Simulation Data Lake

FAIR aspects

CINECA

20 May 2025

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.*

Agenda

1. Simulation Data Lake: **brief overview**

2. Using the SDL:
   - **Web portal**
   - **Command Line Interface (CLI)**
   - **API**

3. SDL **FAIR** support

1. Simulation Data Lake: brief overview

What is the main purpose for which SDL was created?

Store large amounts of **simulation** data

Allow **access** to data in order to facilitate **reusability**

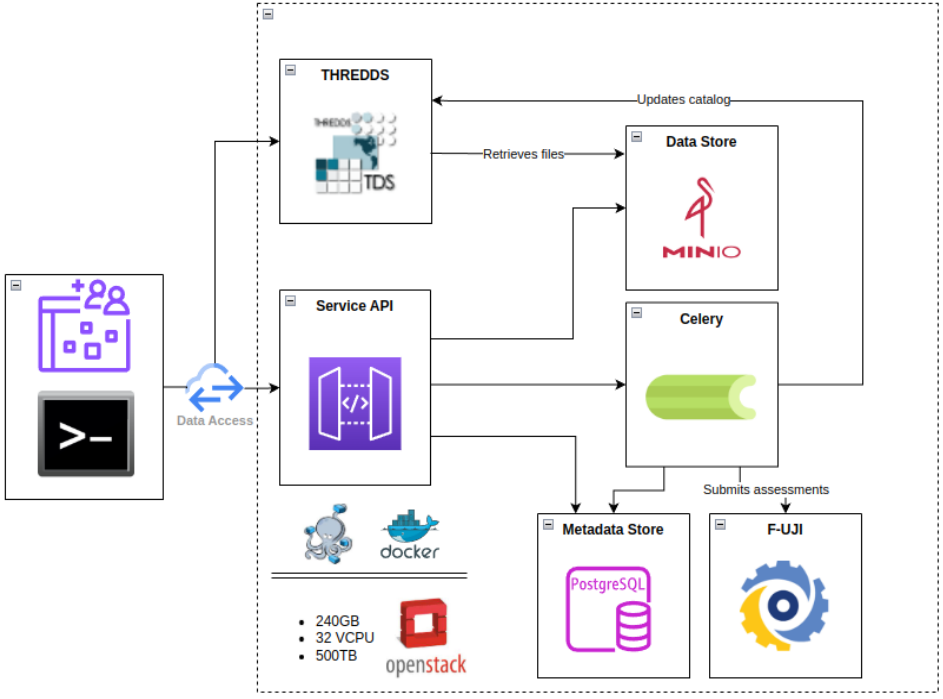
Provide **search** capabilities to find data more **efficiently**

SDL Architecture

Modular architecture and Object Storage

Hosted on **CINECA's Cloud Infrastructure**

- Restful APIs
  - Containerized services: IAM, Catalog, Async Processing
- Data Store: MinIO
  - Compatible with S3
  - Object Storage
- Metadata Store: PostgreSQL - PostGIS
  - Data model
  - Enable data search
- IaaS & Container
  - Docker, Docker-Compose
  - OpenStack

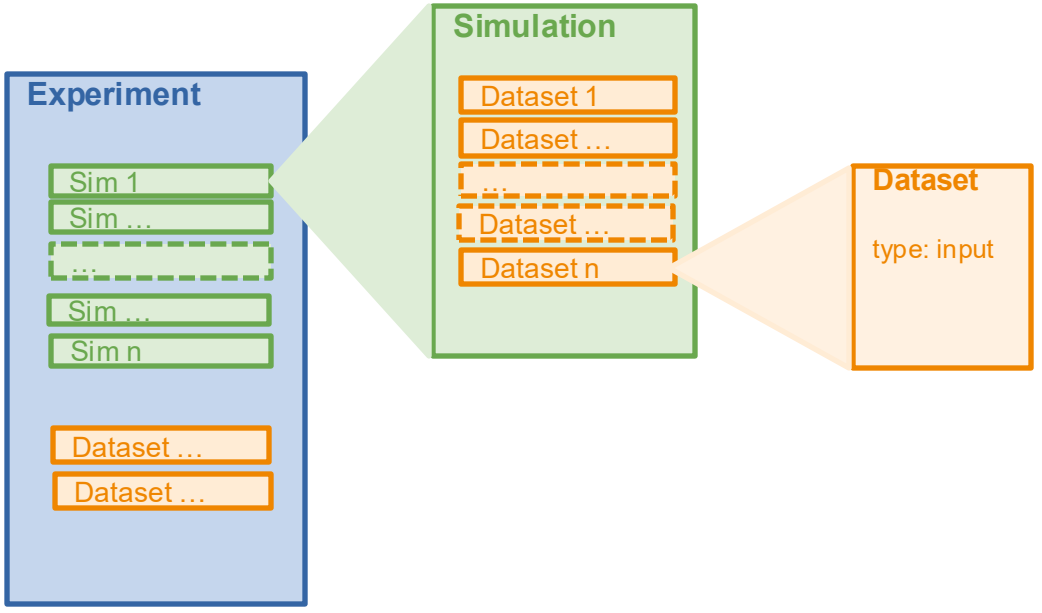


Data Model

- Main Entities:
  - Experiment
  - Simulation
  - Workflow
  - Dataset

- A Dataset can be of the following types:
  - Input
  - Output
  - Data product
  - Input parameter

- A Simulation is the execution of a Workflow and contains input data and the result of that specific workflow run.



Mapping the data/metadata to the SDL data model and v.v.

Input metadata in **json** format with **schema** described at the following link:

https://sdl-dev.hpc.cineca.it/api/bulk-import/schema?ajvStrict=true

Or using a RO-Crate metadata file

The metadata file is output in:

- JSON format
- RO-Crate format to improve fairness aspects especially interoperability

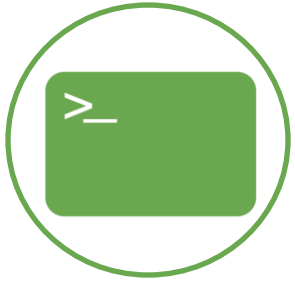
Additional functionalities

- Experiment publication: Restricted, Open access, Embargoed
- DOI minting and management
- Integration with EPOS Data Portal
- OGC Services
- Uploading and downloading both data and metadata (RO-Crate,…)
- Workflow management (CWL files and diagram visualisation)
- Dashboard of KPIs: automated FAIR assessment with **F-UJI tool**
- Data proximity and integration with CINECA HPC systems
- Advanced search functionalities including search by spatial and temporal coverage

How to access the SDL

WEB PORTAL

- User friendly interface
- Simple access to data
- For non-technical users



Command Line Interface

- For large quantities of experiments, simulations and datasets
- For technical users

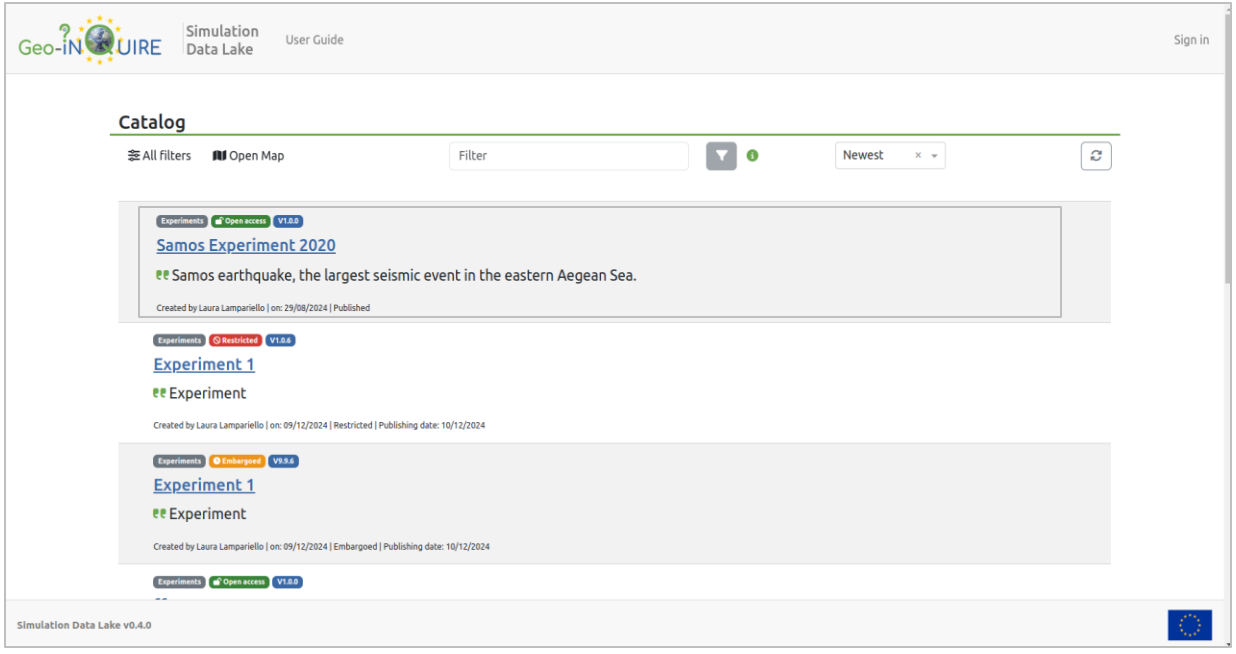
API

- Interact programmatically
- Integration with other services
- For technical users

2. Using the SDL: Web Portal

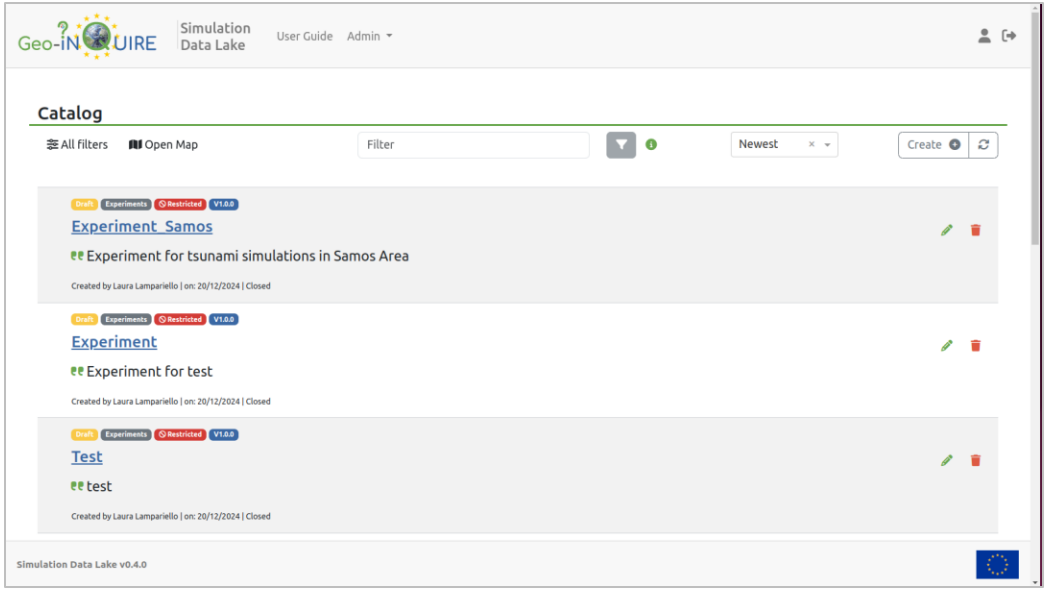
Web portal: Public Catalog

- Everyone can access
- It is possible to see the list of:
  - Experiments
  - Simulations
  - Datasets
- Only published experiments are visible
- The experiment can be published with some policies:
  - Restricted access
  - Embargoed
  - Public access
- Available only limited features



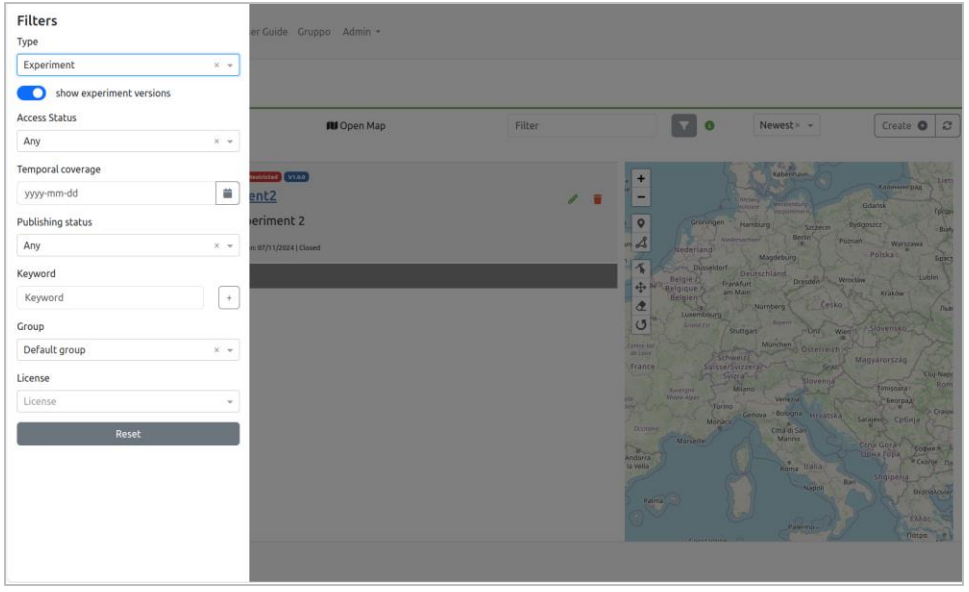
Web portal: Catalog

- Logged users can see also non published experiments
- Unpublished experiments can be seen by:
  - the creator
  - users who are part of the same group as the creator
  - the collaborators
- The catalog is paginated, the list of elements is divided over several pages

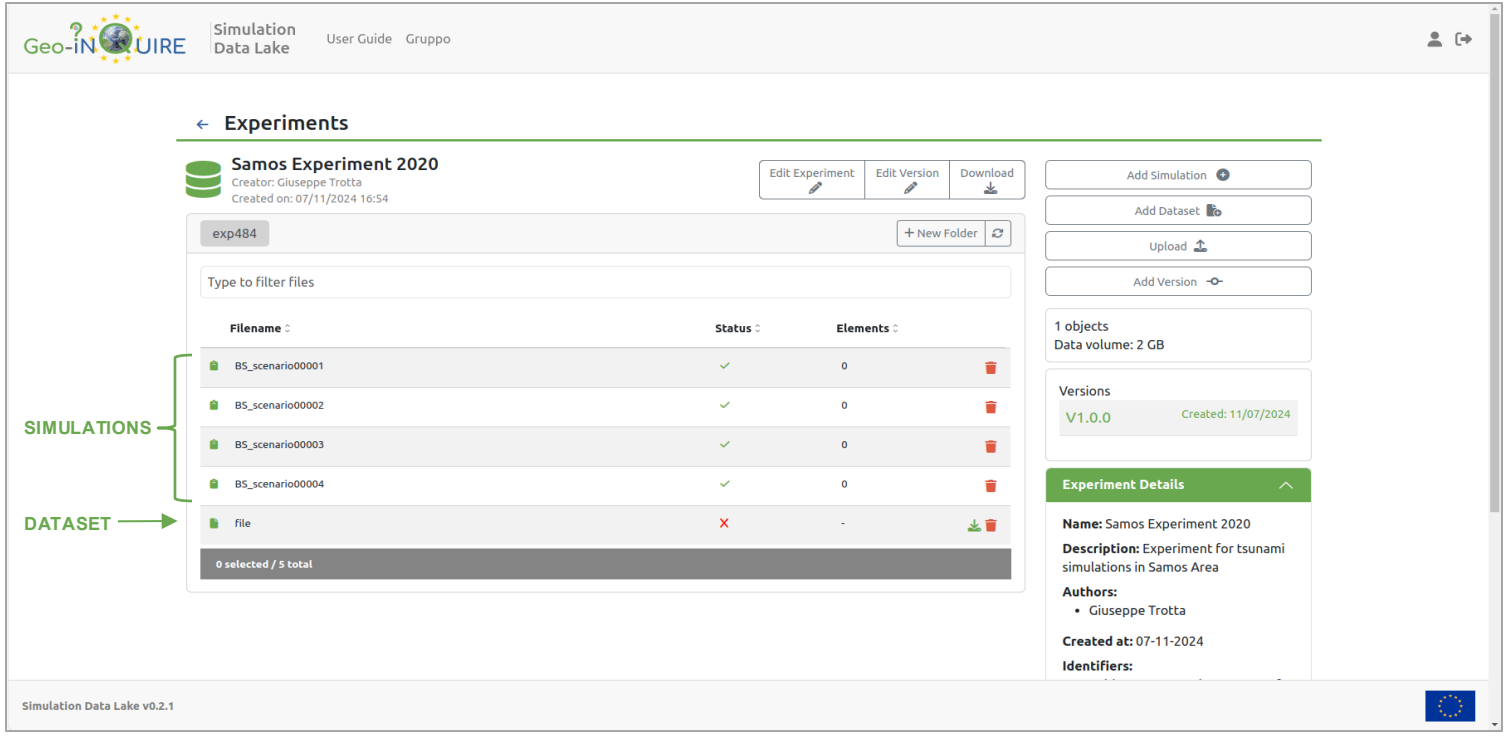


Web portal: Catalog Filters

- Filtering by **data type**: Experiment, Simulation, Dataset
- Show/hide **versions of experiments**
- Filtering by **Access Status** (Open Access, Restricted, Embargoed, Any), **Temporal coverage**, **Keywords** and **License** type
- Only logged users can search by:
  - **Publishing status** (Published, Draft, Any): whether an experiment has been published or not
  - **Group**: group with which the entity is associated



Web portal: Experiment view



2. Using the SDL: Command Line Interface

Command Line Interface (CLI): Getting started

- SDLCTL (Simulation Data Lake ConTroL)
- You can install it using two methods:
  - a remote script
  - in a virtual enviroment
- Installing the CLI you will also install a Python 3 module (SDK) to perform CLI operations from within Python code
- You can find the details of the steps to follow for installation at the following documentation link:

https://sdlctl.readthedocs.io/latest/getting_started/installation/

Command Line Interface (CLI): Commands

Look for experiments

```
(.venv) gtrotta@NGTROTTA0205932:~/Projects/geo-inquire/sdlctl$ sdlctl experiment list
                                    Experiments
```

ID	Name	Description	Created At
44	Cinzia test	Cinzia experiment for metadata ONLY: pls do NOT load any content	2024-05-03 14:16:45
30	Tsuanmi GC12	An example experiment.	2024-04-11 10:23:32
26	Lucia	Tsunami silumations	2024-04-05 14:52:31
23	Davide's Experiment	Testing upload service	2024-03-04 13:47:09
21	New CLI Experiment	This is another stupid test.	2024-02-29 18:04:03
12	Another Experiment	A test experiment after some refactoring.	2024-02-28 09:48:47
11	My second CLI experiment	This is another experiment created with the CLI in a more fancy way.	2024-02-27 18:10:41
9	My Experiment	Just a simple test.	2024-02-27 13:19:41
7	test	My first test experiment	2024-02-01 22:14:37

```
                                    Press 'q' di quit.
```

Delete confirmation

```
 % sdlctl file upload                                          ~ NMPUCCINI205808
Experiment ID: 12
Filename: sample50M
  Calculating MD5... ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 0:00:00

Starting new upload of file: sample50M

  Calculating MD5... ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 0:00:00
  Uploading...       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 0:00:00

Upload completed successfully!

 % sdlctl file delete                                          ~ NMPUCCINI205808
Experiment ID: 12
Filename: sample50M
Are you sure you want to delete the file? [y/N]:
```

Upload resume

```
 % sdlctl file upload -id 12 -f sample50M                      ~ NMPUCCINI205808
  Calculating MD5... ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 0:00:00

Starting new upload of file: sample50M

  Calculating MD5... ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 0:00:00
  Uploading...       ╸━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━  30% 0:00:05^C
Upload interrupted by user.
  Calculating MD5... ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 0:00:00
  Uploading...       ╸━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━  40% 0:00:04
 % sdlctl file upload -id 12 -f sample50M                      ~ NMPUCCINI205808
  Calculating MD5... ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 0:00:00

Resuming upload of file: sample50M

  Calculating MD5... ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 0:00:00
  Uploading...       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 0:00:00

Upload completed successfully!

 %                                                             ~ NMPUCCINI205808
```

See the following documentation for the complete list of commands: https://sdlctl.readthedocs.io/latest/reference/commands/

Wizard

- Help researchers generate the metadata file
- Through a guided procedure, it automatically generate an experiment metadata JSON file by analysing the structure of an input folder.
- It maps folder hierarchy to metadata, detects patterns in files and folders to minimize repetitive input, and guides users through the classification process.
- It is intended to be used with high item count for better results.
- Beta version
- Documentation:
  - https://sdlctl.readthedocs.io/latest/getting_started/wizard/
  - https://sdlctl.readthedocs.io/latest/tutorials/5-wizard/

3. Using the SDL: API

API specs

OpenAPI Specification
https://sdl.hpc.cineca.it/api/docs

upload

- POST `/api/experiments/{experiment_id}/init-upload/{filename}` init upload of a file
- POST `/api/experiments/{experiment_id}/upload` upload file
- POST `/api/experiments/{experiment_id}/complete-upload/{upload_id}/{key}` complete upload

download

- POST `/api/experiments/{experiment_id}/download/{filename}` download a file
- GET `/api/experiments/{experiment_id}/init-download/{filename}` Initialize download of a file

experiment

- POST `/api/experiments` Create a new experiment
- GET `/api/experiments` Get experiments
- GET `/api/experiments/{experiment_id}` Get an experiment by id
- DELETE `/api/experiments/{experiment_id}` Delete an experiment
- PATCH `/api/experiments/{experiment_id}` Update experiment data
- DELETE `/api/experiments/{experiment_id}/files` Delete files whose name begin with a certain prefix in a given experiment (if path ends with '/' it will look for a directory to delete, otherwise, a file)
- GET `/api/experiments/{experiment_id}/files` Get list of files for an experiment
- POST `/api/experiments/{experiment_id}/collaborators` Add a collaborator to the experiment

simulation

- GET `/api/experiments/{experiment_id}/simulations` Get all simulations of an experiment
- POST `/api/experiments/{experiment_id}/simulations` Create a new simulation
- PATCH `/api/experiments/{experiment_id}/simulations` Modify a new simulation
- POST `/api/experiments/{experiment_id}/simulations/{folder_name}` create a simulation folder or a simple folder

3. SDL FAIR support

Improving FAIRness in the SDL

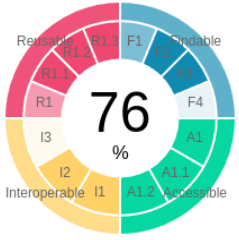
- Automated periodic FAIR assessment
  - Ensures continuous monitoring and improvement
  - F-UJI tool used to assess experiments FAIRness

- Incremental improvements:
  - Agile, step-by-step enhancements to FAIR compliance

SDL FAIR details

- The first FAIR assessments using FUJI gave us a 5% score

- The result of a FAIR assessment performed recently on an SDL experiment:



	Score earned:	Fair level:
Findable:	4 of 7	moderate
Accessible:	7 of 7	advanced
Interoperable:	4 of 6	moderate
Reusable:	5 of 6	moderate

SDL FAIR details / Findable

- Unique identifiers:
  - UUID
- Persistent identifiers:
  - DOI (minting will be available soon)
- Metadata contains descriptive data to support findability:
  - creator, title, data identifier, publisher, publication date, summary, keywords...
- EPOS Data Portal integration

SDL FAIR details / Accessible

- Data access levels specified in metadata
- Metadata and data are retrievable by their identifier
- Standardized communication protocol used to access data and metadata

SDL FAIR details / Interoperable

- Metadata is represented using a formal knowledge representation language
  - RO-Crate

Why RO-Crate?

- Lightweight, machine-readable metadata format
- Facilitates interoperability and reusability
- Widely adopted in research data management
- Allows for CWL workflows representations using profiles such as Workflow RO-Crate

SDL FAIR details / Reusable

- Metadata specifies the content of the data
- Metadata includes license information under which data can be reused
- Metadata includes provenance information about data creation or generation

SDL FAIR details / User perspective

Findable:

- Searchable metadata catalog: Users can easily discover datasets using keywords, text search, or other filters.
- SDL catalog explorable using:
  - Web application
  - CLI
  - Python module
  - REST API
  - EPOS Data Platform (coming soon)
- DOI assignment (coming soon): Enables citation and persistent access to datasets.

SDL FAIR details / User perspective

Accessible:

- Clear access policies
- Standardized resource download mechanism: makes it easy to retrieve data programmatically or manually

SDL FAIR details / User perspective

Interoperable:

- Standard/well known formats: users can work with data using familiar tools
- RO-Crate packaging: ensures metadata and data are bundled in a machine-readable, structured way

SDL FAIR details / User perspective

Reusable:

- Rich metadata: users understand the context, origin, and structure of the data
- Clear licensing: users know how they can use the data

Useful information

- Dev: https://sdl-dev.hpc.cineca.it
  - Development environment with 1T of storage
  - It is the most updated enviroment with features in development
- Prod: https://sdl.hpc.cineca.it
  - Production environment with 500T of storage
- Contacts: sdl@cineca.it

- Request an SDL account using the following link:
  https://sdl-userguide.readthedocs.io/tutorials/Introduction/
- Documentation
  - Web Portal: https://sdl.hpc.cineca.it/docs
  - Command Line Interface: https://sdl.hpc.cineca.it/cli/docs

Thank you for your attention

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.