



Geo-Inquire Simulation Data lake

Gabriella Scipione (CINECA)

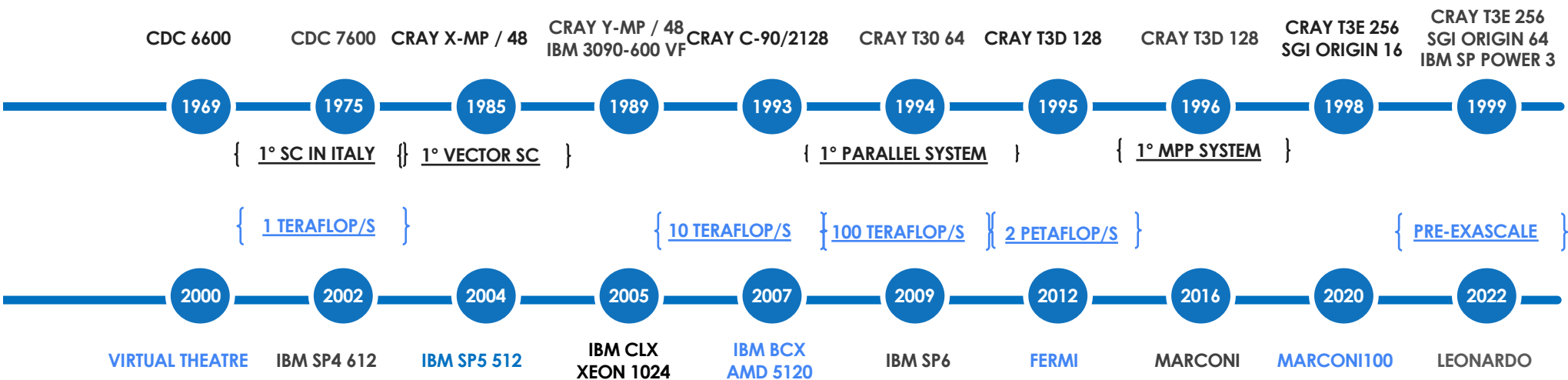
22 June 2023

Geo-INQUIRE is funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

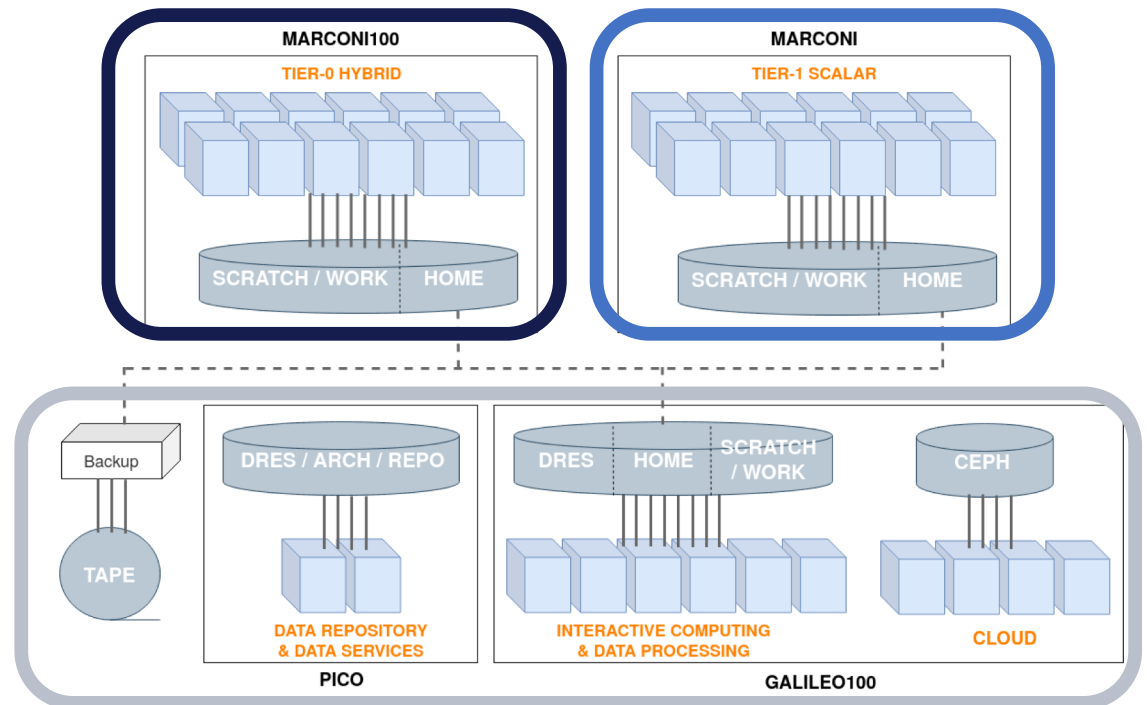
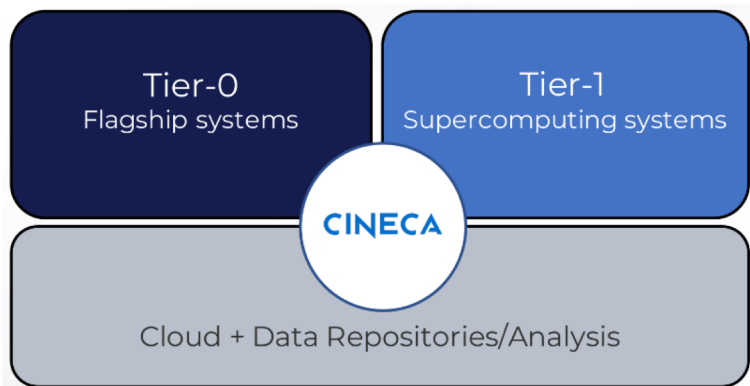


50 YEARS OF SUPERCOMPUTERS

TIMELINE OF CINECA'S SUPERCOMPUTERS



CINECA HPC INFRASTRUCTURE



2022 OVERVIEW

HPC SYSTEMS

CINECA enables world-class scientific research by operating and supporting leading-edge supercomputing technologies and by managing a state-of-the-art and effective environment for the different scientific communities.

CINECA



LEONARDO | 2022

4992 nodes
Booster Module:
32 core per node
4 GPU NVidia Ampere custom
Data Centric Module:
56 cores per node
110 PB Storage
250 PFlops

SOON IN PRODUCTION



MARCONI | 2016

3188 nodes
48 cores per node
612 TB RAM
10 PFlops



MARCONI100 | 2020

980 nodes
32 cores per node
4 GPU Nvidia V100 per node
8 PB Storage
32 PFlops



DGX | 2021

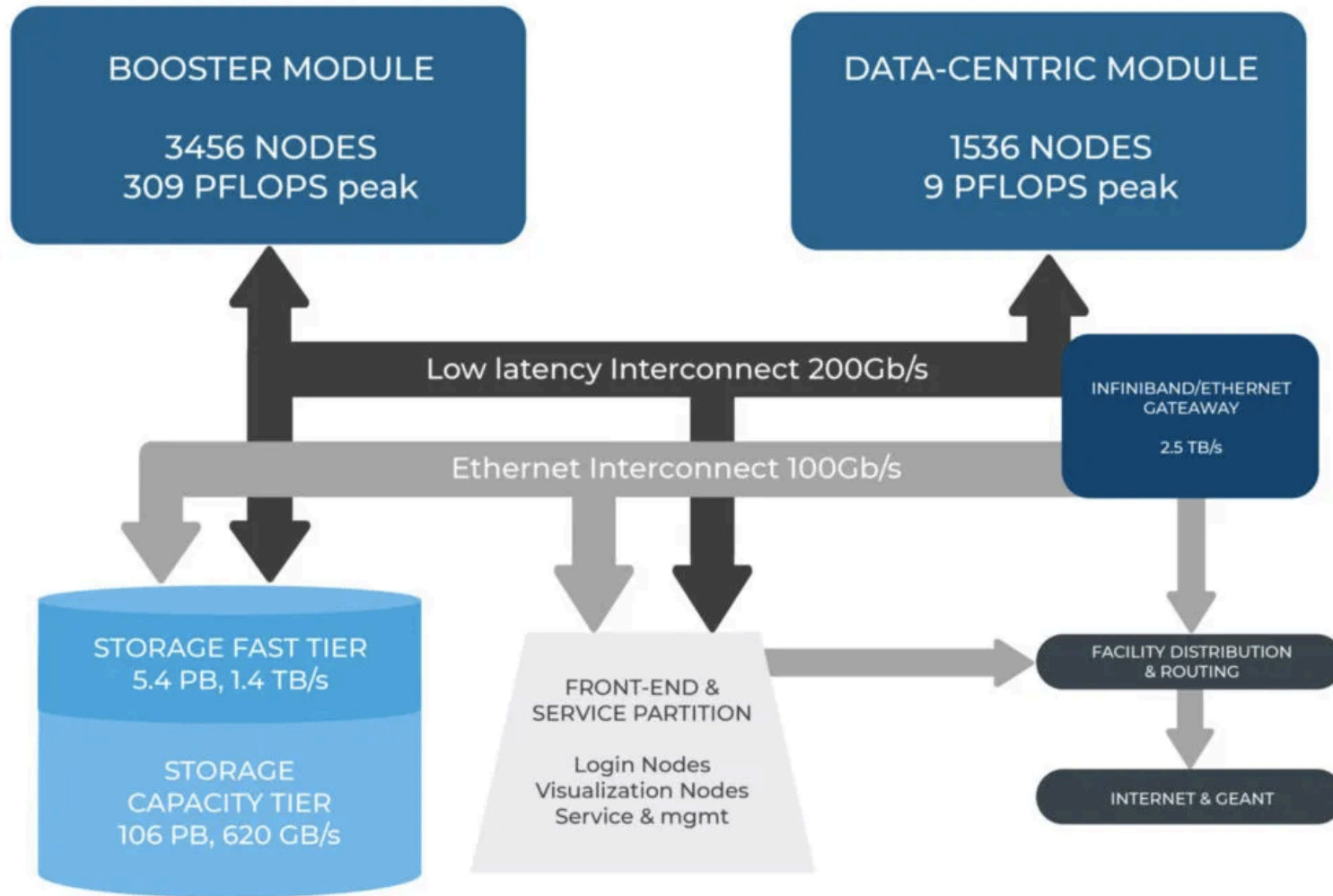
3 nodes
128 cores per node
8 GPU NVIDIA A100 per node
100 TB Storage
15 PFlops



GALILEO100 | 2021

564 nodes
48 cores per node
2 GPU NVIDIA V100 per node
~22 PB Storage
2 PFlops

Leonardo Modular Computing



Tecnopolo di Bologna
1950s structure designed by Ing. Pier Luigi Nervi

ECMWF DC relocation

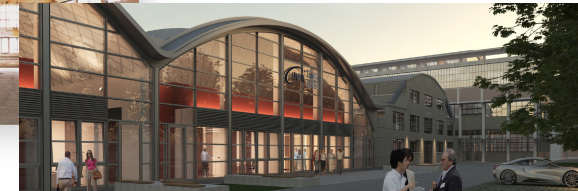
CINECA



Capannone Miscela C2 - LEONARDO



Data Hall Leonardo



Capannone
Miscela C2

Ballette building



Technological center G1



-INQUIRE is funded by
the European Union

Technological tunnels

TUNNEL TECNOLOGICI INTERRATI



MEP connections

CINECA for Geo-INQUIRE

WP5 T5.5

CINECA provides remote Transnational Access using Galileo100 and Leonardo system

WP5 T5.2

Simulation Data Lake (SDL): to provide a solution for the VA to the Simulazion Data Lake as a container for 'raw' simulation results and **in-silico experiment**, some of which run in the HPC@CINECA clusters.



Geo-Inquire Partners requirements for the Simulation Data Lake

User Stories

Questionnaire and survey:

- to better understand datasets (e.g., metadata, formats)
- to better understand the expected functionalities

USERS

- Researcher
- Developer, responsible for software in a scientific community
- Responsible for TA service (providing access to software and computational resources)
- Designer of service provider for data processing

DATA TYPE

- Inputs and outputs of simulations
- Post-processing of simulations
- Intermediate step of a WaaS
- Results of numerical benchmarks and verification tests (published previous works)



Partners requirement: experiment **reproducibility**

- **Dataset definition: Dataset ~ Experiment** (including 1-to-many simulation runs).
- **Data formats:** numerous. We need a standard, at least for outputs (e.g., netCDF, HDF5).

To have experiments reproducibility both

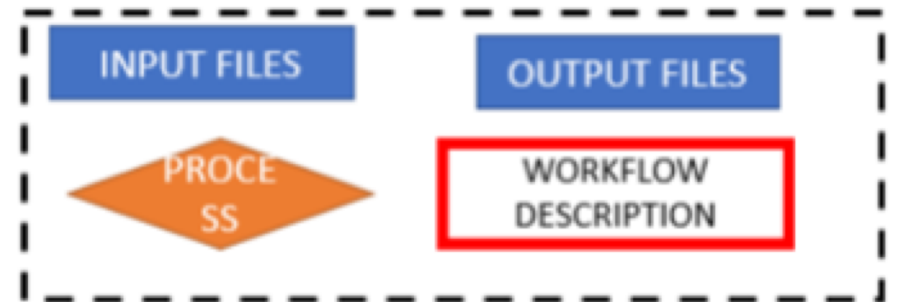
input (e.g., parameter files, config files, input data) and

output of simulations

must be stored in the SDL.

Additionally, the dataset could potentially include scripts/pipelines and a description of the workflow.

So, a dataset should store all the data of an experiment.



Partners requirements for the Simulation Data Lake

Use case/Functionality	Description
AAAI Authentication, Authorization, Accounting, Infrastructure	AA & tracking the access of users and the functions/functionalities they make use of
Data access policy	Define user custom data access policies: public, private, embargo
Create dataset (up to ~1TB)	Create a data collection
Add new data (single file up to ~ 1GB) to dataset	Add data to a data (files) collection
Assign a DOI	Create and assign a DOI to a data collection or object
Assign a PID	Create and assign a DOI to a data collection or object
Assign metadata (add single item and scheme)	Assign metadata. Schemes are TBD by the communities
Query data by metadata	List data matching some metadata criteria
View list of all data owned by a user	List data with a certain author
View list of all data (general usage)	List all data collections stored
View details of a specific data	View details/metadata summary
Get dataset (up to ~1TB)	Download data collection
Get individual data (single file up to ~ 1GB) from dataset	Download data object
Get data from file	Extract data from single file (temporal, geographical slice)
API exposure	API exposure for integration in post-processing pipelines



What is EUDAT?

The **EUDAT Collaborative Data Infrastructure** (or EUDAT CDI) is one of the largest infrastructures of integrated data services and resources supporting research in Europe. It is sustained by a network of more than 20 European research organisations, data and computing centres.



 B2ACCESS	 B2DROP	 B2FIND
B2ACCESS Identity & authorisation View service	B2DROP Sync and share research data View service	B2FIND Find research data, research data portal View service
 B2HANDLE	 B2SAFE	 B2SHARE
B2HANDLE Register your research data with a persistent identifier View service	B2SAFE Keep research data safe via data management policies View service	B2SHARE Store and publish research data View service



Geo-Inquire SDL solution based on B2SHARE



- storing and publishing scientific data



- Find dataset through Metadata



- Assign PID and DOI
- API exposure and Web-platform

Geo-Inquire community created in B2SHARE

The screenshot shows the B2SHARE interface for the Geo-INQUIRE community. At the top, there are logos for B2SHARE and EUDAT, a search bar, and navigation links for HELP, COMMUNITIES, UPLOAD, and CONTACT. Below the navigation, the page title is "Geo-INQUIRE". The main content area includes a description of the community, its creation and update dates, and a table of statistics. A sidebar on the right shows the community logo and name. At the bottom, there is a section for "Community latest records" with a list of recent uploads.

Created at 20/06/2023, 08:18:41
Last updated at 21/06/2023, 07:16:52

Geo-INQUIRE aims to establish a strategic framework for groundbreaking progress in Earth system research. It focuses on improving access to diverse datasets, observations, and data products, promoting a comprehensive approach to geoscience, especially in studying integrated geohazards. The Geo-INQUIRE Community aims to gather and enhance the usability of numerical simulations and in-silico experiments conducted within the geoscience community, driving advancements and innovation in the field.

Identifier: [b8cbc1e9-0dca-4306-97a5-b7d3bb690ee4](#)

Using root schema version: 2

Record views	File downloads
9	1

Records	Files	File size
1	2	4.3 GB

Community latest records

Observational data
21 Jun 2023 by Caroli, Cinzia
Data from ground stations



B2SHARE demo: Communities

The screenshot shows the B2SHARE website interface. At the top, there is a navigation bar with the text "GO TO EUDAT WEBSITE" and a search bar. Below the navigation bar, there are logos for B2SHARE and EUDAT, and a search bar with the text "Search records for...". The main content area is titled "Communities" and displays a grid of community cards. The cards are for Aalto, BBMRI, Geo-INQUIRE, CompBioMed, DRIHM, and EISCAT. The Geo-INQUIRE card is highlighted with a blue border. The Aalto card features the Aalto University logo and the text "Aalto-yliopisto" and "Aalto University". The BBMRI card features the BBMRI logo and the text "Biomedical Research.". The CompBioMed card features the CompBioMed logo and the text "CompBioMed is a European Commission H2020 funded Centre of Excellence focussed on the use and development of computational methods for biomedical applications". The DRIHM card features the DRIHM logo and the text "Meteorology and climate data.". The EISCAT card features the EISCAT logo and the text "Incoherent scatter radar data".


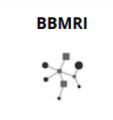

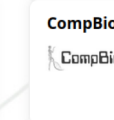
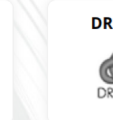











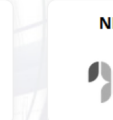
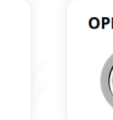



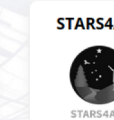



B2SHARE demo: Upload dataset

← → ↻ <https://b2share.eudat.eu/records/new#> ☆ 🔔 ⬇️ 📄 🗑️ ☰

Title

Community

 Aalto Aalto-yliopisto	 BBMRI	 CLARIN	 CompBioMed	 DRIHM	 EISCAT
 EPOS EUROPEAN PLANT OBSERVING SYSTEM	 Geo-INQUIRE	 EUON	 Europlanet-VESPA	 GBIF	 HPC-Europa3
 InGRID	 KiCoS	 LIFE+Respira	 LTER	 NRM	 OPENCoastS
 OpenEBench	 R3PACK	 RDA	 STARS4ALL	 TBopen	

Please select a target community

You can also update the data in an existing record by creating a new version of that record. Search for the 'Create new version' button



B2SHARE demo: Load files

The screenshot displays the B2SHARE web interface for editing a record. The browser address bar shows the URL: <https://b2share.eudat.eu/records/bc6b4b9f996b4cf29aa8e8991b523753/edit>. The page header includes the B2SHARE and EUDAT logos, a search bar, and navigation links for HELP, COMMUNITIES, UPLOAD, and CONTACT. The user profile is identified as [I.rodriuezmunoz@cineca.it](#). The breadcrumb trail indicates the current location: RECORDS > BC6B4B9F996B4CF29AA8E8991B523753 > EDIT. The main content area is titled "Editing draft version" and features a "Delete draft" button. Below this, the "Add files" section contains a large dashed box with the text "Drop files here, or click to select files" and a button labeled "Add B2DROP files". The "Basic fields" section is partially visible, showing a "Community *" dropdown menu with "Geo-INQUIRE" selected.



B2SHARE demo: Metadata (general + community)

← → ↻ <https://b2share.eudat.eu/records/bc6b4b9f996b4cf29aa8e8991b523753/edit> 110% ☆

Titles *

Descriptions

Type *

Creators

Open access * True

Embargo date

License



B2SHARE demo: Save draft or publish directly

The screenshot shows a web browser window with the URL <https://b2share.eudat.eu/records/bc6b4b9f996b4cf29aa8e8991b523753/edit>. The page is titled "Temporal coverages" and "Funding references".

Temporal coverages

- Ranges**
 - Start date:
 - End date:
 - Buttons:
- Spans**
 -
 - Buttons:

Funding references

- Funder name ***
- Buttons:

Submit draft for publication

When the draft is published it will be assigned a PID and a DOI, making it publicly citable. Please note that the published record's files can no longer be modified by its owner.

This publication will get the following DOI: [10.23728/b2share.bc6b4b9f996b4cf29aa8e8991b523753](https://doi.org/10.23728/b2share.bc6b4b9f996b4cf29aa8e8991b523753)

EUDAT Collaborative Data Infrastructure

Acceptable Use Policy | Data Privacy Statement | About EUDAT



B2SHARE | EUDAT Extended Metadata Schema

(https://schema.eudat.eu/eudatcore_metadataelements/)

Metadata elements:

- [Community \(O\)](#)
- [Title \(M\)](#)
- [Description \(R\)](#)
- [Keywords \(R\)](#)
- [Identifier \(M\)](#)
- [RelatedIdentifier \(O\)](#)
- [Creator \(R\)](#)
- [Publisher \(M\)](#)
- [Contributor \(O\)](#)
- [Instrument \(O\)](#)
- [PublicationYear \(M\)](#)
- [Language \(R\)](#)
- [Contact \(O\)](#)
- [Rights \(R\)](#)
- [ResourceType \(O\)](#)
- [Format \(O\)](#)
- [Size \(O\)](#)
- [Version \(O\)](#)
- [FundingReference \(O\)](#)
- [Discipline \(O\)](#)
- [Spatial Coverage \(O\)](#)
- [Temporal Coverage \(O\)](#)

The Extended MD schema
can be further customised
adding elements
community specific



Next Steps



Simulation Data are:

- BIG: in Geo-Inquire dataset size of the order of 10-30 TB
- HPC proximity important, not easy to move the data
- Easy sharing of data with the HPC-CLOUD-IAC
- In Cineca we are providing tools for visualisation, Interactive Computing, and more
- Extraction of data from single datasets and files

**SDL to become the archive for the data assets
in the geoscience community
also supporting the Digital Twins**



Next Steps

At CINECA we are assessing the possibility of **developing a Data Lake solution based on:**

Object-storage and cloud-native technologies:

- High availability and fault tolerance.
- Scalable.
- Updating/maintenance is easier.

- It can potentially be integrated with S3 compatible storage systems in other computing centers or research institutions.

- Early stage of design.



Thank you!

Geo-INQUIRE is a joint effort of 51 institutions



Geo-INQUIRE is funded by the European Commission under project number 101058518 within the HORIZON-INFRA-2021-SERV-01 call.



Dataset example

- Tsunami simulations from INGV: Experiment_Samos.
- 40000 scenarios (simulation runs).
- There's a folder for each scenario.
- Input is grouped in a folder.
- Subfolders need to be compressed.
- JSON metadata (EUDAT Core Schema).
- User needs to download sim_setup and any single simulation to be able to reproduce it.

```
Experiment_Samos
├── BS_scenario00001
│   ├── out_ts.nc
│   ├── out_ts_ptf.nc
│   └── parfile.txt
├── BS_scenario00002
│   ├── out_ts.nc
│   ├── out_ts_ptf.nc
│   └── parfile.txt
├── BS_scenario00003
│   ├── out_ts.nc
│   ├── out_ts_ptf.nc
│   └── parfile.txt
├── BS_scenario00004
│   ├── out_ts.nc
│   ├── out_ts_ptf.nc
│   └── parfile.txt
├── BS_scenario00005
│   ├── out_ts.nc
│   ├── out_ts_ptf.nc
│   └── parfile.txt
├── sim_setup
├── Step1_scenario_list_BS.txt
├── Step2_extract_ts.py
├── Step2_local_domain_2020_1030_samos.grd
├── Step2_local_domain_2020_1030_samos_POIs_deth.dat
└── Step2_ts.dat
```

