additional data. They rely on commercial clouds; namely CloudFerro, Orange, OVH and T-System.
- Several SME initiatives – for example Sinergise Sentinel Hub or Terradue Ellip.
- Amazon and Google propose a wide range of satellite data
    - All sentinel 1&2 (Copernicus) data are already hosted by Amazon. US agencies (NASA, USGS and NOAA) are moving their data to Amazon to promote its use and facilitate treatment with very large data by users (research & downstream).

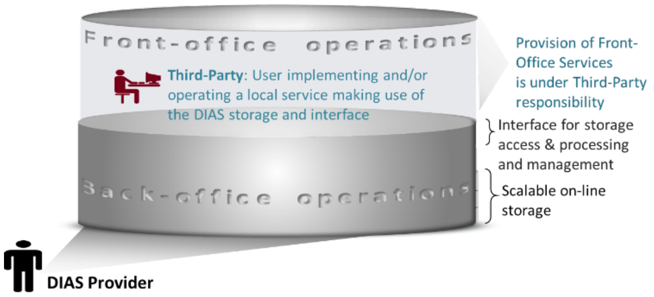

Front-office operations
Third-Party: User implementing and/or operating a local service making use of the DIAS storage and interface
Back-office operations
Provision of Front-Office Services is under Third-Party responsibility
Interface for storage access & processing and management
Scalable on-line storage
DIAS Provider

Figure: DIAS concept

For the research sector in France, satellite data are exploited in five thematic Data & Services hubs:

- AERIS for the atmosphere thematic
- FOR@MATER for the solid Earth thematic
- ODATIS for the ocean thematic
- PNDB for the biodiversity thematic
- THEIA for the land-surface thematic

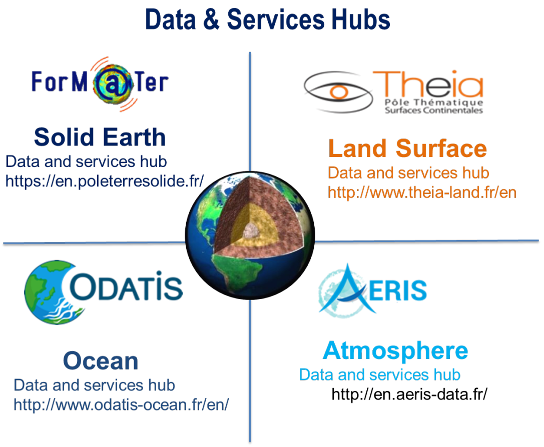


Figure: Earth System Data & Services Hubs in France

Each Data & Services Hub is geographically distributed across multiple data & services centers that all have their own computing capabilities (which mostly are HTC/HDA type clusters). They do not only deal with satellite data, but also a very large amount of in-situ heterogeneous data (ground, sea, airborne…). These data are less bulky than satellite data, but they are much more varied in terms of content, size and formats.

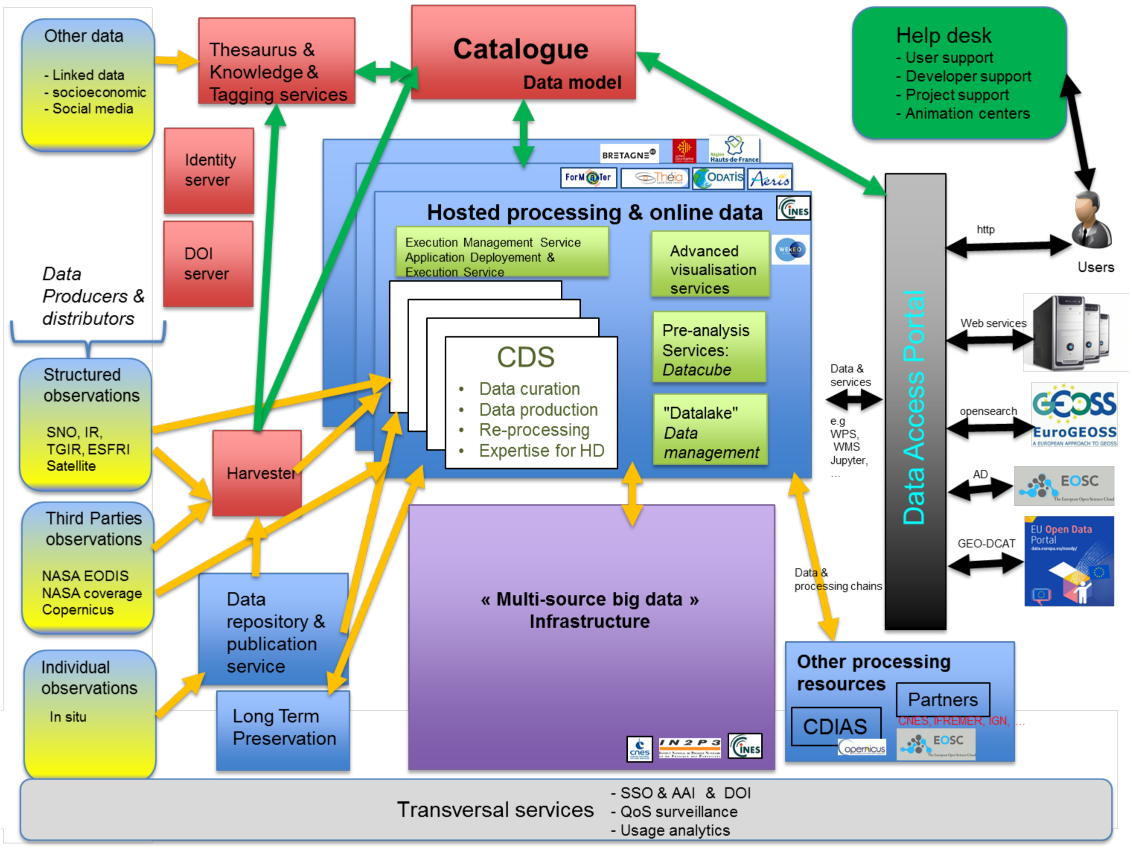
With the rise of IoT and 5G, the volume of in-situ data may explode. In this case it may be necessary to fully review the computing hierarchy of this data and turn to technologies such as Edge Computing, data compression/reduction, use of software defined networks for a smart orchestration of the successive level of resources, support of workflows from end to end (from the edge to the tape of the data center), …
In that sense, some Data & Services Centers use data from models that are processed in HPC centers located in other research infrastructures.

Globally, our capacity to adopt an integrated inter- and trans-disciplinary approach is hampered by the fact that Earth Science data is today highly compartmentalised between these different scientific disciplines and communities.

3. DATA TERRA : toward a fully integrated earth sciences data distributed platform

The Earth is a living system encompassing multi-scale and multi-physics internal dynamical processes and interactions with its external fluid envelopes (e.g., ocean, atmosphere) and continental interfaces (lands, biosphere and anthroposphere). Understanding, monitoring, and predicting the evolution of the Earth's systems in their environments is a fundamental scientific challenge with important societal and economical applications in terms of natural hazards (e.g. volcanoes, earthquakes, tsunamis, landslides), environmental and climate change, new energy resources, sustainable development.

Some countries have undertaken major efforts to restructure this heterogeneous ecosystem, mutualise resources and expertise, and to provide platform of services enabling from end-to-end (edge to the tape) the efficient data logistics including discovery, access, interoperability and wider reuse of data within and beyond Earth Science communities, to society as a whole. The French “DATA TERRA” Research Infrastructure is an example, and a recognized global leader. System Earth was established in France in 2017 to integrate existing data and service hubs and provide easy access to Earth Science data and associated products across the board for scientists and decision makers. With the core mission to facilitate and foster integrated & interdisciplinary research to understand DATA TERRA processes and Global Changes, DATA TERRA is committed to implementing the FAIR principles, creating distributed services, tools and workflows for data management, curation and scientific use; and promoting dialogue and international agreement on best practice.



DATA TERRA High level Architecture

DATA TERRA will have to face several challenges:

- the level of FAIRness of the different centers is heterogeneous
- Bulky data are geographically distributed in heterogeneous computing infrastructure with different level of services; there is a need to develop crosscutting applications then to combine different data in different location (in a distributed software platform) with implication on:
  - Workflows
  - Data logistics
  - Network
  - Computing infrastructure
- Stay open to European cooperation (e.g. ESFRI & ENVRI) and international cooperation (e.g. RDA)
- It will not be possible to build a monolithic and centralized system with all data and processing resources
- In a context of convergence of HPC, HPDA and AI, take into account the moving computing infrastructure landscape in France and in Europe
  - Existing processing capabilities in the Thematic Data and Services Hubs
  - French INFRANUM project led by the French Ministry of Research to concentrate the processing infrastructures at the regional and at the national level (GENCI)
  - EOSC which is the natural solution for ESFRI
  - DIAS

Field
Delet

PRACE the HPC European Research Infrastructure, EuroHPC and EDI (European Data Infrastructure)

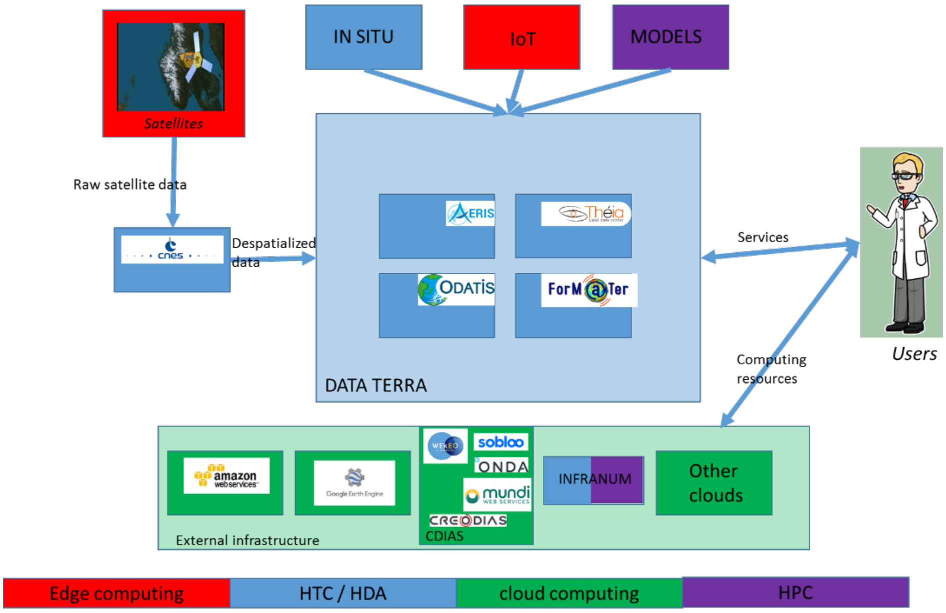


Figure: TERRA DATA transcontinuum workflow

The proposed BDEC proposal is a prototype of an architecture that will allow DATA TERRA to fulfill its objectives.

1. What innovative capabilities/functionalities will the proposed candidate platform demonstrate (e.g. transcontinuum workflow, edge computing, data logistics, distributed reduction engine, etc.)?
The prototypes addresses transcontinuum workflows (cf figure above), edge computing, data logistics.

2. What applications/communities would/could be addressed?
The communities are the Earth Science communities.

3. What is the “platform vision,” i.e. what kind of shared cyberinfrastructure (CI) for science would the further research/design/development of this platform lead to?
Cf two previous figures/

4. How available/ready/complete is the set of software components to be used to build the demonstrator?
To be developed. The CEF (Connecting European Facilities) OpenData/HPC EC project PHIDIAS (duration 3 years from July 2019 with 16 European partners including CNES, CINES, CSC, CERFACS, IS Terre, IRD, IPSL, SPACIA, …) will allow to develop the first elements of the system.

5. As far as one can tell at this early date, to what extent can this be done with existing and/or otherwise available hardware/software/human resources?
Cf point .4

6. What is the potential international footprint of the demonstrator?
System Earth is by nature an international topic. It can be derived at European level in the frame of ESFRI/ENVRI. It can be derived at international level in the frame of GEO or RDA.