IS-ENES/ESGF
Virtual Workshop on Compute and Analytics

An open "data-side" platform for climate analytics and services

Guillaume Levavasseur

December 2nd, 2019

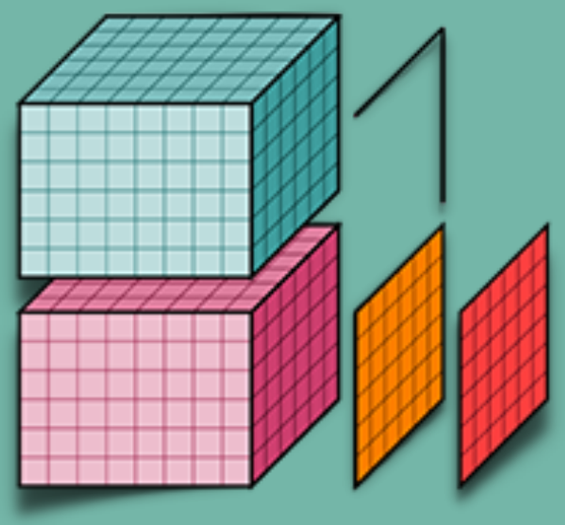
CLIMERI-France

What's ESPRI?

ESPRI = "Common Services for Research at IPSL"

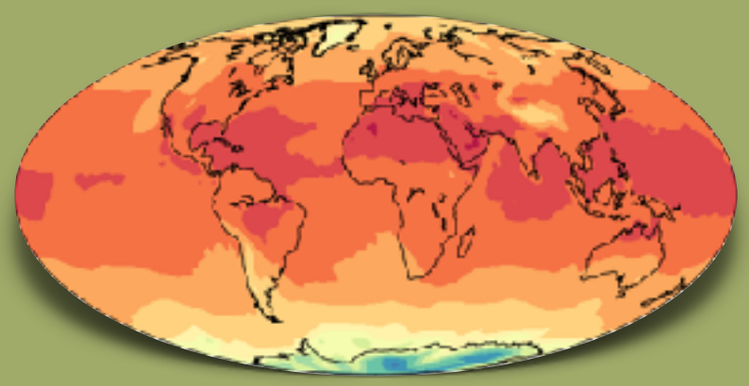
Our mission:

- Ensure the **pivotal position** at the interface **between researchers and data**,
- Design, develop and deploy applications to support the scientific climate community.



Data

VS



Informations

VS

Knowledge

What's ESPRI?

ESPRI = "Common Services for Research at IPSL"
Our mission:

- Ensure the **pivotal position** at the interface **between researchers and data**,
- Design, develop and deploy applications to support the scientific climate community.
- In one word... or four: **FAIR**

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and **reuse** by the community after the data publication process.

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

The ESPRI platform

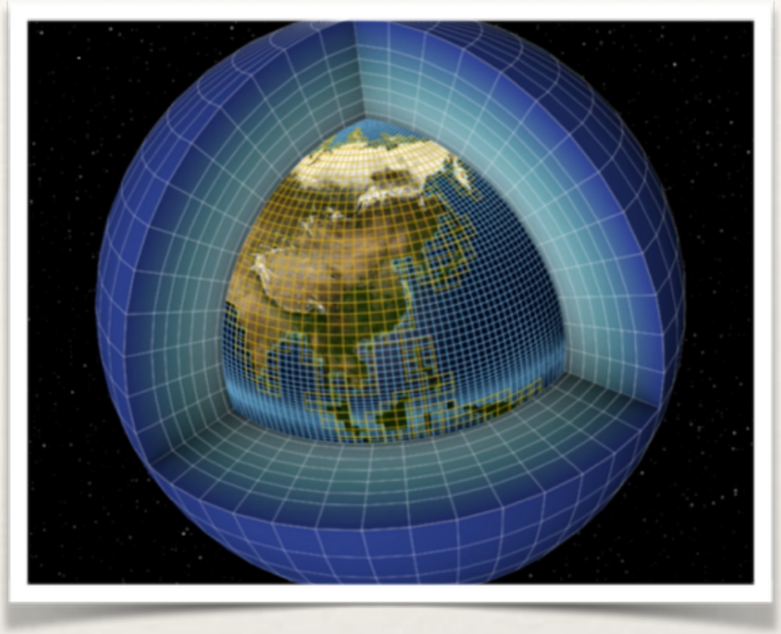
ESPRI, the "local" level:

- Facilitate the **distribution**, **access** and **analysis** of international **climate data,**
- **"on-demand" IPSL data requests.**

The ESPRI platform : ***Model*** *data pools*

Coupled Model Intercomparison Project (CMIP)

- Several scientific topics,
- Several physical processus to explore,
- Several (interconnected) communities
- Several tied projects



Coordinated Regional Climate Downscaling Experiment (CORDEX)

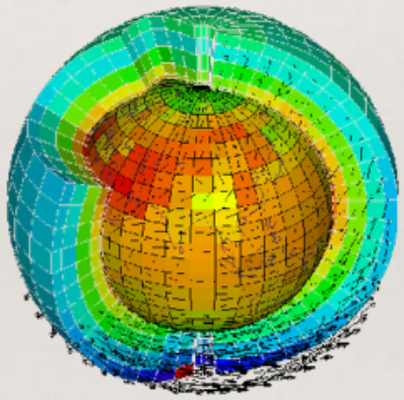
- Focus on regional climate variability and its impacts,
- Several geographical domains,
- Several regional models,
- More and more bias correction methods.

Observations for Model Intercomparison Project (obs4MIP)

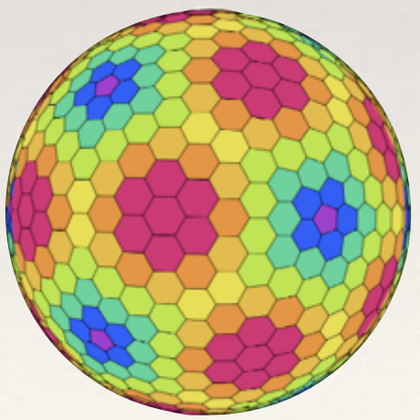
- Collection of observational data ensembles well established and documented for MIP comparison,
- Comply CMIP requirements

Input Datasets for Model Intercomparison Project (input4MIP)

- Boundary conditions and forcings for CMIP6



LMDZ

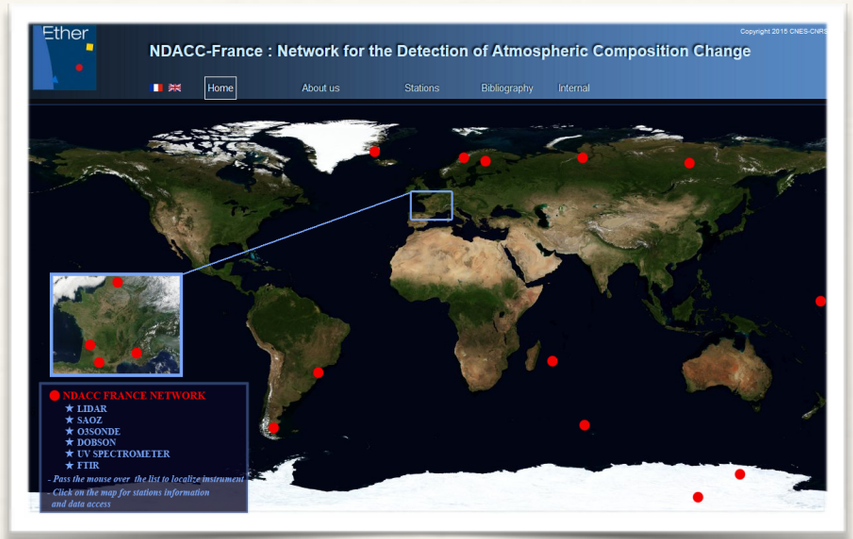


DYNAMICO

The ESPRI platform : ***Observation*** *data pools*

Ground-based and in situ

- Campaigns measurement (CAL/VAL MT, balloons, etc.
- Atmospheric components from 17 stations
- Systematic measurements at SIRTA
- Post-processing of radiosounding (ARSA, TIGR)



Ground-based data

Satellite

- Products from level 1 to 4 (POLDER, PARASOL, CFMIP-obs., etc.
- Model outputs for INDOEX, AMMA, HyMex (+ radar), ChArMEx WFR forcings

Reanalyses

- ERA
- MERCATOR-OCEAN
- NCEP
- FCDR (AMSI, SSMI, GridSat)

Native model data

- 50 atmospheric constituent fields 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

The ESPRI platform : ***Crossing*** *data pools*

ESPRI is a **mutualized** data analysis **platform** providing **optimal** access to climate observations **and** model results, together **close** to the computing facility used by IPSL community(ies).

Federated infrastructure:

- **2 sites** at Sorbonne University (Paris) and Polytechnique Campus (Palaiseau)
- **Shared computing ressources** (over 1,200 cores w/ 3 TB RAM),
- Server virtualization for optimization and reactivity.

Data facilities:

- **Shared data** (≈ **2.7PB**) with dedicated access between sites (SSHFS, FTP, SAMBA) and organized archives,
- **Shared computation framework/environments** (CliMAF, etc.),
- **Data documentation** (DOIs, ES-DOC, errata, etc.)

Services:

- User and project support, close to the scientific teams,
- Acquisition and archival of data produced by the IPSL scientific community (observation + model results)
- Automated replication of data from other data centres or communities,
- **ESGF nodes hosting**,
- Transfer to the civil society through "climate services" (**Copernicus program**, etc.).

"CICLAD" cluster (Paris)

"ClimServ" cluster (Palaiseau)

The ESPRI platform

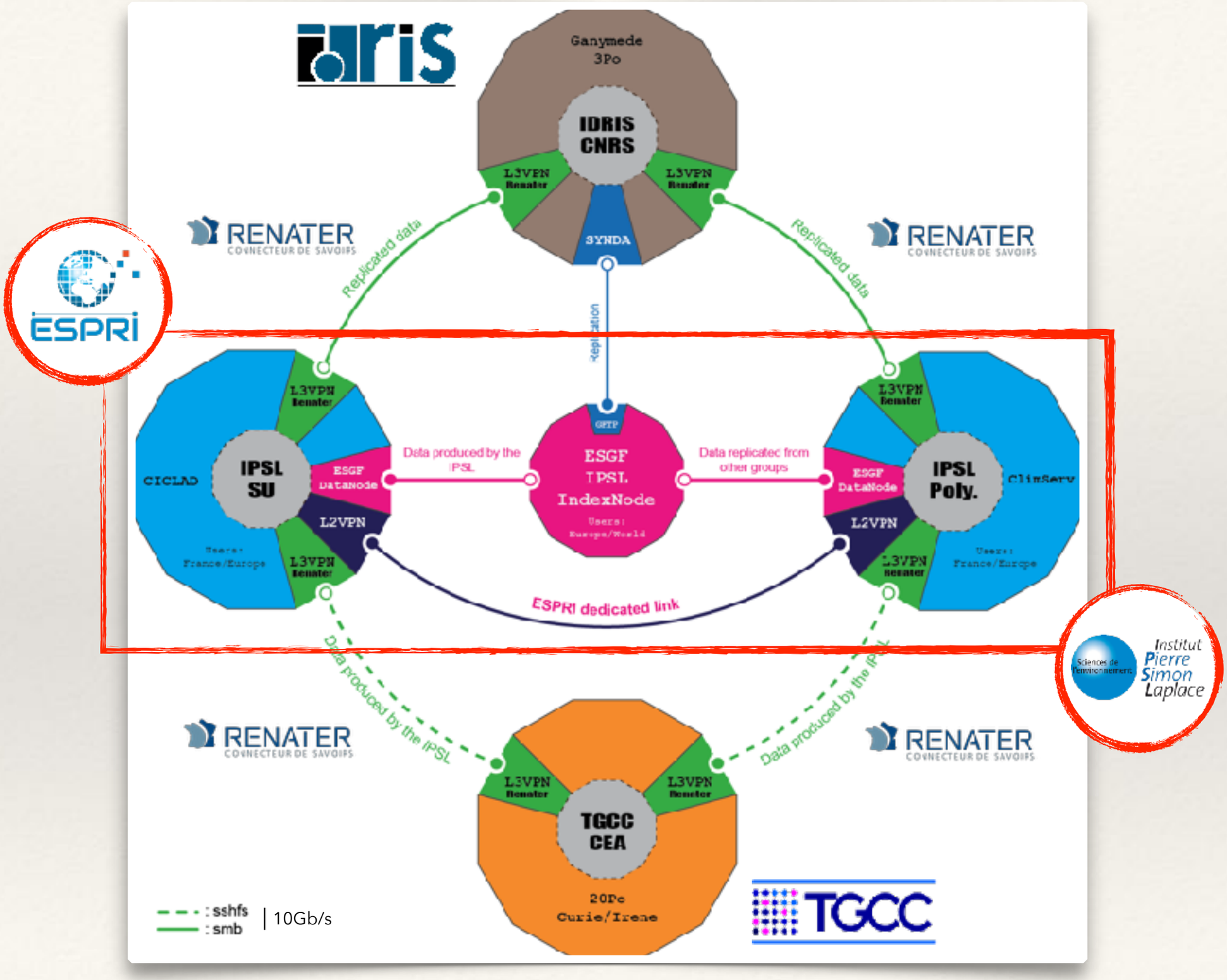
ESPRI, the "local" level:

- Facilitate the **distribution**, **access** and **analysis** of international **climate data,**
- **"on-demand" IPSL data requests.**

CLIMERI-France, the national level:

- Accompanying the community,
- **Coordination between french partners,**
- **Relies on ESPRI plateform**.

CLIMERI-France: a national infrastructure dedicated to climate modeling



The ESPRI platform

ESPRI, the "local" level:

- Facilitate the **distribution**, **access** and **analysis** of international **climate data,**
- **"on-demand" IPSL data requests.**

CLIMERI-France, the national level:

- Accompanying the community,
- **Coordination between french partners,**
- **Relies on ESPRI plateform**.

IS-ENES, the European level

- Coordination between European partners (BADC, DKRZ, etc.),
- Raise the needs of french partners on the forefront of discussions,
- **Strengthening the infrastructure, through operational implementation of the ESGF.**

The ESPRI platform

ESPRI, the "local" level:

- Facilitate the **distribution**, **access** and **analysis** of international **climate data,**
- **"on-demand" IPSL data requests.**

CLIMERI-France, the national level:

- Accompanying the community,
- **Coordination between french partners,**
- **Relies on ESPRI plateform**.

IS-ENES, the European level

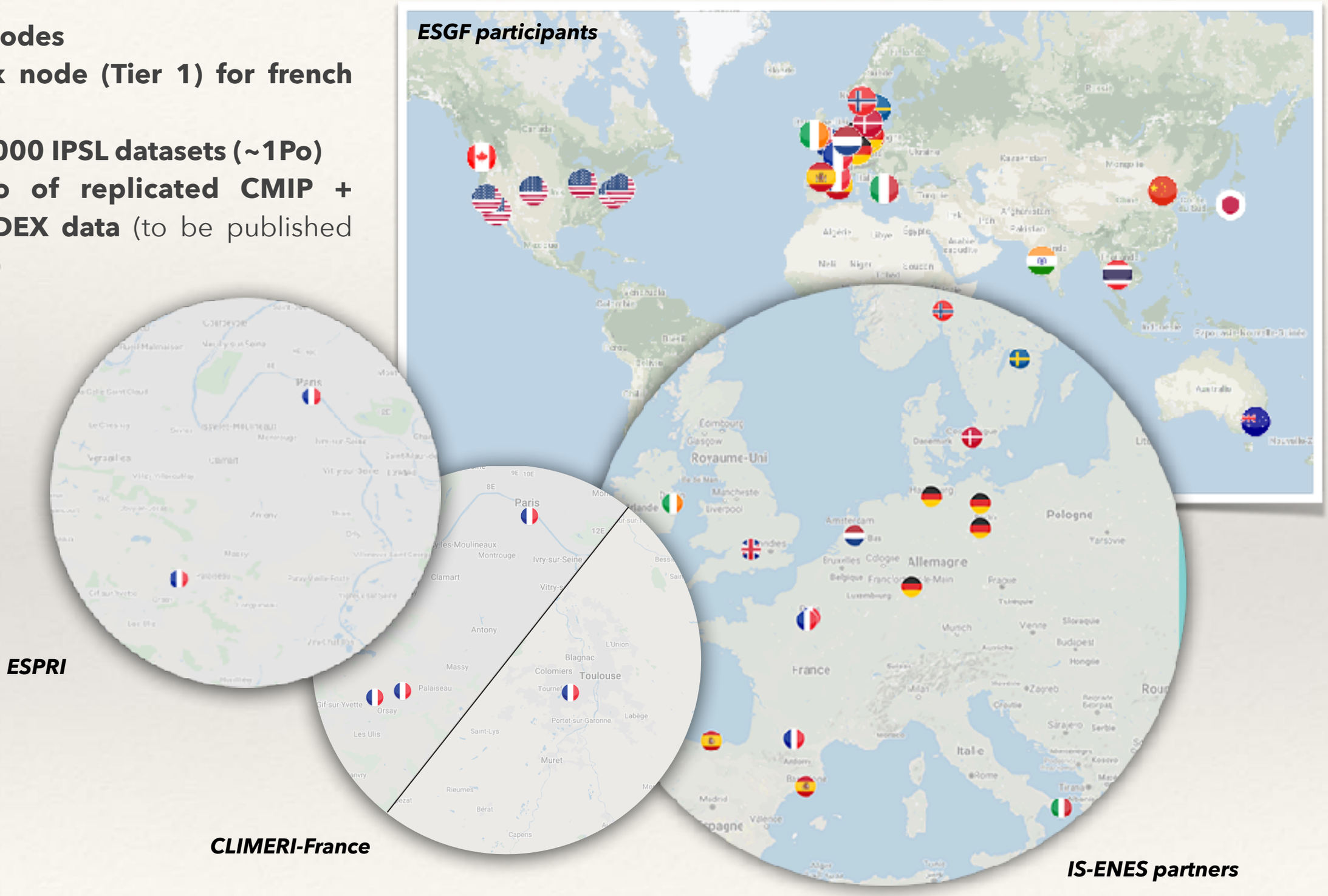
- Coordination between European partners (BADC, DKRZ, etc.),
- Raise the needs of french partners on the forefront of discussions,
- **Strengthening the infrastructure, through operational implementation of the ESGF.**

ESGF, the international level

- CMIP Data Node Operation Team (Member)
- ESGF Governance (Executive Committee),
- ES-DOC Governance (Principal Investigator).

The ESPRI platform into ESG Federation

- **2 datanodes**
- **1 index node (Tier 1) for french results**
  - **230 000 IPSL datasets (~1Po)**
  - **1,5Po of replicated CMIP + CORDEX data** (to be published soon)



Climate "Big" Data

Couple Model Intercomparison Project Phase 6 (CMIP6)

Many, many processes… many, many (interconnected) communities !

CMIP3 archive: **24** models **x 12** experiments = **39TB** (82 340 files)

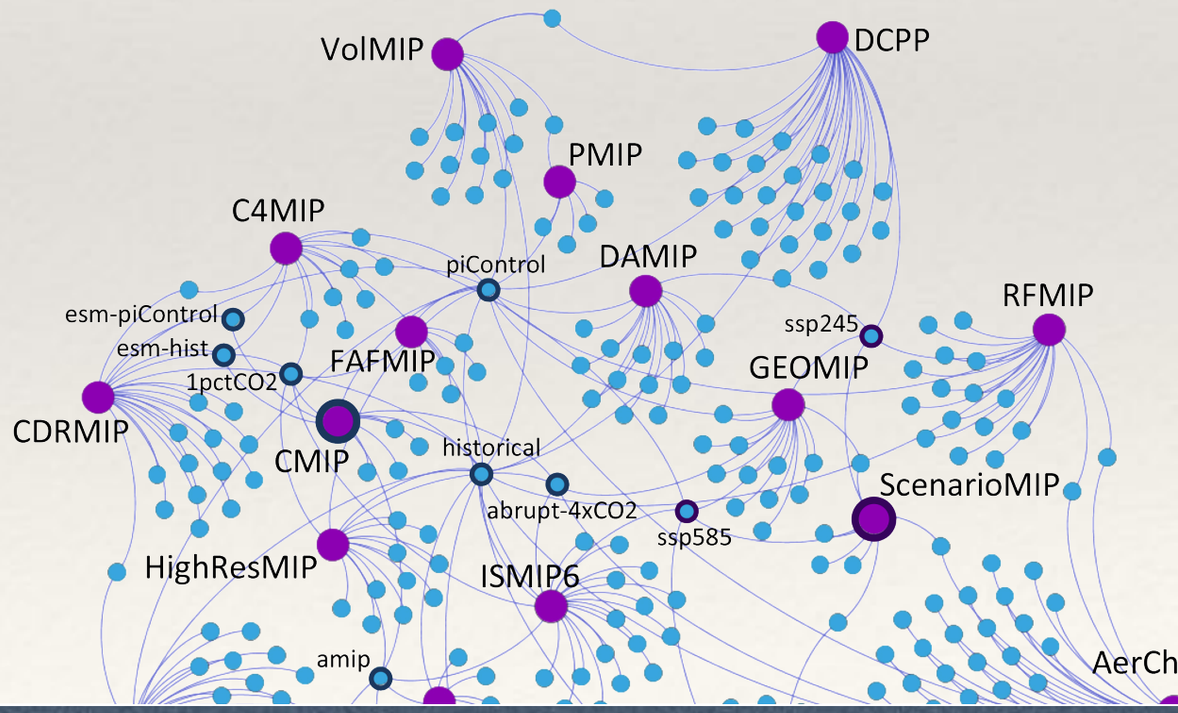
x50

CMIP5 archive: **63** models **x 101** experiments = **1,8PB** (4,3 millions of files)

x20-50

CMIP6 archive: **126** models **x 299** experiments **x** finer resolution **x** larger ensembles =

36PB (86 millions of files) < ??? < 90PB (215 millions of files)



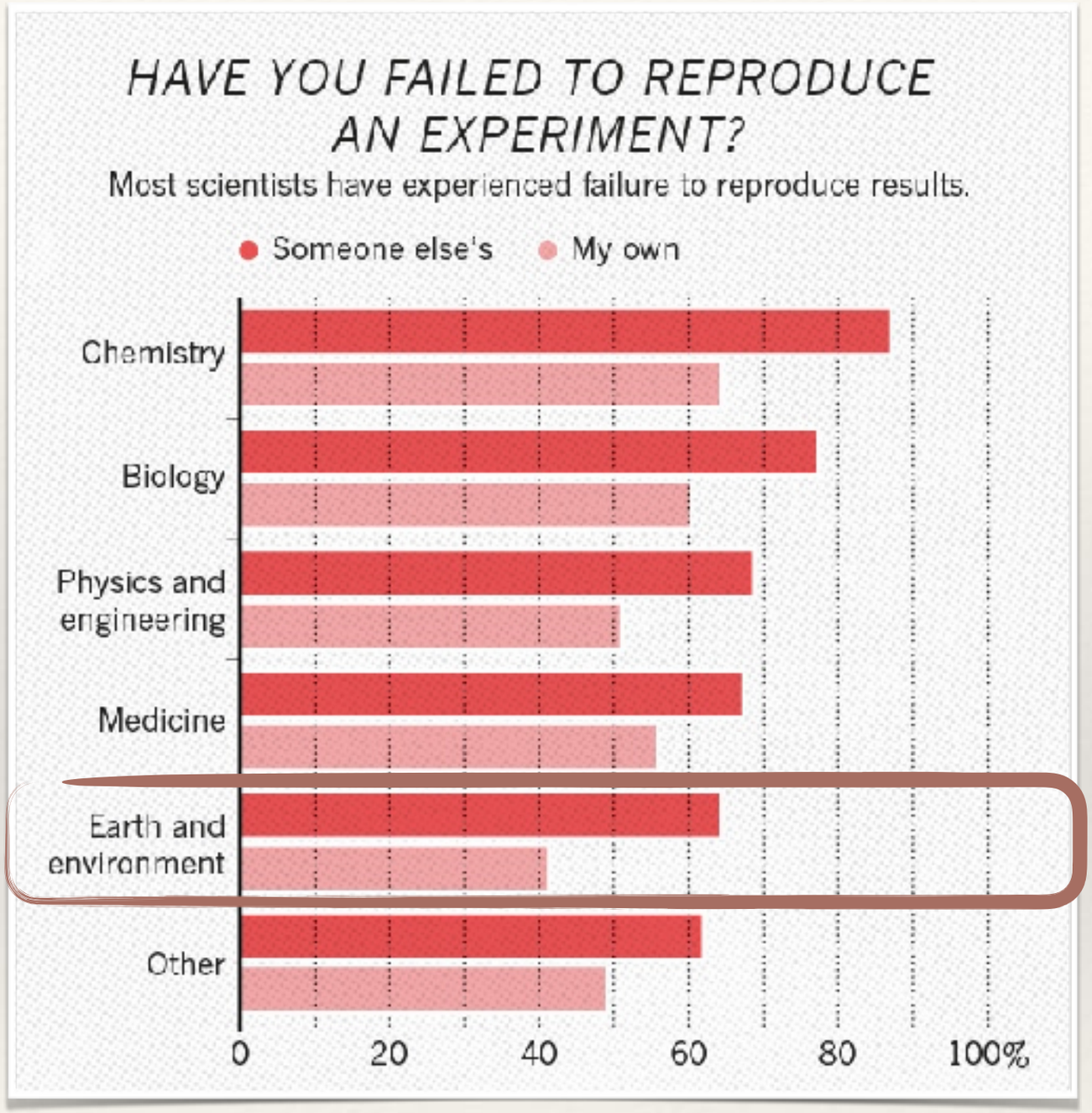
Some lessons from the past... and from users...

- Vocabularies discrepancies
- Inhomogeneous indexes
- Few web processing services
- POSIX filesystems limitations
- Lost in the UI
- Data discovery not user-friendly

...

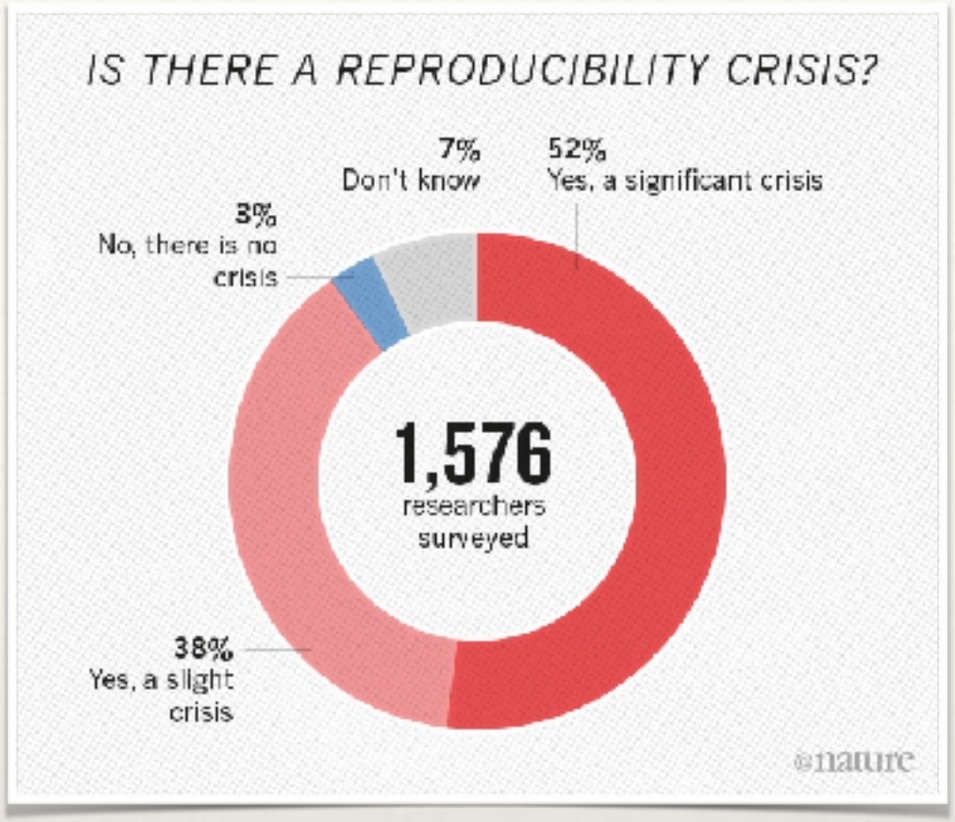
« *I have not failed, I have just found 10000 ways that don't work.* » *(A. Einstein)*

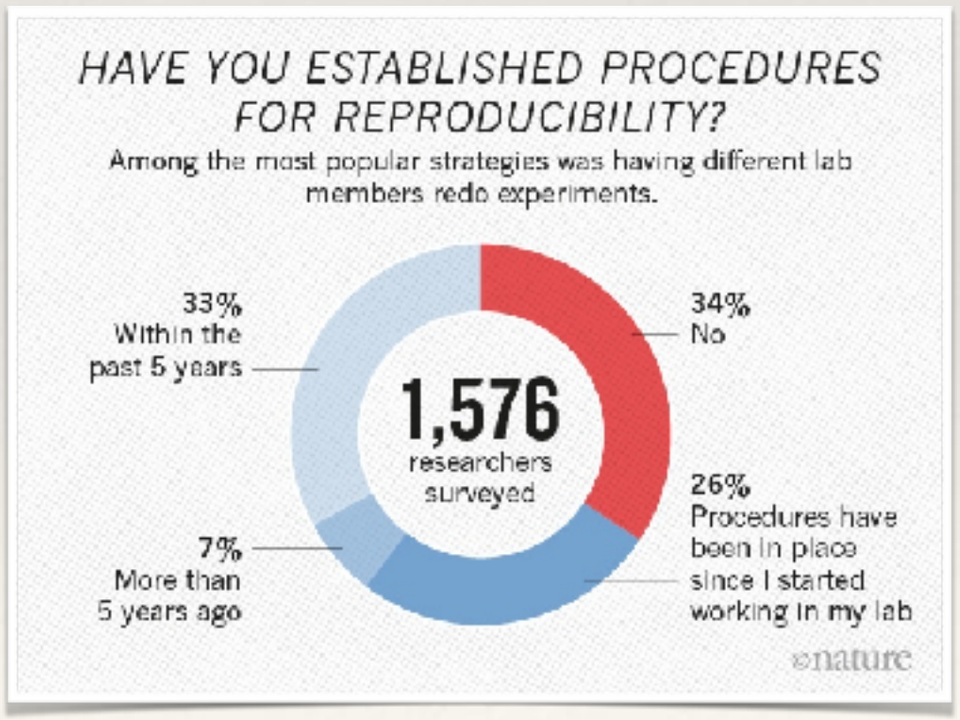
Reproducibility crisis?



1,500 scientists lift the lid on reproducibility

Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a





How ESPRI aims to address the user needs?

- Improve/complete our analytic environment

`CliMAF` (Climat Assessment Framework) is an open source software that aims to ease the common steps that separate scientists from their diagnostics. `CliMAF` is able to deal with:

- Several DRS
- Share diagnostics (cache mechanism)
- Plug and play with homemade scripts
- Usual data treatments (subsetting, regridding, ensemble mean, etc.)
- Cache mechanism to not recompute a whole treatment chain.

Climate modelers
Model assessment,
Model intercomparison,
Home-made scripts and methods.

Future plan for CliMAF under discussion:

- Improve discovery on filesystem using `pyessv` CV manager
- Rely on `xarray` + `Dask` instead of `CDO` operators?

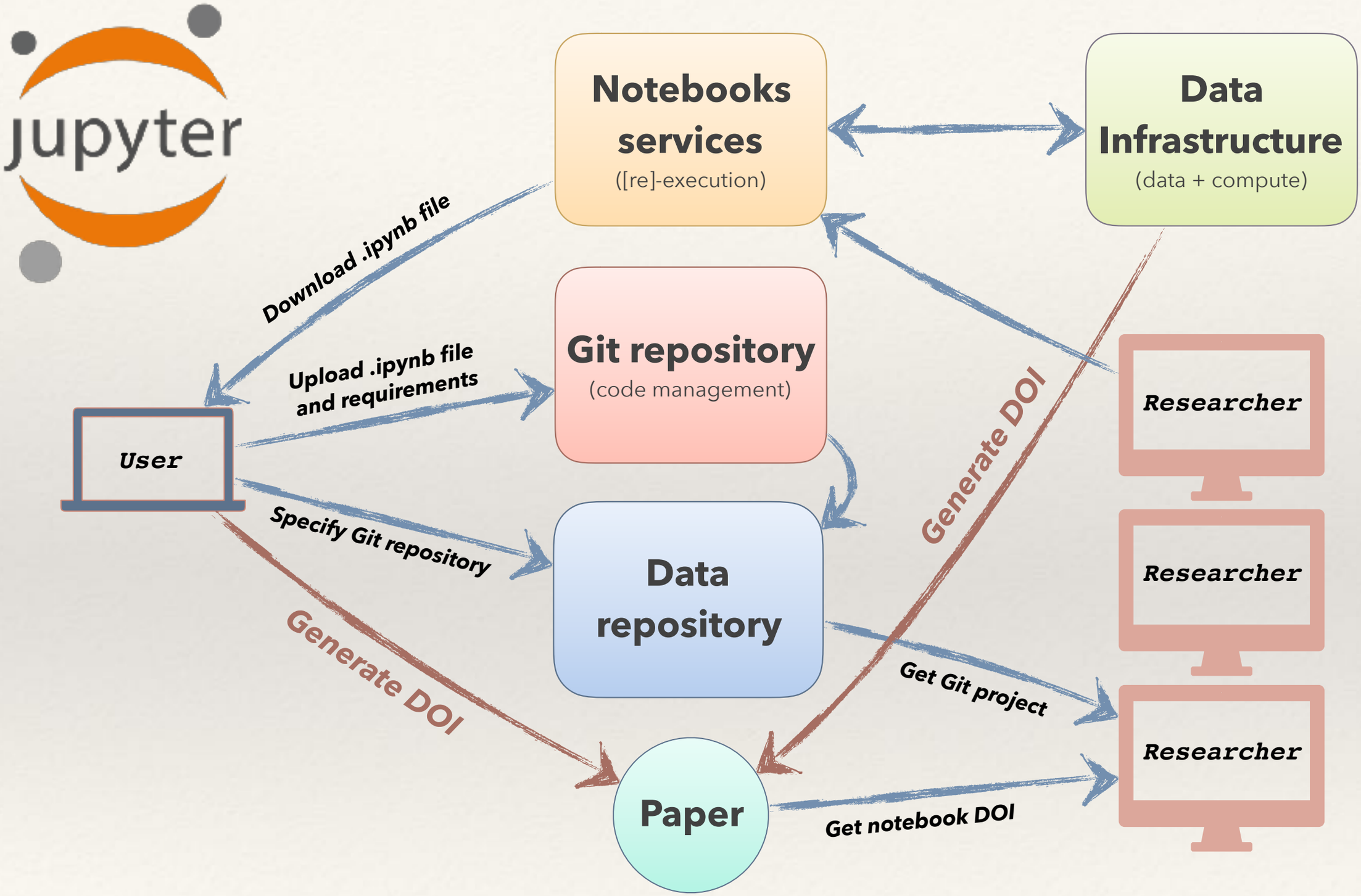
How ESPRI aims to address the user needs?

- Improve/complete our analytic environment
- `Kubernetes` instance soon in production on ClimServ cluster
  - `Jupyter` Notebooks for training purposes and analysis traceability,
  - ESGF data node hosting (vesg.polytechnique.ipsl.fr ?).

Scientific researchers in climate
but who are not climate modelers themselves

Interdisciplinary studies,
Limited time to learn for data usage,
Need to be straight to the goal.

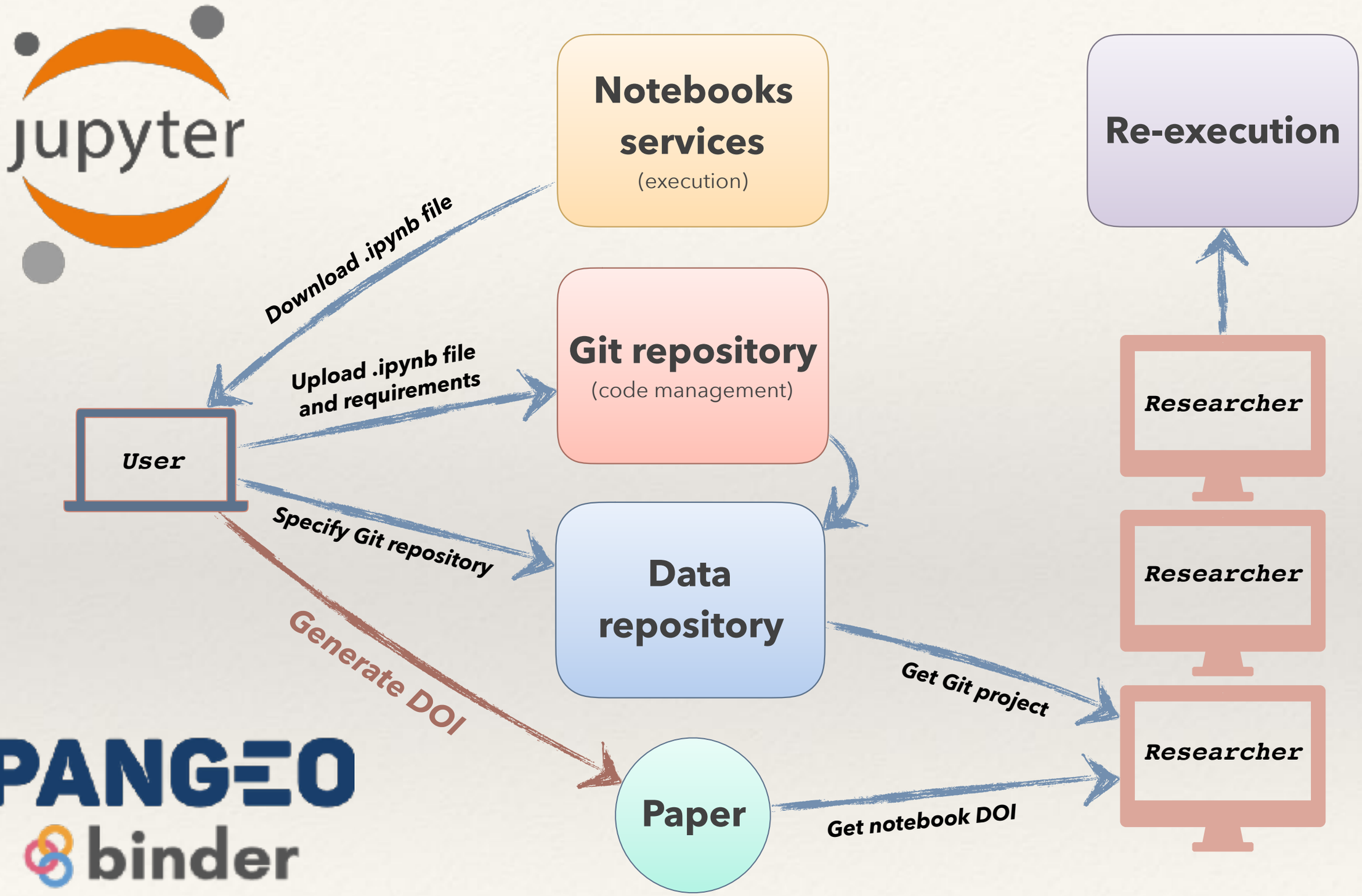
Reproducibility needs "environment traceability"



How ESPRI aims to address the user needs?

- Improve/complete our analytic environment
- `Kubernetes` instance soon in production on ClimServ cluster
  - `Jupyter` Notebooks for training purposes and analysis traceability,
  - ESGF data node hosting (vesg.polytechnique.ipsl.fr ?).
- Explore "object stores" technologies and Pangeo Binders

Reproducibility needs "environment traceability"

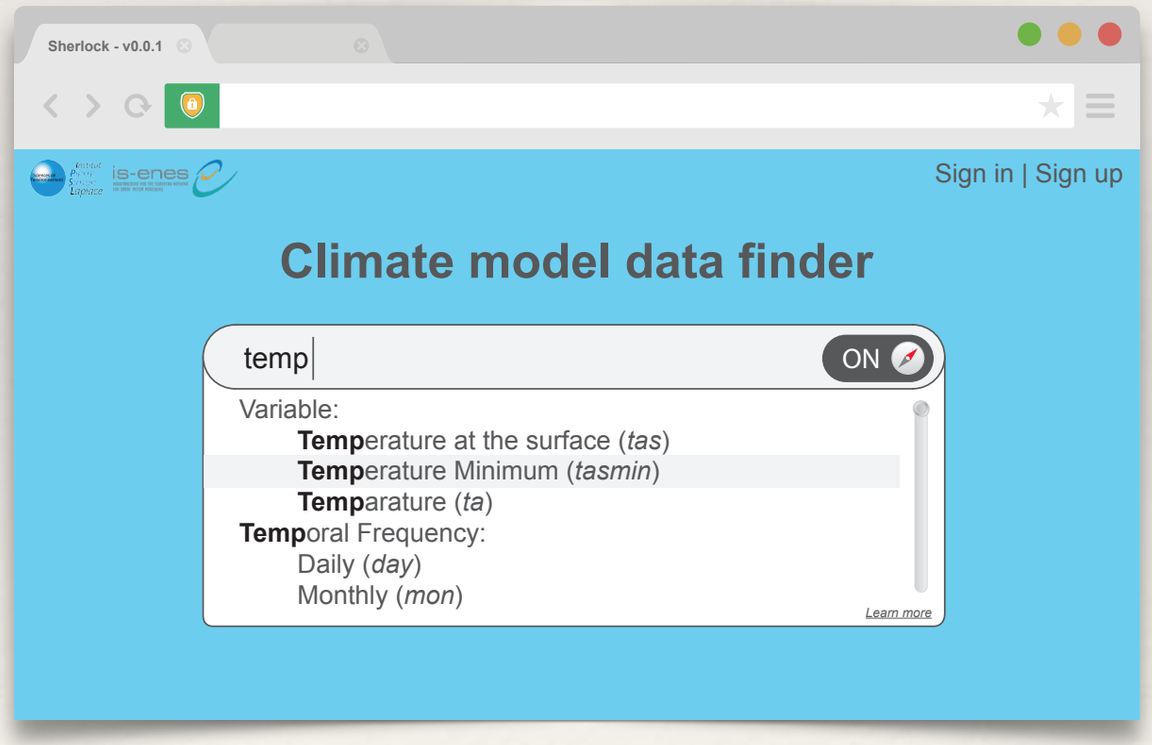


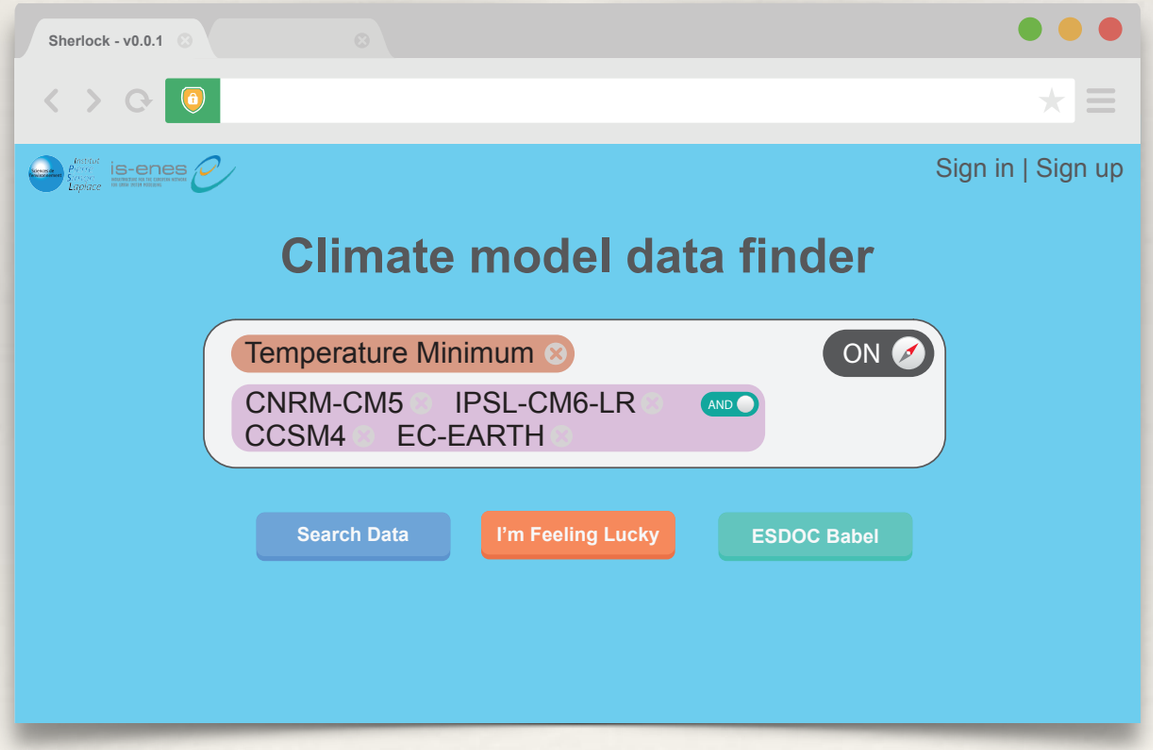
How ESPRI aims to address the user needs?

- Improve/complete our analytic environment
- `Kubernetes` instance soon in production on ClimServ cluster
  - `Jupyter` Notebooks for training purposes and analysis traceability,
  - ESGF data node hosting (vesg.polytechnique.ipsl.fr ?).
- Explore "object stores" technologies and Pangeo Binder,
- Work on PoC of a "Climate Dataset Finder" to improve data discovery on the platform.

elasticsearch

INTAKE





How ESPRI aims to address the user needs?

- Improve/complete our analytic environment
- `Kubernetes` instance soon in production on ClimServ cluster
  - `Jupyter` Notebooks for training purposes and analysis traceability,
  - ESGF data node hosting (vesg.polytechnique.ipsl.fr ?).
- Explore "object stores" technologies and Pangeo Binder,
- Work on PoC of a "Climate Dataset Finder" to improve data discovery on the platform.
- WPS deployed for climate services (Copernicus) which needs standard processes (e.g., reggridding, subsetting, etc.):
  - `Birdhouse` solution from DKRZ
  - Load-balanced WPS

Scientific researchers (or non-researchers) from other domains

Assess the impacts of climate change on ecosystems, economic activities, industry and other applications,
Don't know about data vocabularies, architecture and/or standards.
Need for guidelines or expertise.

Thank you for your attention.