GAIA Data IN
GlobAl Integrated EArth Data Implementation Network

Richard Moreno[1], Frédéric Huynh[2], Gilbert Maudire[3], Jean-Christophe Desconnets[4], Damien Boulanger[5], Emilie Deschamps-Ostanciaux[6], Michel Diament[7], Nicolas Baghdadi[8], Nicole Papineau[9], Caroline Blanke[10]

Earth System Research Infrastructure
CNRS, CNES, IFREMER, IRD, IGN, IRSTEA, Météo-France, IPGP

Introduction

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

The Earth sciences share long-term observation-driven (land, sea, air, spatial) research practices, integrating advanced statistical analysis – with increasing use of machine learning techniques – and rapidly growing volumes of data from diverse sources (e.g. from satellite observation, large observational systems, in-situ campaigns, distributed environmental sensors). Observation data is routinely combined with data from experimentation and numerical simulations over a wide spectrum of spatial and temporal scales.

Data flows in the Earth sciences are exploding. They are generated both by observational systems in edge-environments (land, sea, air, space), where resources (energy, communication, storage, compute) are limited, and by numerical simulations in centralised-environments (e.g. HPC, Cloud) where high-performance resources are concentrated. The data flows generated (continuous or sporadic) are heterogenous, in terms of volumes, velocity, value, variety, variability, and veracity. Establishing efficient data management processes is challenging. One of the key challenges is the management of time-sensitive positioning and encoding or formatting of data, relative to its intended end users and the computational resources that they can apply. This challenge includes the FAIRisation of data in increasingly sophisticated analysis and processing workflows.

A growing number of scientific challenges demand predictive understanding and monitoring of a common object or system. They require interdisciplinary and transdisciplinary approaches demanding seamless access to multi-source data generated in different scientific contexts by different observational systems (e.g. land, sea, air, spatial). Composite data is used together with translational tools for collating, processing and analysis in automated workflows, often involving numerical simulations and machine learning techniques. Discovering these data, associated services and resources requires harmonisation (across different disciplines) of data models; as well as standardisation (within disciplines) of data quality norms, integrity and veracity control tools and their documentation; metadata and provenance information; information systems and registries.

Globally, our capacity to adopt an integrated transdisciplinary approach has, until recently, been hampered by the fact that Earth Science data is highly compartmentalised between the many different scientific disciplines and communities active in the field. Datasets are generated, archived and managed in a great variety of formats and volumes by different platforms and research infrastructures. Some fields generate highly structured data; space, seismological and geodetic data, for example have a long history of standardized data and metadata formats and distribution protocols facilitated by international agreements and organisations such as CEOS, GEO, FDSN, UNAVCO and OGC. The situation is much more complex for data observed in-situ (ground, sea, airborne), due to the large variety of observation or monitoring systems and manual sample analysis techniques (physics, chemistry, geology, biology and biodiversity, genomics) deployed in laboratories across the world.

[1] *CNES, Toulouse, France*
[2] *Infrastructure de Recherche système Terre - CNES-CNRS-IFREMER-IGN-IRD-IRSTEA-METEO France -, France*
[3] *Ifremer, Odatis, Brest, France*
[4] *IRD, THEIA, Maison de la Télédétection, Montpellier, France.*
[5,9,10] *CNRS, AERIS, OMP Toulouse, IPSL Paris, France.*
[6, 7] *IPGP, ForM@Ter, Université Paris-Diderot, Paris, France.*
[8] *IRTEA, THEIA, Maison de la Télédétection, Montpellier, France*

Several international initiatives have been established to make in-situ data more accessible and reusable. In Europe, examples include research infrastructures for marine and oceanic observation (e.g. SeaDataNet, EMODNet, JERICO-Next and the ESFRI ARGO and EMSO); ACTRIS, ICOS and IAGOS ESFRIs, or the International Programme MISTRALS, for atmospheric data; and EPOS and EMSO for geophysics and geology. These infrastructures have standardised data management systems and distribution protocols, data and metadata standards, along with thematic and integrated core services. Recently, all these research infrastructures have been federated within a pan-European cluster called ENVRI.

Some countries have gone further and undertaken major efforts to restructure this landscape, mutualise resources and expertise, and to provide service platforms enabling the efficient discovery, access, interoperability, and reuse of data, beyond Earth Science communities. The French "Système Terre" Research Infrastructure (ES-RI) is an example of such efforts, and a recognized global leader. Système Terre was established in 2017 to integrate data and service hubs and provide easy access to Earth science data and tools for scientists and decision makers. Système Terre integrates and interlinks data and services hubs focussing on; atmosphere (AERIS http://en.aeris-data.fr/), solid Earth (FORM@TER https://en.poleterresolide.fr/), ocean (ODATIS http://www.odatis-ocean.fr/en/) and land surfaces (THEIA http://www.theia-land.fr/en), and collaborates closely with the French national biodiversity platform (PNDB https://cat.opidor.fr/index.php/PNDB). As part of its core mission to facilitate and foster integrated & interdisciplinary research on global processes and changes, Système Terre is committed to implementing the FAIR principles, creating distributed services, tools and workflows for data management, curation and scientific re-use; as well as promoting dialogue and international agreements on best practices.

Overarching principle of operation

One of our main driving forces is to avoid re-invention of the wheel by integrating and harmonising developments and beacons of best practice, and improving the interoperability of the existing systems and data. The GAIA Data IN will build on, and promote, extensive pre-existing coordination work, services, tools, and recommendations, whilst seeking to improve transdisciplinary and cross-domain data interoperability. A key objective will be to facilitate access to both space and in-situ data (land, sea, air). To encourage re-use and the sharing of best practices, the GAIA Data IN will extend the scope of working groups exploring technical topics of common interest (DOI, Long Term Preservation, Catalogues, Format, SSO, Data curation and processing, etc.), promote dialogue to seek broader and deeper international agreements, and where appropriate implement recommendations established by the Working Groups of the Research Data Alliance, and other international fora. Our long term goal is to actively participate in the seismic shifts in the global data landscape that is driving the co-design and co-implementation of the EOSC, DIAS and the global Internet of FAIR Data & Services. The GAIA Data IN is open to any new partner or interested party and will actively seek to include or work with organisations across the globe.

Primary objectives/tasks include :

Objective 1: **Define and implement common catalogue standards and distributed services to enable cross-disciplinary data discovery, access, interoperability, and reuse**

Harmonizing and harvesting existing catalogues in different fields and data infrastructures, and aggregating FAIR data to populate common catalogues. This will entail:

- Adoption and maintenance of commonly agreed vocabularies and ontologies.
- Setting up standardized vocabulary servers.
- Seeking agreement on, and harmonization of, metadata standards (ISO or Inspire recommendations for example).
- Seeking agreement on community defined best practice for metadata structure.
- Building and maintaining user interfaces and services to facilitate browsing for discovery using the latest indexing techniques.

Objective 2: **Promote the FAIR principles and implement FAIR distributed services using common or interoperable solutions wherever possible**

The GAIA Data partners and hubs are actively involved in most European environmental & Earth research infrastructures (eg ACTRIS, EPOS, EMSO, etc.) and will contribute to, and implement standards established by ENVRI-FAIR.

Objective 3: **Facilitate the aggregation, the logistics, and the analysis of composite datasets and secondary data products across Earth science fields**.

To achieve this objective GAIA Data will:

- Adopt standards and services both for data interoperability in processing purposes, and for data management and handling along software pipelines and workflows.
- Propose an architecture and software platform of distributed services enabling the combination of large and multi-source data distributed in different data and services centres. This should result in a kind of Earth System Virtual observatory associated to data lakes at different sites.
- Make it easier for any researcher, or interested party, to develop customised data-handling, analysis, aggregation, or visualisation workflows. Different approaches that will be explored include Pangeo / Datacubes and New products: Analysis Ready Data.

Objective 4: **Promote dialogue, exchange and agreements within International initiatives**

The partners of the GAIA Data IN are either partners, or actively participating members in major international fora and organisations including, GEOSS, RDA, ESA Data Coordination Body "EO innovation Europe"', H2020/EOSC projects : ENVRI FAIR, EOSC Pillar, NextGEOSS, and EuroGEOSS, amongst others. These networks, and international meetings or working groups, will be used to diffuse and share best practices, seek agreements at the international level, and promote dialogue to increase the FAIRness of Earth science data.

List of current Partners

The GAIA Data consortium currently includes 34 partners (CNES, CNRS, IFREMER, IRD, IGN, IRSTEA, Météo-France, …), 4 Data Hubs (*www.theia-land.fr ; www.odatis-ocean.fr ; www.aeris-data.fr ; www.poleterresolide.fr*), along with 20 Data & services infrastructures.

Signatories

Frédéric Huynh
Jean-Pierre Villotte
Richard Moreno