French Data Analysis Platform for ENES (aka. ESPRI)

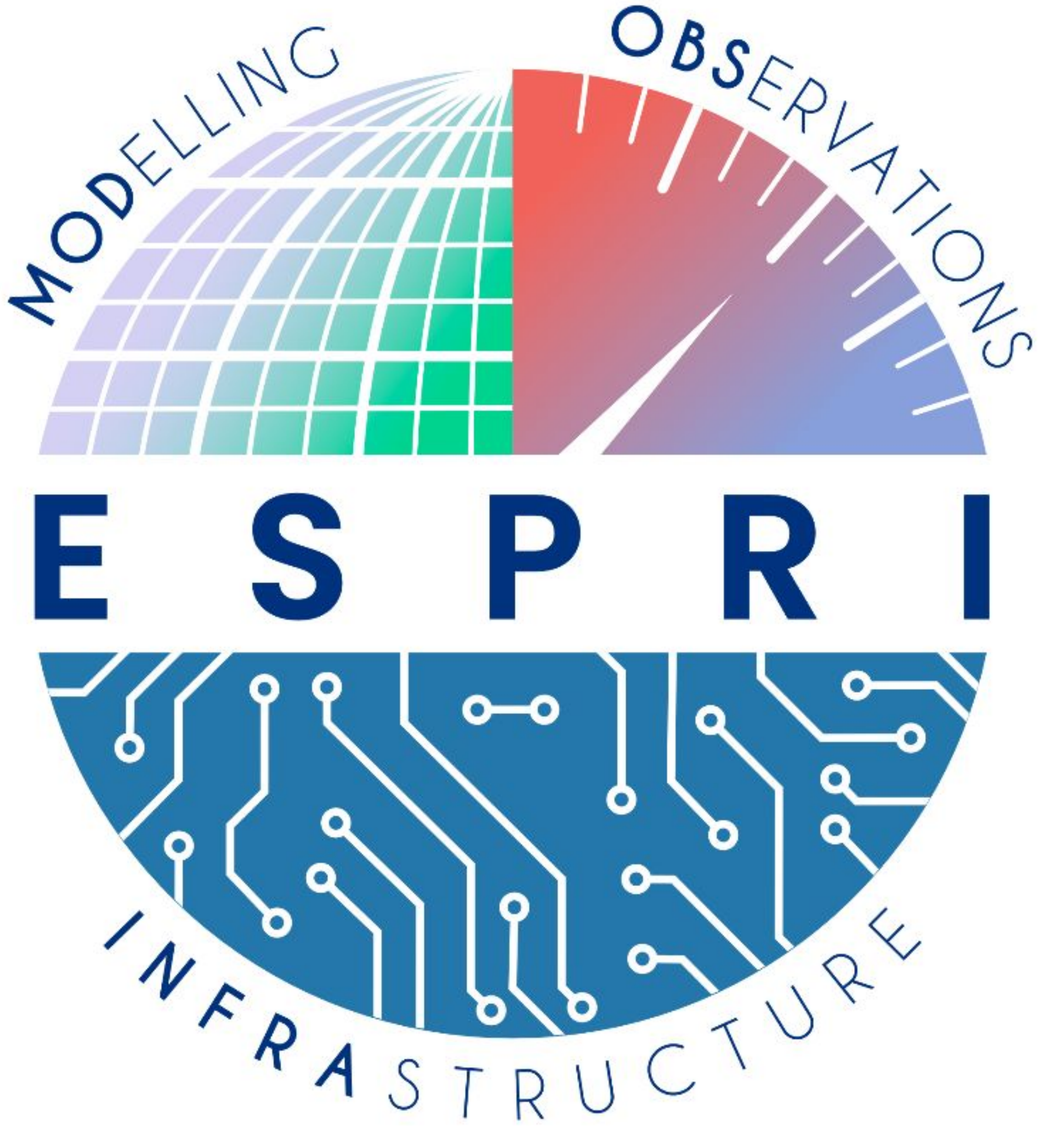
Levavasseur G.

What's "ESPRI"?

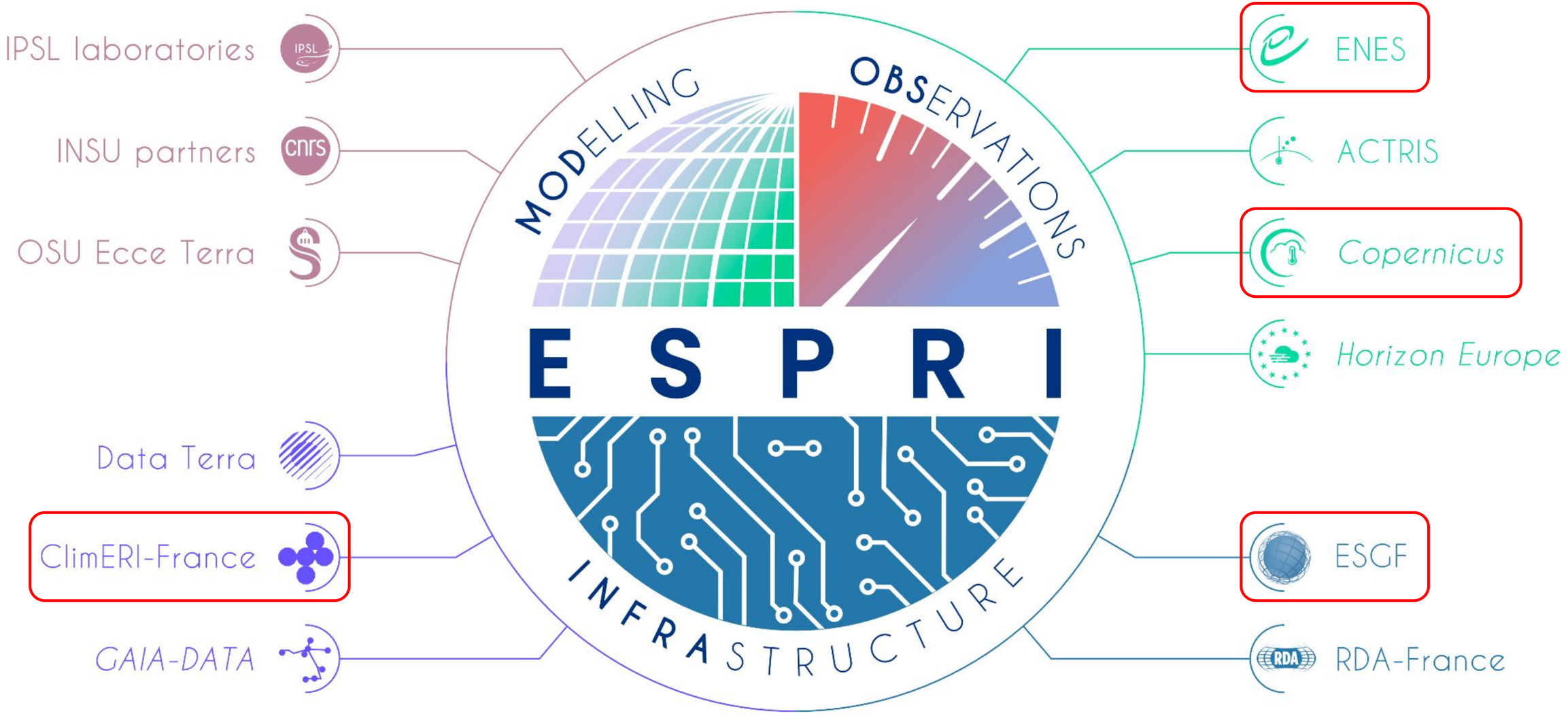
The Institut Pierre-Simon Laplace provides its laboratories with coordination resources and services to help develop major projects and disseminate the results.

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

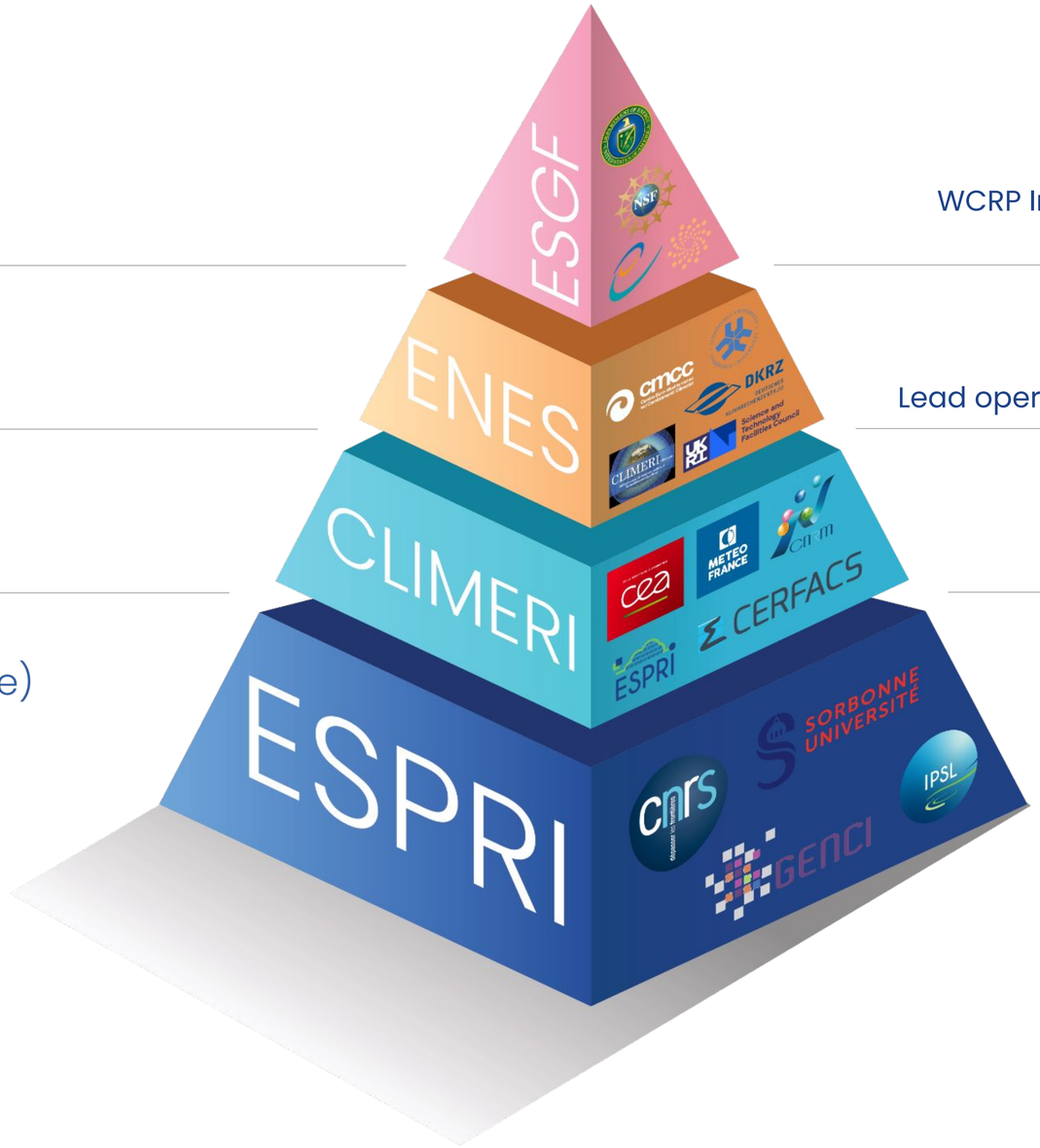
ESPRI-Mod community infrastructure

International

European

National (France)

Regional (Ile-de-France)



Largest contributor (> 1,2M datasets ≈ 1.7PB)
Tier 1 ESGF server
WCRP Infrastructure Panel contributor (Task Team co-leader)
ES-DOC governance member (Principal Investigator)

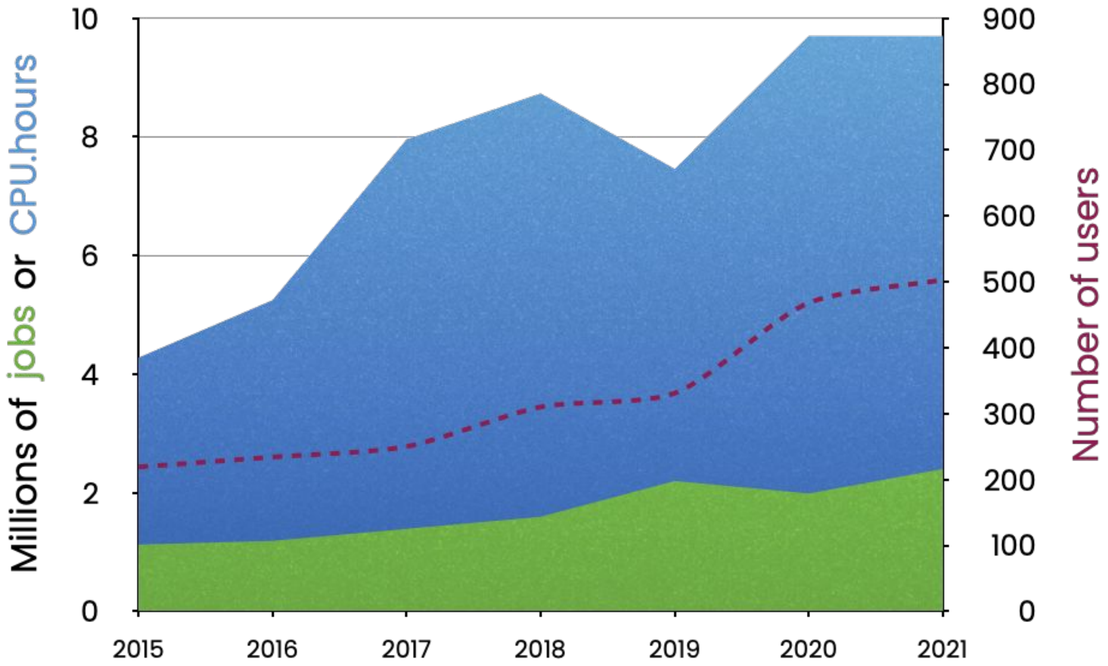
Coordination of the ENES infrastructure
Feeding national needs into large-scale projects
Maintain and consolidate the ESGF
Lead operator of the climate data archive for the C3S and EOSC

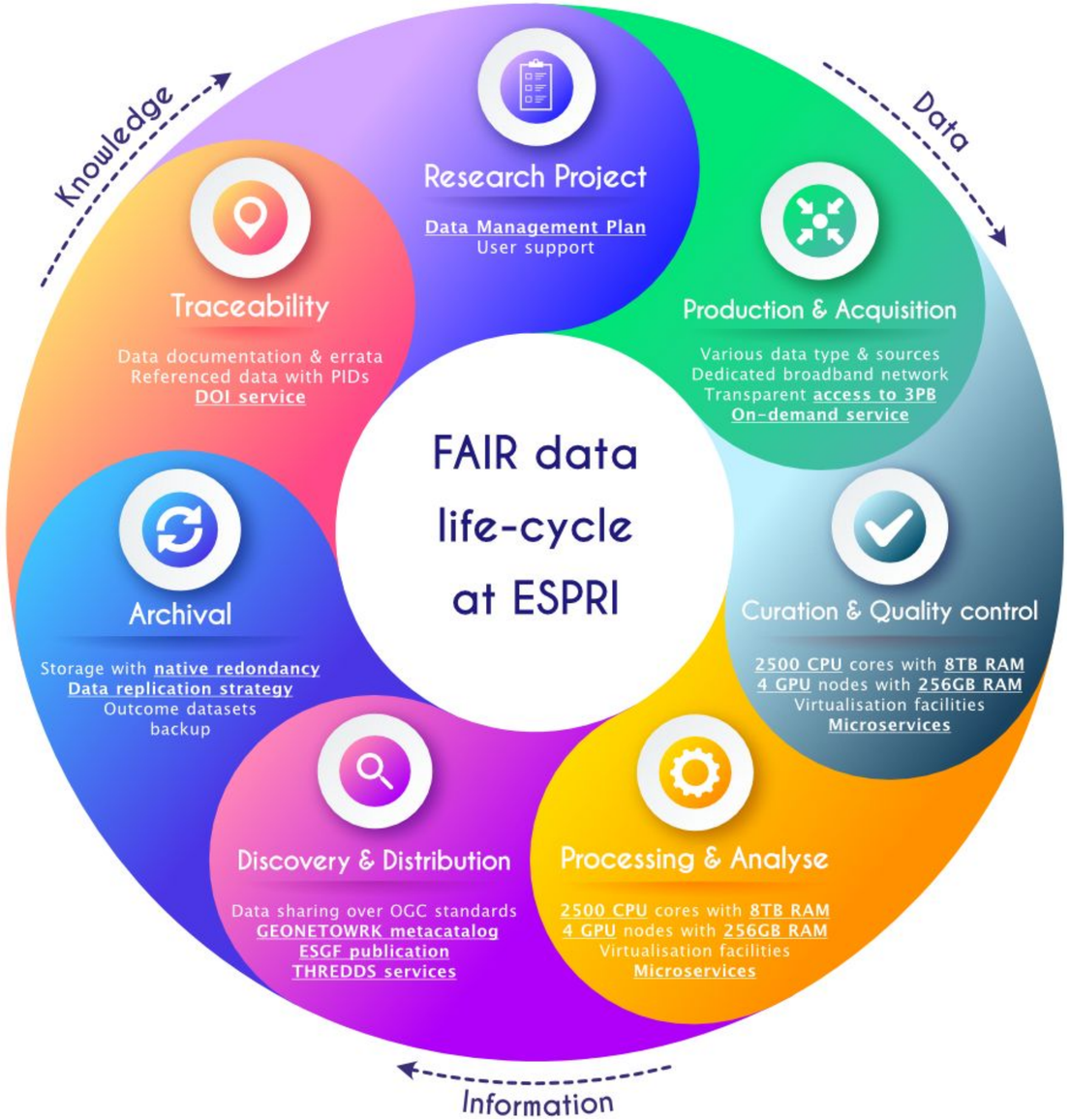
Supporting the French community
Coordination with HPC partners
Replication & multi-model caching
Response to climate services

Distribution of French climate simulations
Analysis & software environments
Processing workflow (bias correction, etc.)
Documentation & traceability of simulations

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.





Resources access

Free access

- IPSL community
- IPSL partners from national IR
- Non-EU partners through ENES VA calls

Other access:

- Project funded access
- Quote on request according to the ESPRI pricing

Registration & recommended contact to get started

- http://mesocentre.ipsl.fr
- Guillaume.Levavasseur@ipsl.fr

Login through SSH

Expertise and support:

- Data analysis (discovery, access, dissemination, etc.)
- ML/DL (through ESPRI-IA)
- Bias correction

ESPRI services

Research environment

- Computing
  - 2 500 CPU cores with 8TB shared RAM
  - 36 GPU cores
- Data archives
  - ~3Po of various climate data
  - Dedicated high bandwidth access
- Storage: 20GB on `/home` + 1TB on `/data`
- Compilers: GNU, Intel, etc.
- Librairies: HDF5/netCDF4/openMPI, etc.
- Products: Python2/3, CDO/NCO, Ferret/GraDs, Matlab, ncl, R, SciLab, Git/SVN
- Shared python environments: ClimAF, “PANGEO-like”

On-demand data sharing

- Group shared space
- THREDDS shared space with OGC protocols
  - http://thredds-su.ipsl.fr
- GridFTP
- DOI/PIDs

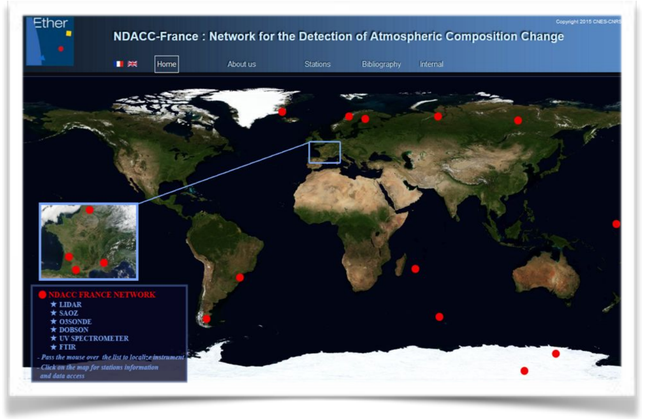
SPIRIT cluster - Sorbonne Université (Paris)

SPIRITX cluster - Polytechnique (Palaiseau)

Let’s cross the bytes

Earth observations

- Soil & in-situ data
  - Campaigns (CAL/VAL MT, balloons, etc.)
  - Airborne compounds from 17 stations
  - SIRTA measurements
  - Radio soundings (ARSA, TIGR)
- Satellite
  - Level 1 to 4 products (POLDER, PARASOL, CFMIP-obs)
  - Numerical outputs (INDOEX, AMMA, HyMex ChArMEx)
- Reanalyses
  - ERA
  - MERCATOR-OCEAN
  - NCEP
  - FCDR (AMSI, SSMI, GridSat)
- Atmospheric model outputs
  - 50 airborne compounds from REPROBUS
  - Potential vorticity and temperature from MIMOSA

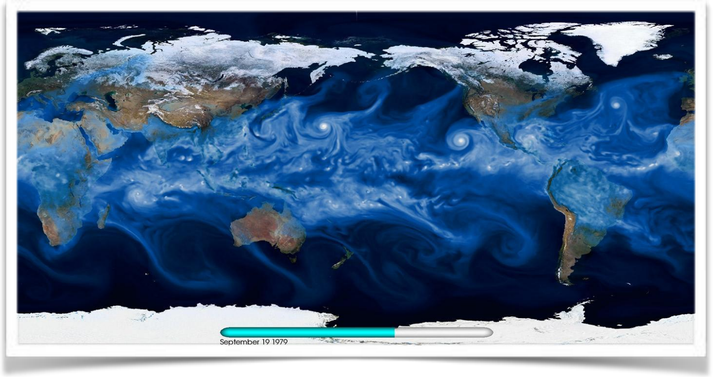


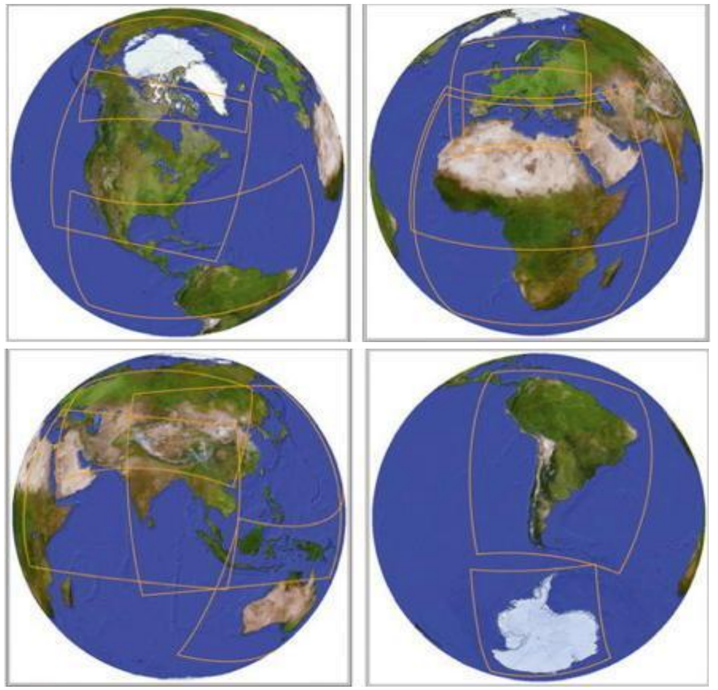
Satellite data
IASI level 1C (METOP-A-B)
IASI level 2 (O3, CO, SO2, CH4, HCOOH, NH3)
AMSUA-MHS-HIRS4 level 1C (METOP-A-B)
GOME2 level 1B (METOP-A-B)
GOME2 level 2 (METOP-A-B)
GOSAT level 1B / FTS/CAI
GOSAT level 2 / FTS/CAI
SAGE II, UARS, SPOT3, SPOT4, ODIN, ENVISAT

Let’s cross the bytes

Climate simulations

- Coupled Model Intercomparison Project (CMIP)
  - CMIP3, CMIP5, CMIP6
  - Tied projects (PMIP3, GeoMIP, etc.)
  - With bias correction
- Coordinated Regional Climate Downscaling Experiment (CORDEX)
  - Several geographical domains
  - Several RCMs
  - With bias correction
- Observations for Model Intercomparison Project (obs4MIP)
- Input Datasets for Model Intercomparison Project (input4MIP)
- Climate projections for Copernicus Climate Data Store (including IPCC Climate Atlas)



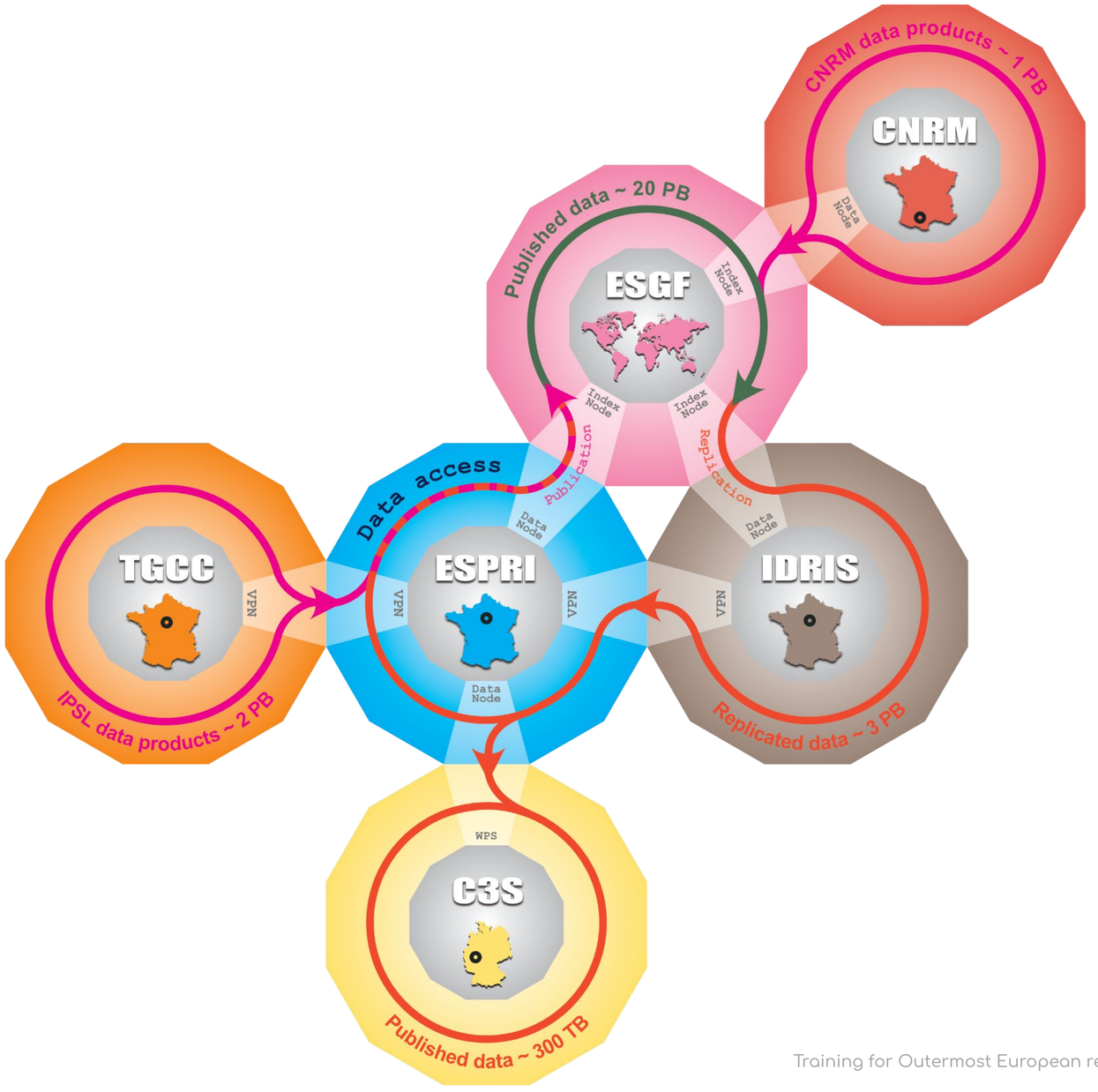


ESPRI centralizes ClimERI-France data archives

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

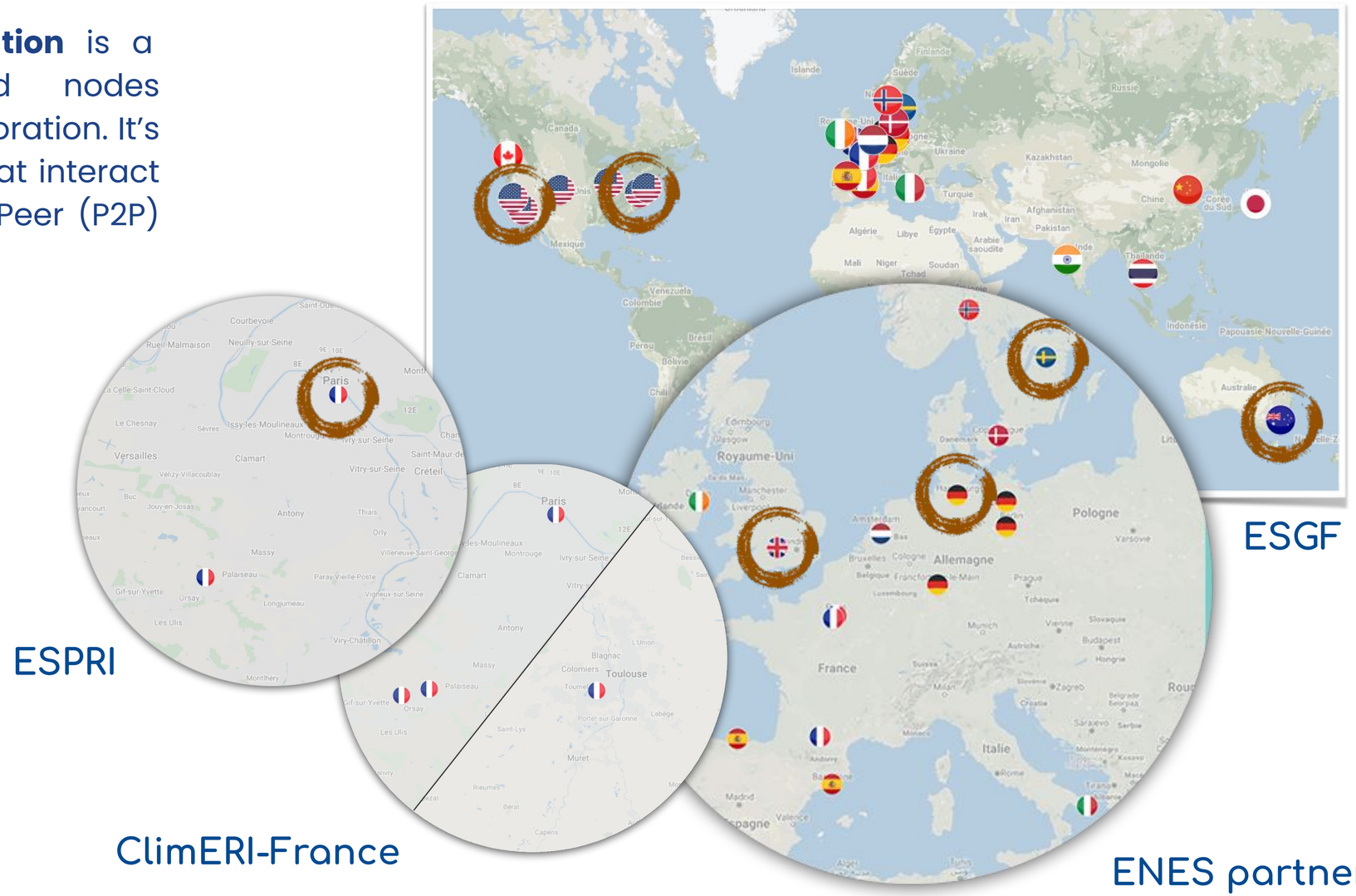
Replica pools:

Project	*Thousands of files*	*Volume*
CMIP6	4 916	1.5 PB
CMIP5	811	360 TB
CORDEX	458	223 TB
input4MIPs	8	3 TB
C3S-CMIP5	174	25 TB
C3S-CORDEX	582	275 TB
C3S-CMIP6	235	18 TB
C3S-IPCC-Atlas	*(coming soon)*	
Total	7 184	2,6 P



ESPRI hosts the French ESGF Tier 1 node

The **Earth System Grid Federation** is a decentralized and federated nodes network with international collaboration. It's a system of distributed nodes that interact dynamically through a Peer-To-Peer (P2P) paradigm.



ESPRI provides data services for CLIMERI-France and ENES

Local data discovery through single point of access: /bdd

CMIP6 directory structure

Pattern

```
root
└── mip_era
    └── activity_id
        └── institution_id
            └── source_id
                └── experiment_id
                    └── variant_label
                        └── table_id
                            └── variable_id
                                └── grid_label
                                    └── version
```

Example

```
/bdd
└── CMIP6
    └── CMIP
        └── IPSL
            └── IPSL-CM6A-LR
                └── historical
                    └── r1i1p1f1
                        └── Amon
                            └── tas
                                └── gr
                                    └── v20180803
```

```
# Referenced data

$> 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
$> ls /bdd/C3S-CMIP6
```

CMIP6 file naming

Pattern

`<variable>_<table_id>_<source_id>_<experiment_id>_<variant_label>_<grid_label>[_<period_start>-<period_end>].nc`

Example

`tas_Amon_IPSL-CM6A-LR_historical_r1i1p1f1_gr_185001-201412.nc`

Constraints

- Users must know the directory structure of data archives.
- Users must know the controlled vocabulary of data archives.

ESPRI provides data services for CLIMERI-France and ENES

Local data discovery through `intake-esm` catalogs

```
import intake
```

```
catMaster = intake.open_catalog("/modfs/catalogs/master.yml")
catMaster
```

```
master:
  args:
    path: /modfs/catalogs/master.yml
  description: ClimERI-France master catalog for all data pool catalogs available
  driver: intake.catalog.local.YAMLFileCatalog
  metadata: {}
```

```
subcats = list(catMaster.items())
subcats
```

```
[('CMIP6', <CMIP6 catalog with 13482 dataset(s) from 7089602 asset(s)>),
 ('CMIP5', <CMIP5 catalog with 3678 dataset(s) from 938400 asset(s)>),
 ('CORDEX', <CORDEX catalog with 1942 dataset(s) from 428146 asset(s)>)]
```

```
('CORDEX', <CORDEX catalog with 1942 dataset(s) from 428146 asset(s)>)]
```

ESPRI provides data services for CLIMERI-France and ENES

Local data discovery through `intake-esm` catalogs

```
catCMIP6 = catMaster["CMIP6"]
catCMIP6
```

CMIP6 catalog with 13482 dataset(s) from 7089602 asset(s):

	unique
path	7089602
project	1
activity_id	19
institution_id	40
source_id	107
experiment_id	237
member_id	738
table_id	39
variable_id	1034
grid_label	12
version	867
init_year	63
period_start	26094
period_end	32849
climatology	2
latest	2

ESPRI provides data services for CLIMERI-France and ENES

Local data discovery through `intake-esm` catalogs

```
mydata = catCMIP6.search(
    experiment_id=["historical", "ssp585"],
    table_id="Amon",
    variable_id="tas",
    latest=True)
mydata
```

CMIP6 catalog with 111 dataset(s) from 15882 asset(s):

	unique
path	15882
project	1
activity_id	2
institution_id	33
source_id	63
experiment_id	2
member_id	201
table_id	1
variable_id	1
grid_label	3
version	185
init_year	0
period_start	258

ESPRI provides data services for CLIMERI-France and ENES

Local data discovery through `intake-esm` catalogs

```
mydata.serialize("myCat")
```

```
Writing catalog with 15882 entries into: myCat.json
Writing ESM collection json file to: myCat.json
```

Main features

- User-friendly search interface
- Record your own catalog from the referenced ones
- Directory structure agnostic
- netCDF driver (with xarray)
- Aggregation over time, simulation members and initial year (decadal predictions)
- And more at https://intake-esm.readthedocs.io/

Constraints

- Users must know the controlled vocabulary of data archives.

ESPRI provides data services for CLIMERI-France and ENES

Interactive research environment (Shell based - default)

Famous gridded data analyzing tools

- NCO/CDO operators
- Ferret/GraDs
- Shared Python environments

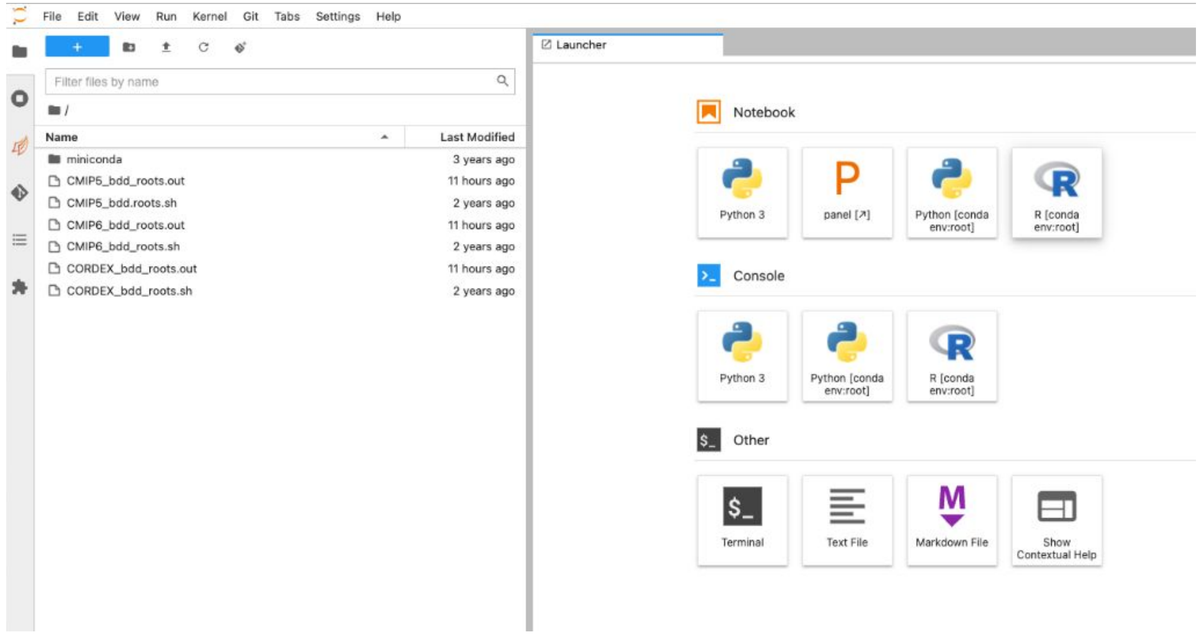
```
$> module load modtools-python2
$> module load modtools-python3
$> source activate analyse_3.6
# xarray, intake-esm, ipython-notebook, Dask, etc.
```

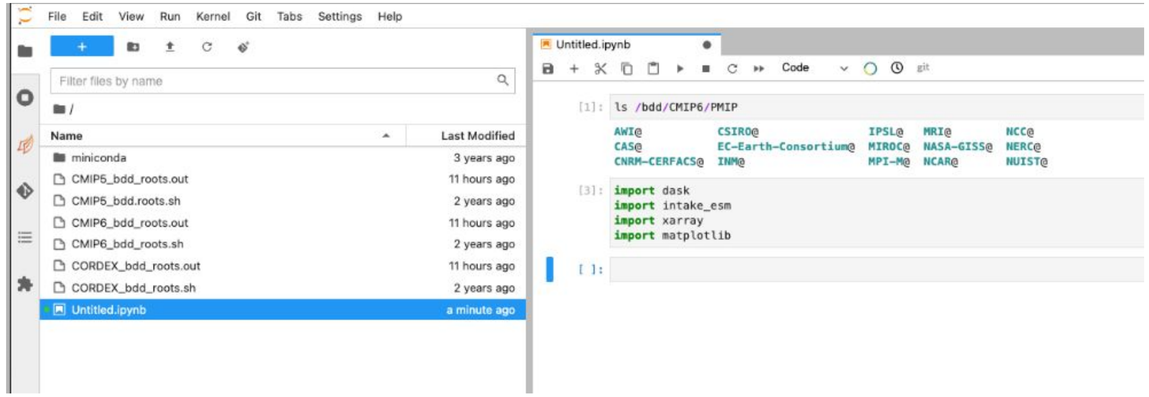
Specific climate data analysing tools

- CliMAF - https://climaf.readthedocs.io/
- ESMValTool (latest version, thanks to Klaus Zimmerman and Stephane Senesi)
- KNMI Climate Explorer (deployment under consideration)

```
$> module load climaf
$> module load esmvaltool
```

ESPRI provides data services for CLIMERI-France and ENES





Remote research environment through JupyterHub

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

Status: Production since April 2022

Documentation: https://documentations.ipsl.fr/

Will be completed and available through the new incoming ESPRI website.

Python environment: PANGEO + useful modules.

Data access: All open data from `/bdd`

ESPRI provides data services for CLIMERI-France and ENES

On-demand data replication service

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. Our tool is able to replicate data from ESGF at a download rate of 4TB/day.

REQUEST

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

ESGF NODES

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

SYNDA
Transfer

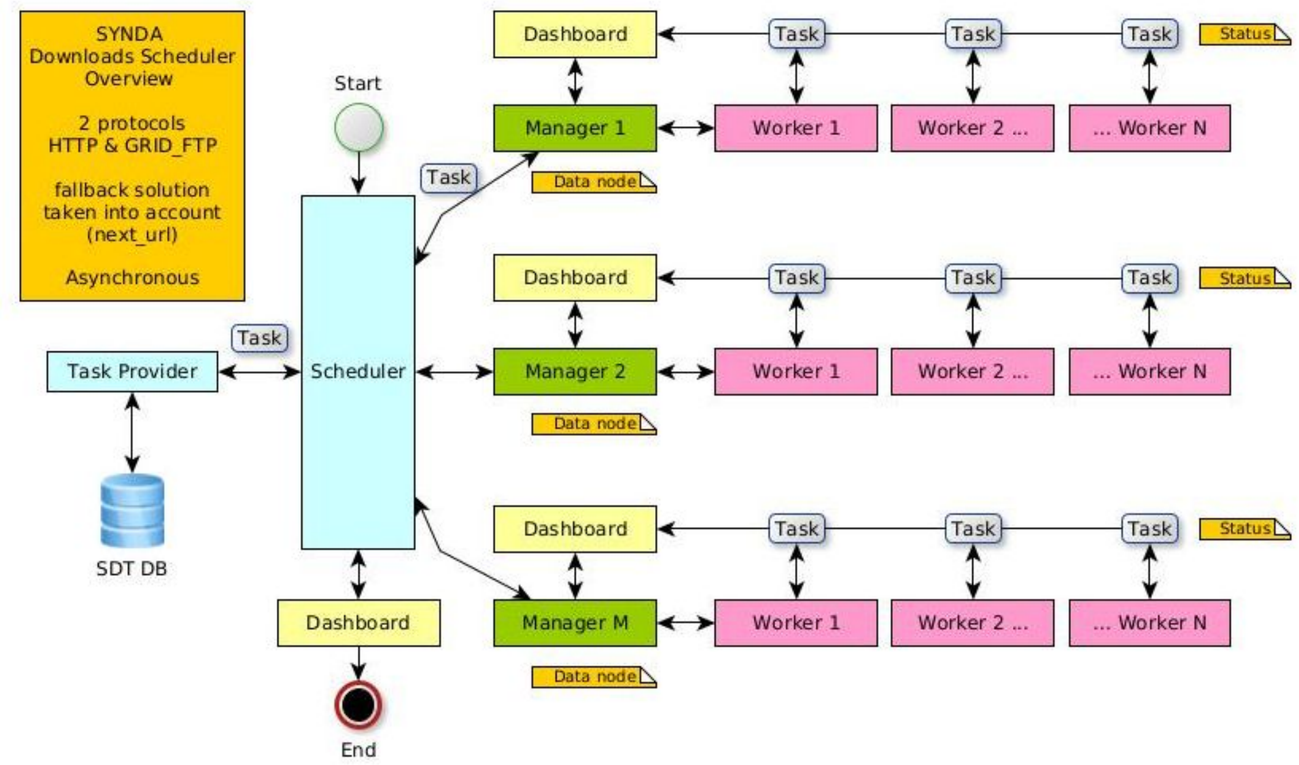
A command line tool to discover and download files from the Earth System Grid Federation (ESGF) archive.

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.



ESPRI provides data services for CLIMERI-France and ENES

On-demand data replication service

How to request additional referenced data?

1. Explore the existing database in `/bdd`
2. If data you need are not locally available, explore the ESGF catalogue: https://esgf-node.ipsl.upmc.fr
   a. If data you need are not available on the ESGF: data does not exist or have not been published.
   b. If data you need is available on the ESGF: create a `.txt` file with the following syntax

```
$> vi my_template.txt
```

```
#login@ipsl.fr
project=CMIP6
experiment=historical amip
model=all
ensemble=r1i1p1f1
variable[day]=clt tas
variable[Omon]=sic evap
```

3. Send your file to espri-mod@listes.ipsl.fr
4. Your request will be examined quickly (space required vs. free space)
   a. We validate your request and run data replication.
   b. We specify together your expectations to satisfy the storage spaces.
5. We notify you as soon as your data is available
6. And of course, we support you!

ESPRI provides data services for CLIMERI-France and ENES

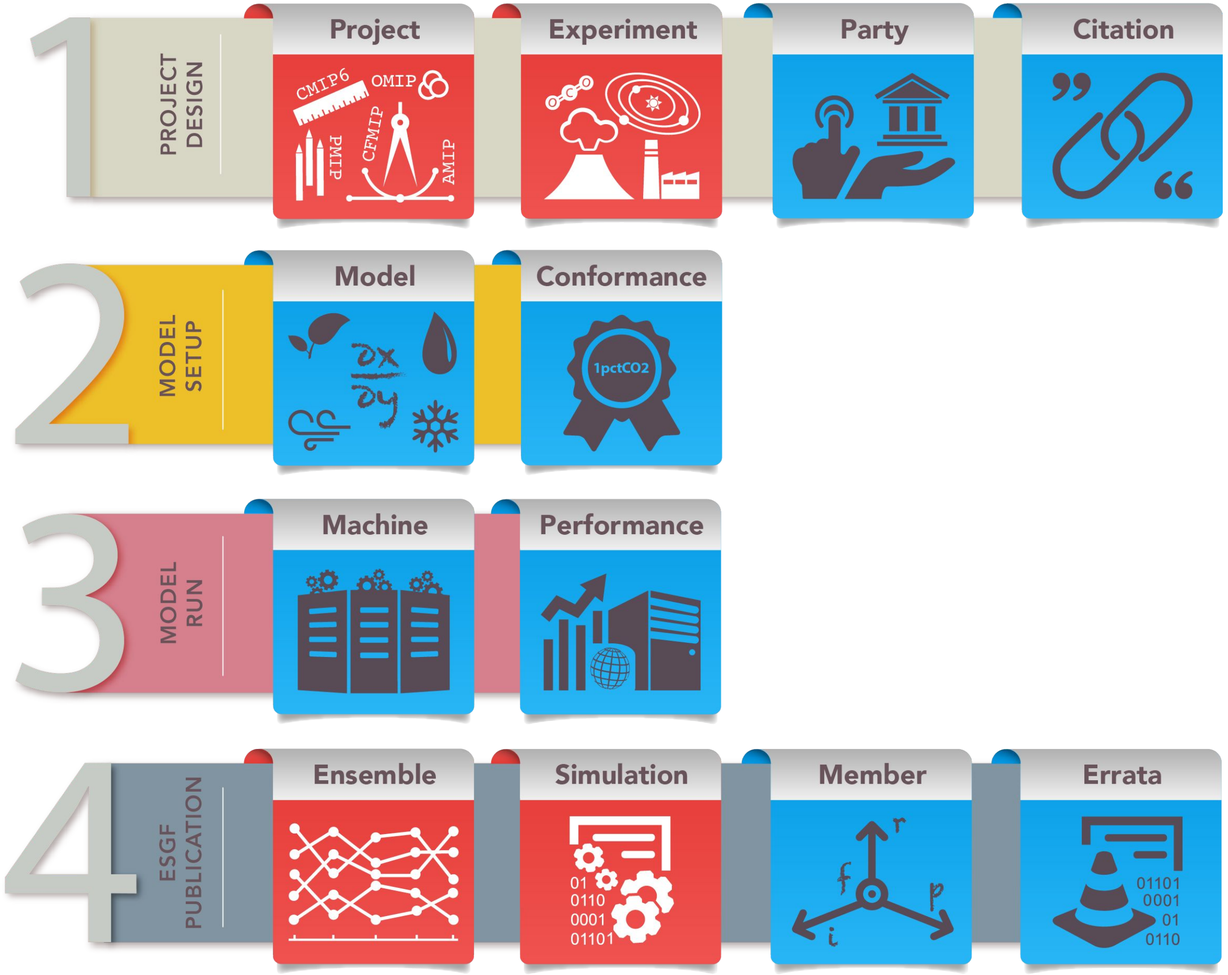
Expertise and support in data documentation and citation

The **Earth System Documentation** (ES-DOC) aims to nurture an ecosystem of tools and services in support of Earth System documentation creation, analysis and dissemination. Such an ecosystem enables the scientific community to better understand and utilize Earth system model data. ES-DOC is coordinated with other community efforts such as CMIP and ESGF via the World Climate Research Programme work group on Climate Modelling (WGCM) and its Infrastructure Panel (WIP).

Home page
https://es-doc.org/

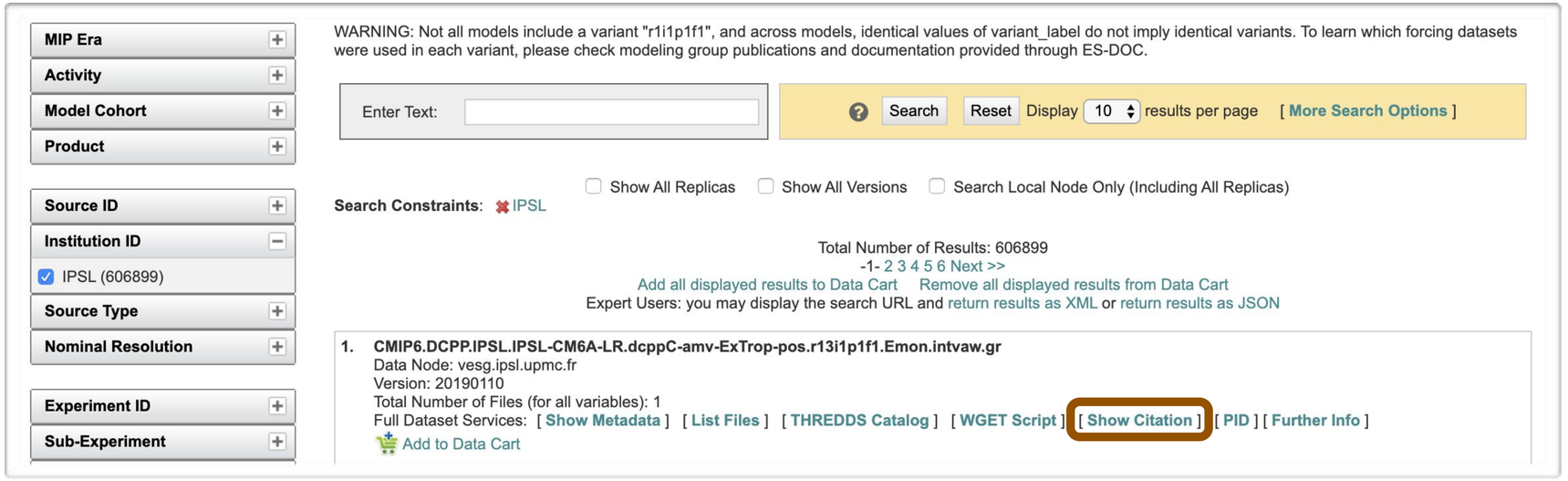
Documentation service
https://search.es-doc.org/

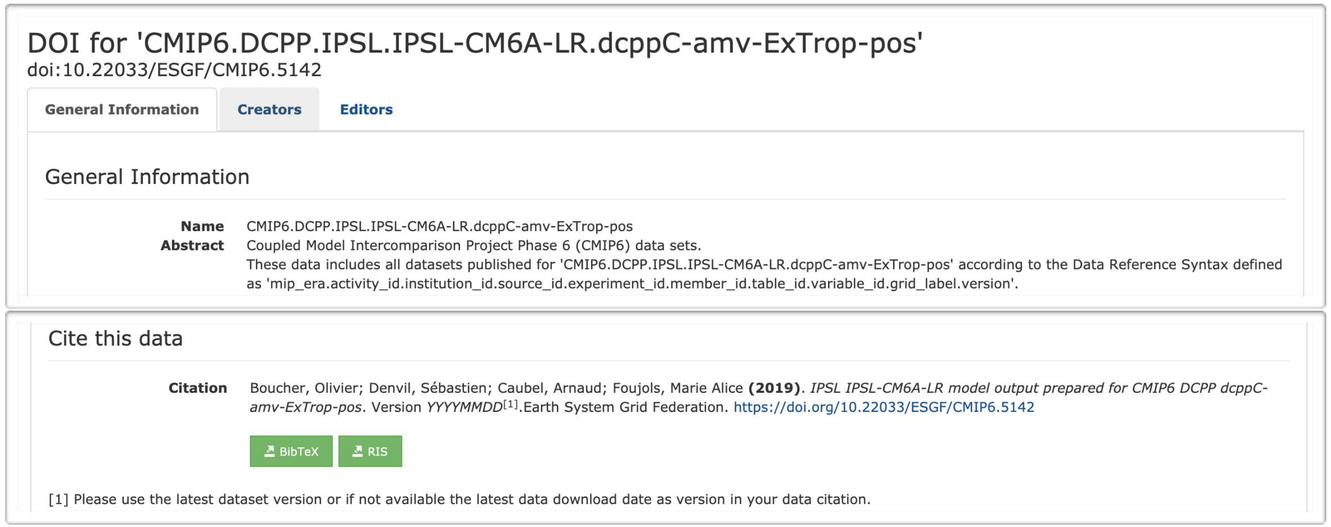
Errata service
https://errata.es-doc.org/



ESPRI provides data services for CLIMERI-France and ENES

Expertise and support in data documentation and citation

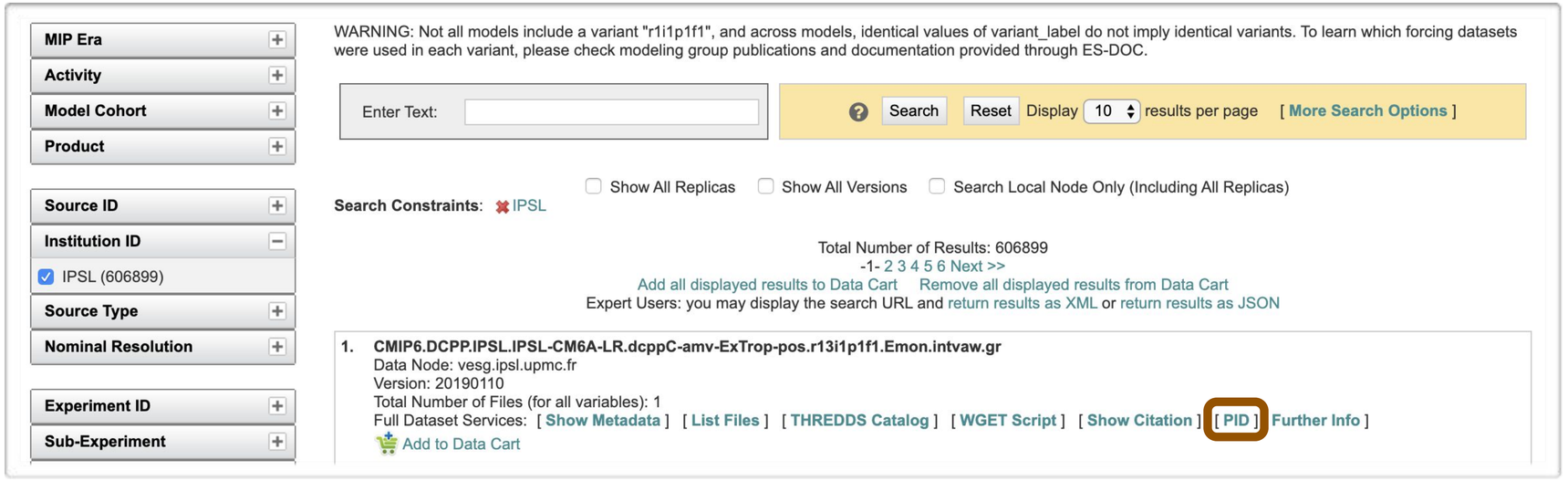




A unique and immutable DOI for each CMIP6 simulation registered at https://cera-www.dkrz.de

ESPRI provides data services for CLIMERI-France and ENES

Expertise and support in data documentation and citation



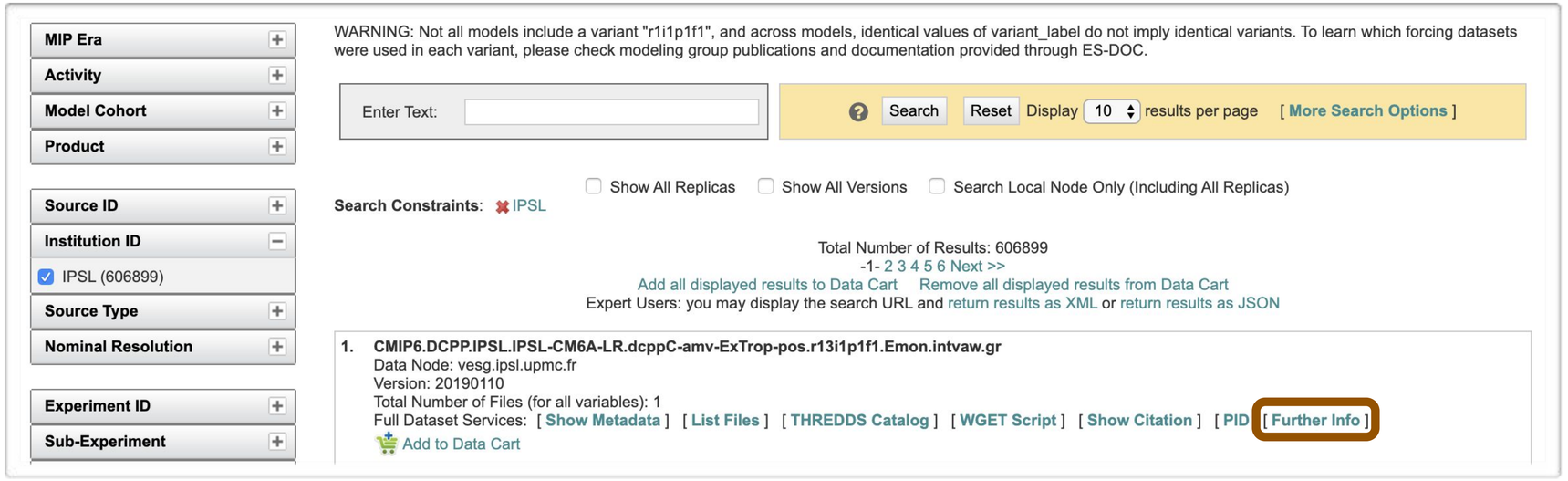
- Identifiant du dataset
- Généalogie des versions
- Errata
- Accès aux réplicas
- Liens vers PID des fichiers

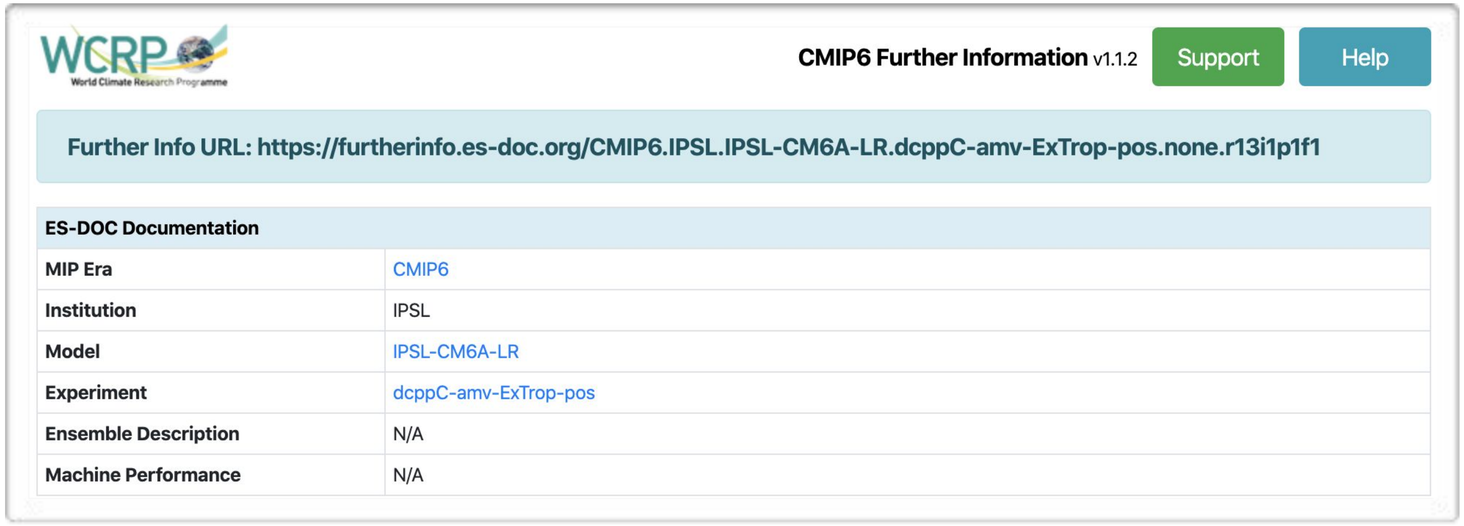
A unique and immutable file identifier, called Persistent IDentifier (PID) or Handle, permanently stored even if the data is removed or deleted. It provides a landing page that collect information about:

- The dataset ID and PID
- Version genealogy with file PIDs
- Links to the errata
- Access to replicas

ESPRI provides data services for CLIMERI-France and ENES

Expertise and support in data documentation and citation





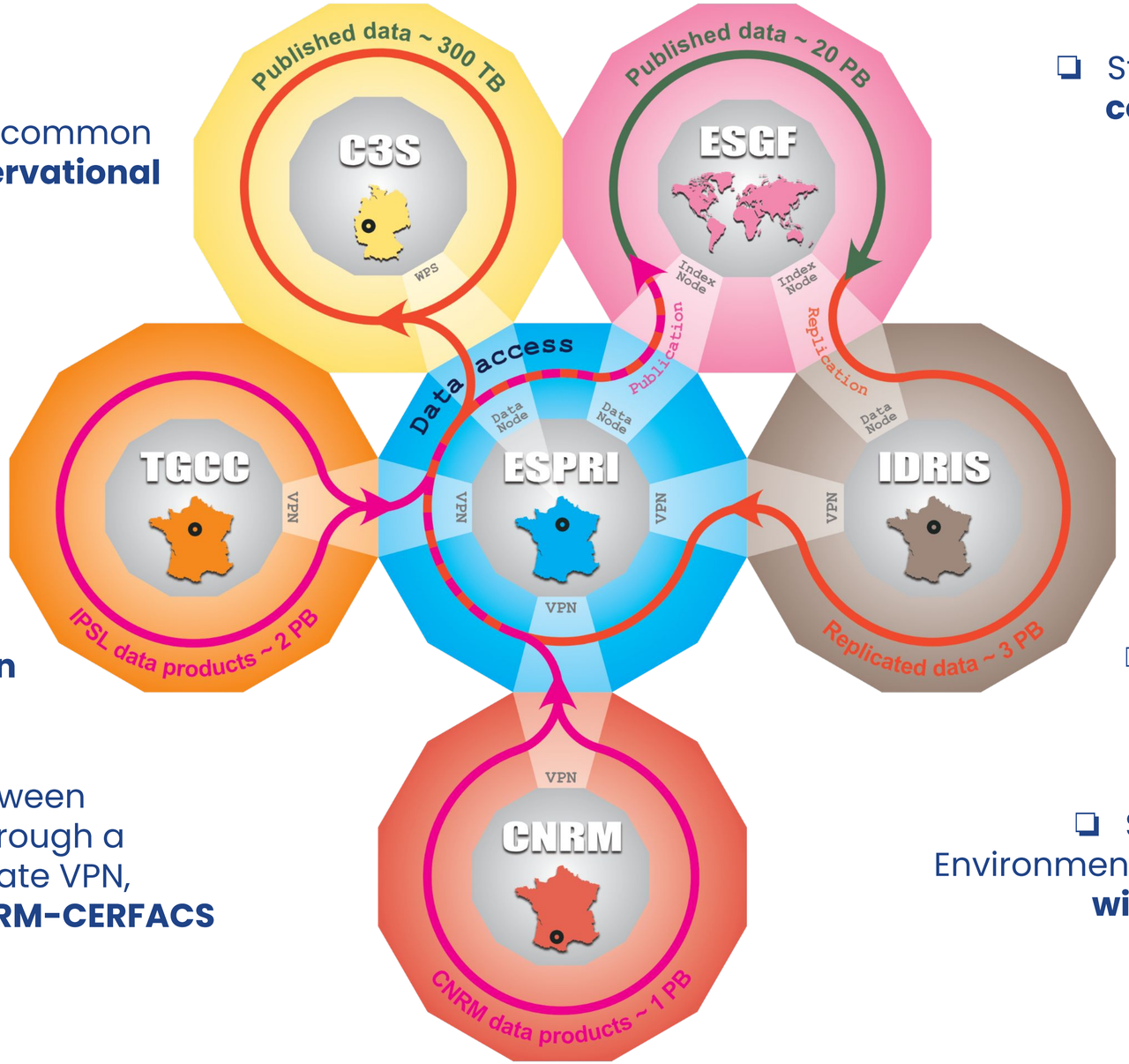
ES-DOC Documentation	
MIP Era	CMIP6
Institution	IPSL
Model	IPSL-CM6A-LR
Experiment	dcppC-amv-ExTrop-pos
Ensemble Description	N/A
Machine Performance	N/A

A dedicated and unique URL, called “further info URL” that collects available information on each CMIP6 simulation:

- Links to the documentation of the model and the experiment
- Links to the errata
- Links to the citation

2023 and beyond with GAIA-Data

- Towards **STAC catalogs** with common **search engine** between **observational and modelling** data sets.
- Extending and automating **DOI services to codes**.
- Implement a **web-based thesaurus** that **guides user in data discovery.**
- **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 Research Data Alliance (RDA).
- **Extending** a multi-thematic **storage at IDRIS.**
- Strengthening Virtual Research Environment based on **Jupyter Notebooks with PANGEO and AI4GEO suites.**

Thank you for your attention

Institut Pierre-Simon Laplace