ClimERI-France Data Analysis Platform

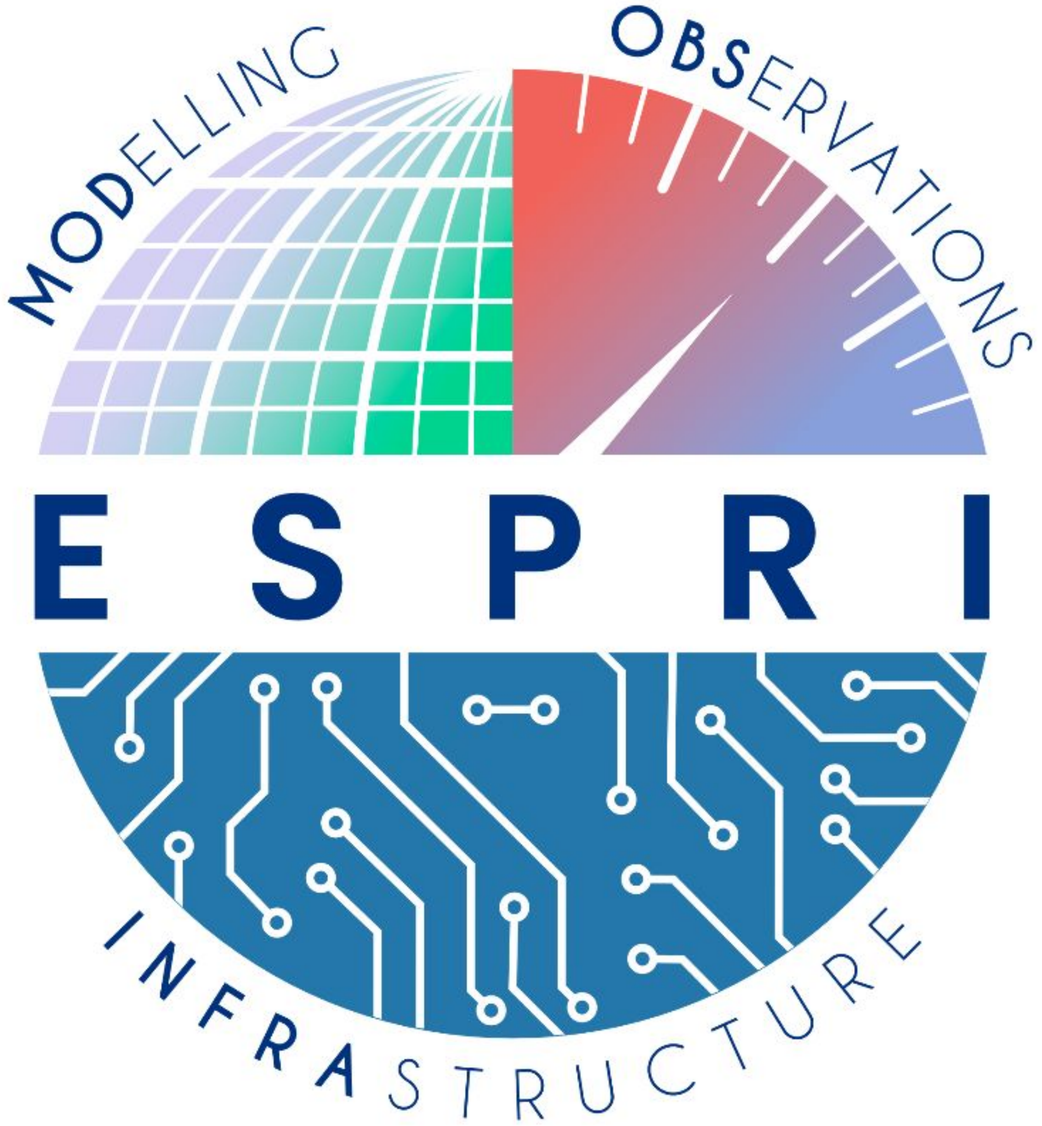
Levavasseur G. and Ramage K.

What's "ESPRI"?

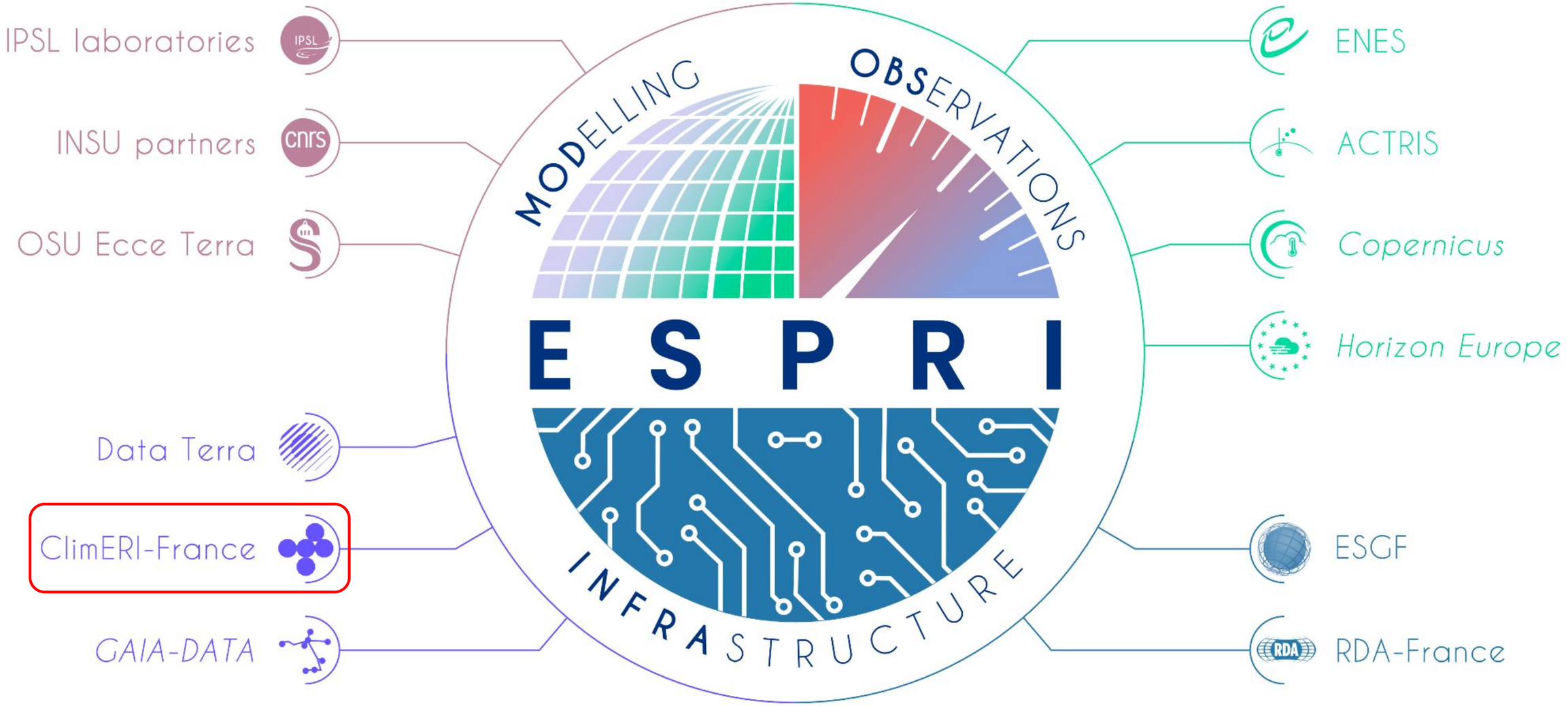
The Institut Pierre-Simon Laplace provides its laboratories with coordination resources and services to help develop major projects and disseminate the results.

The ***IPSL Computing and Data Centre*** was born out from several initiatives at the IPSL and its laboratories, wishing to share IT resources and joint projects through a numerical facility for research.

For more than 20 years, the IPSL computing and data centre has been providing **E**nsemble of **S**ervices for (*tr.* "**P**our") **R**esearch at the **I**PSL - ***ESPRI***.



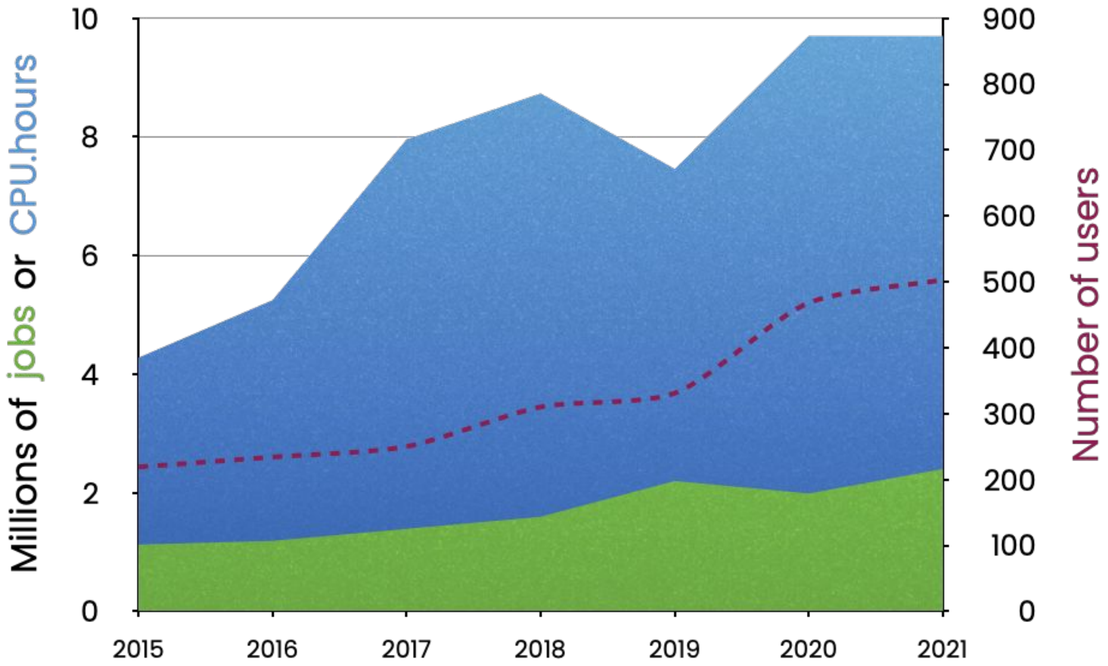
Community infrastructure

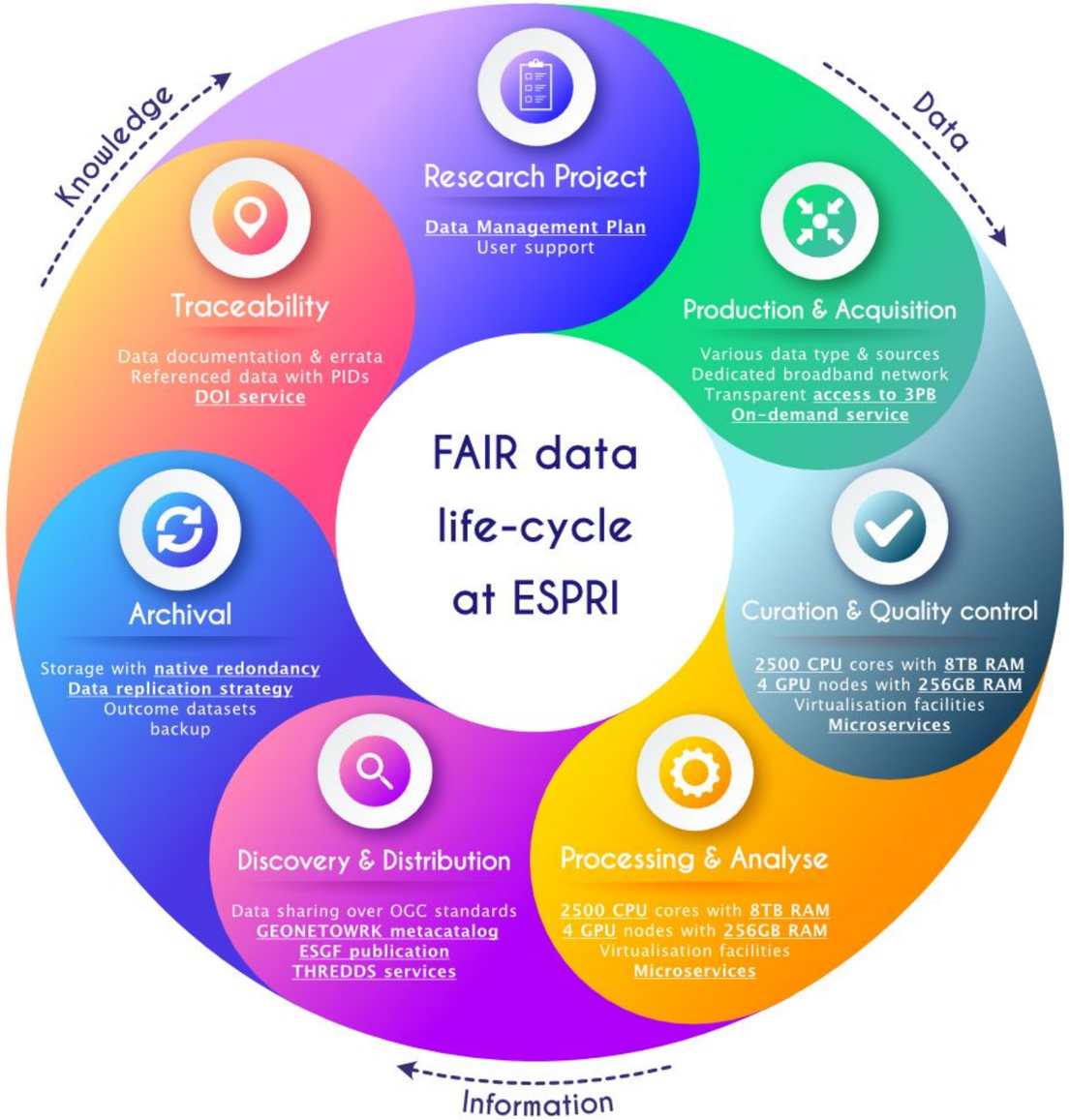


Levels: Regional National European International

Mission & key figures

ESPRI aims to provide a distributed numerical infrastructure for research at IPSL with “FAIR” (**F**indability, **A**ccessibility, **I**nteroperability and **R**eusability) data services.

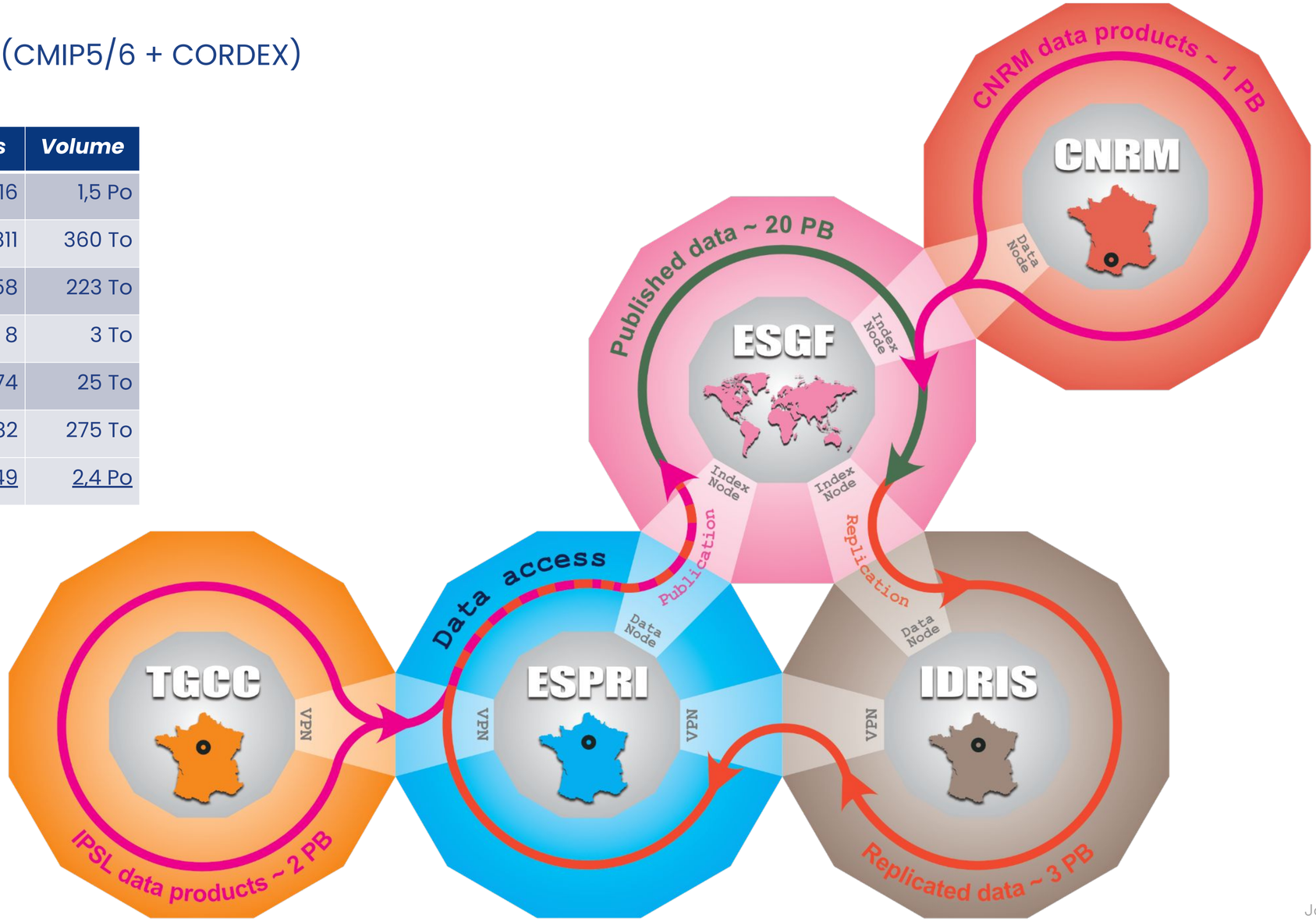




ESPRI centralizes ClimERI-France data services

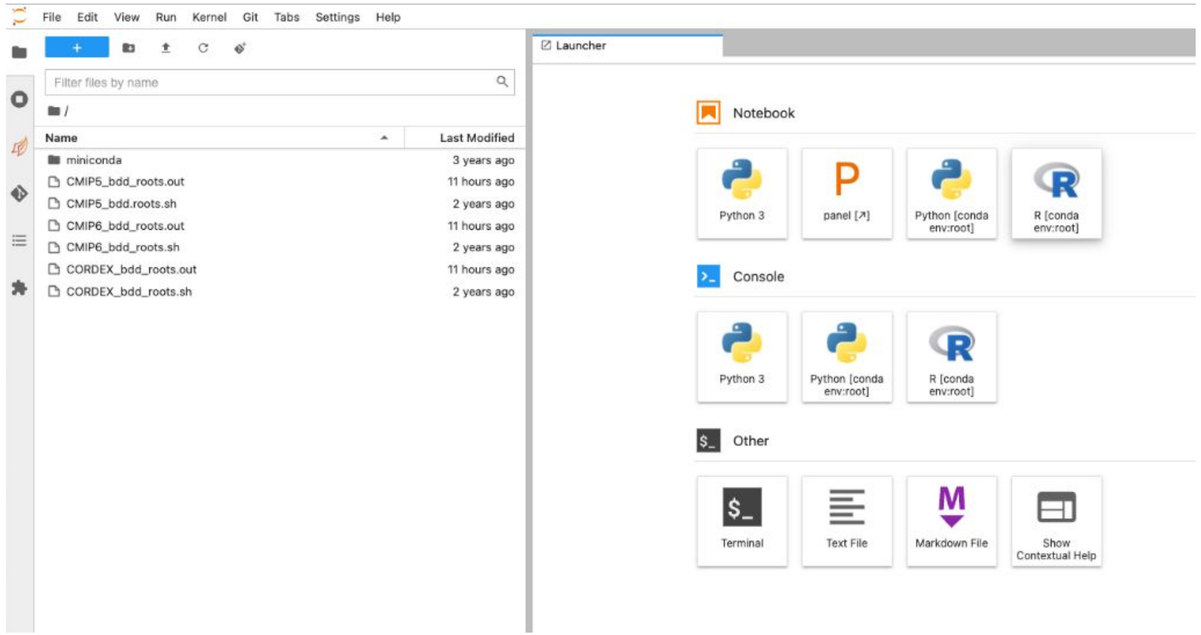
- All IPSL-CM production (CMIP5/6 + CORDEX)
- Replica pools:

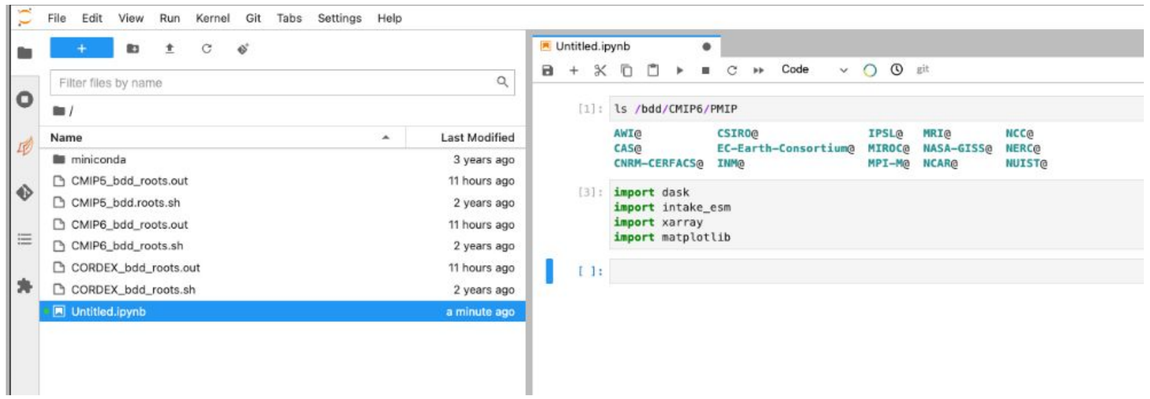
Project	Thousands of files	Volume
CMIP6	4 916	1,5 Po
CMIP5	811	360 To
CORDEX	458	223 To
input4MIPs	8	3 To
C3S-CMIP5	174	25 To
C3S-CORDEX	582	275 To
Total	6 949	2,4 Po



Towards Python Notebook interface

New





Access: https://data.ipsl.fr/jupyter

Status: Pre-production

Users: ~20 “power-users” with extensive usages + Proven with training session.

Documentation: in progress to be available through the new incoming ESPRI website.

Python environment: PANGEO + useful modules.

Service opening: March 2022

GAIA-Data requirements:

- Data catalogs access (including ClimERI ones)
- Online editors (VSCode, R, Matlab)
- Dashboards (Dask, Tensor)

Improved local data discovery

New

Old-fashion discovery

```
# Les données de "référence"

$> ls /bdd/CMIP3
$> ls /bdd/CMIP5
$> ls /bdd/CMIP6
$> ls /bdd/CORDEX
$> ls /bdd/obs4MIPs
$> ls /bdd/input4MIPs
$> ls /bdd/CMIP5-Adjust
$> ls /bdd/CORDEX-Adjust
$> ls /bdd/C3S-CMIP5
$> ls /bdd/C3S-CORDEX
```

Discovery through "intake-esm" catalogs

```
$> python3
>>> import intake

>>> cat = intake.open_esm_datastore('/modfs/catalog/CMIP6.json') # Takes up to a minute to load the catalog
>>> cat
<CMIP6 catalog with 11352 dataset(s) from 6868559 asset(s)>
>>> cat.df.columns
Index(['path', 'project', 'activity_id', 'institution_id', 'source_id', ..., 'latest'], dtype='object')
>>> cat.unique(columns='table_id')
{'table_id': {'count': 39, 'values': ['Emon', 'Amon', 'CFmon', 'day', ..., 'Eday']}}

>>> subcat = cat.search(experiment_id='ssp585',
                       institution_id=['IPSL','CNRM-CERFACS'],
                       variable_id=['pr', 'tas'],
                       table_id='Amon',
                       latest=True)
>>> subcat
<CMIP6 catalog with 4 dataset(s) from 40 asset(s)>

>>> dsets = subcat.to_dataset_dict()
>>> dsets.keys()
dict_keys(['ScenarioMIP.IPSL.IPSL-CM6A-LR.ssp585.Amon.gr', 'ScenarioMIP.CNRM-CERFACS.CNRM-CM6-1.ssp585.Amon.gr',
...])

>>> dsets['ScenarioMIP.IPSL.IPSL-CM6A-LR.ssp585.Amon.gr']
<xarray.Dataset>
Dimensions:      (time: 3432, lat: 143, lon: 144, axis_nbounds: 2, member_id: 7)
[...]
Data variables:
[...]
    pr           (member_id, time, lat, lon) float32 dask.array<chunksize=(1, 1032, 143, 144), meta=np.ndarray>
    tas          (member_id, time, lat, lon) float32 dask.array<chunksize=(1, 1032, 143, 144), meta=np.ndarray>
Attributes: (12/50)
[...]
    tracking_id:            hdl:21.14100/f09be139-ce58-4719-8fe9-2769aad503c...
```

Access:
`/modfs/catalog`

JupyterHub

Supported projects:

CMIP5/6

CORDEX

(other coming soon).

Quality control:
CV checked

Update:

Weekly

Service opening:
February 2022

Improved data replication scheduling

New

ESGF NODES

SDT retrives the certificates and builds the HTTP requests to Solr corresponding to the search criteria.

REQUEST

Search criteria called facets are used to select which files to download. They can be set on command line or using a template.

SYNDA
Transfer

A command line tool to discover and download files from the Earth System Grid Federation (ESGF) archive.

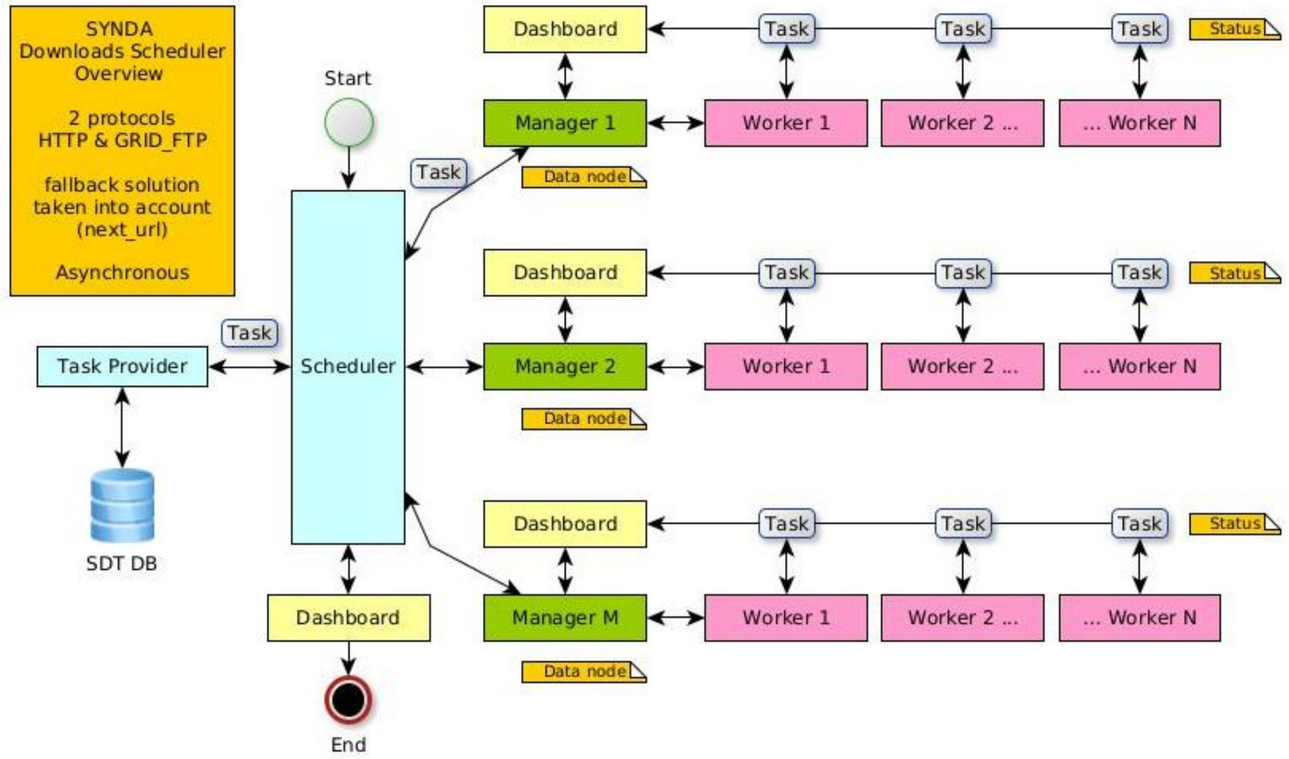
FILESYSTEM

ESGF files are downloaded using the HTTP or GridFTP protocol and managed on the local filesystem following the Data Reference Syntax.

SDT DATABASE

A SQLite database records each downloaded file and dataset. A complete dataset triggers a "dataset_complete" event, which informs the SDP module to start the pipeline.

New paradigm implemented to perform **parallel downloads asynchronously**. Once each workers have their download tasks given out by the scheduler, they are able to asynchronously carry out their duties without having to wait for each other.



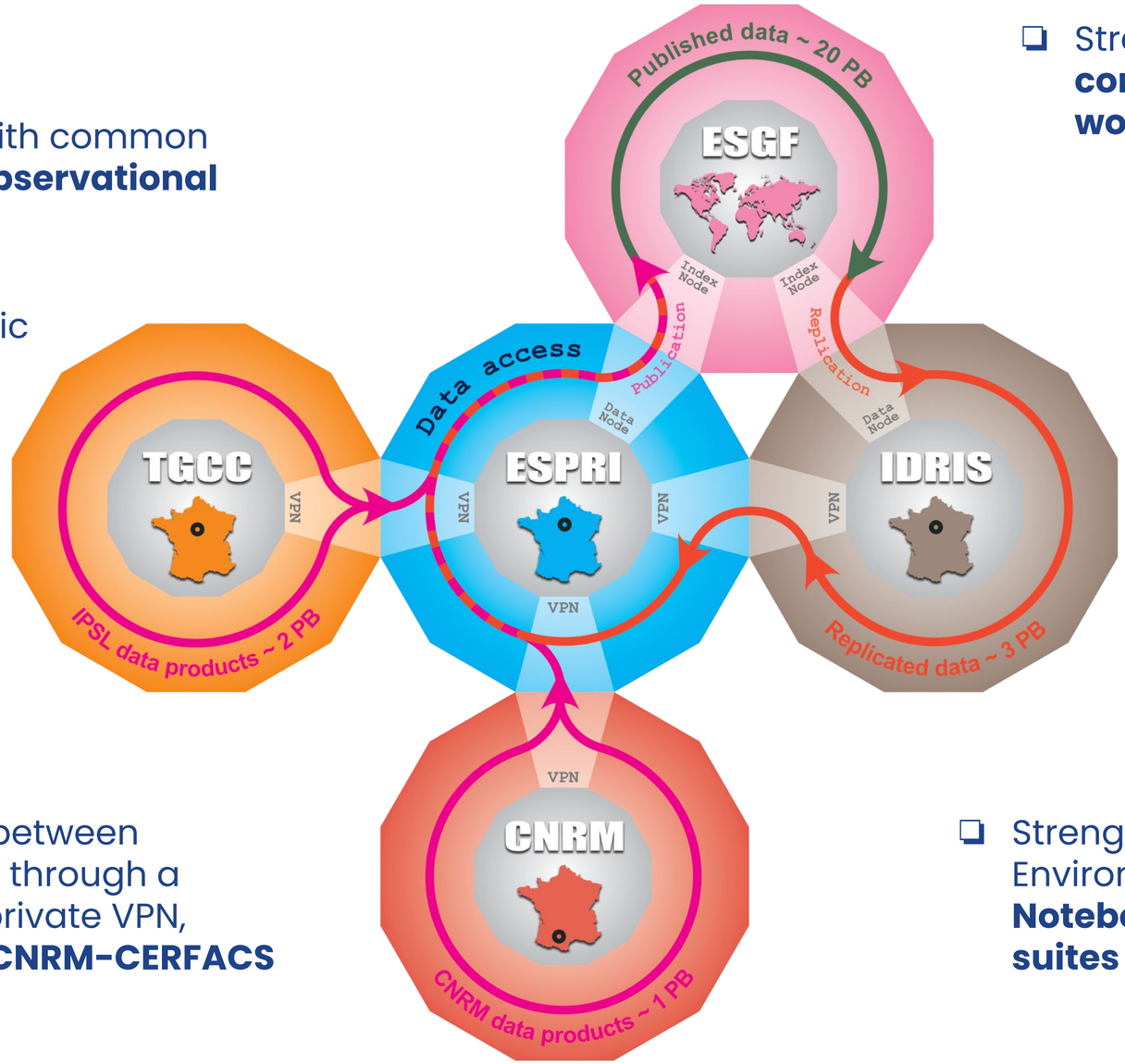
ESPRI awarded

This year, the ESPRI team was awarded by the **Crystal collective prize from CNRS**.

Thank you for your support and feedbacks !

2022 and beyond with GAIA-Data

- Towards **STAC catalogs** with common **search engine** between **observational and modelling** data sets.
- **Extending** a multi-thematic **storage at IDRIS**
- Extending and automating **DOI services to codes**.
- **Network mounting point** between **CNRS-CERFACS and ESPRI** through a **10Gbps** RENATER link and private VPN, allowing **direct access to CNRM-CERFACS climate simulations**.



- Strengthening **on-demand bias correction** processing based on **workflow managers**.
- Getting the ***CoreTrustSeal* label** from the Reasearch Data Alliance (RDA).
- Strengthening Virtual Research Environment based on **Jupyter Notebooks with PANGEO and AI4GEO suites**

2022 and beyond with Copernicus

The Copernicus **Climate Data Store** (CDS) supplies access to a **subset** of **global** and **regional climate projections** from CMIP5/6 and CORDEX exercices (~300TB).

This service relies on **dedicated** and **load-balanced** servers deployed by **ENES partners** which provides a single **resilient point of access** to data delivered through replication and **redundancy** (~500 downloads/day or ~300GB/day).

ESPRI will **coordinate** the **operational maintenance** of the infrastructure for **the next 4 years**.

Thank you for your attention

Institut Pierre-Simon Laplace