is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

EGI Conference 2020
Experience from Around the World with Federated Clouds and Data

ENES climate data infrastructure

Stephan Kindermann
DKRZ

European Network for Earth System Modelling (ENES)

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

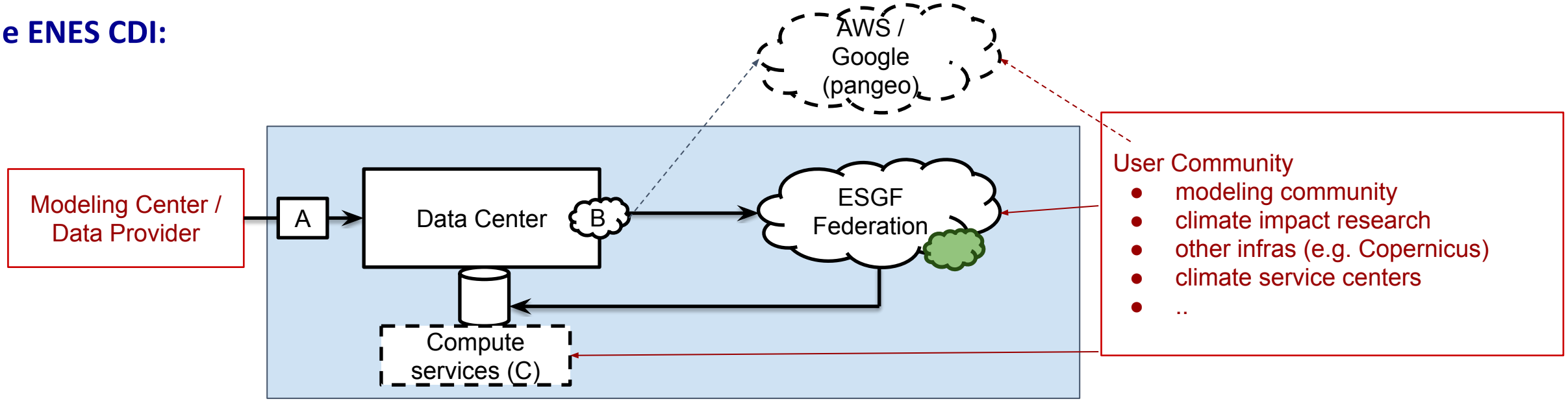
The ENES Climate Data Infrastructure (CDI)

Overview

- Main objectives, user groups, evolution
- Core infrastructural components and services of the data infrastructure
- Evolving aspects:
  - computation near to the data
  - towards cloud based infrastructure(s)

The ENES Climate Data Infrastructure (CDI)

The ENES CDI:



- **A: Metadata and Data standards**
- **B: Data centers integrating services in a federation (ESGF based)**
- **C: Enabling data analysis activities near to the data**

towards cloud deployments:

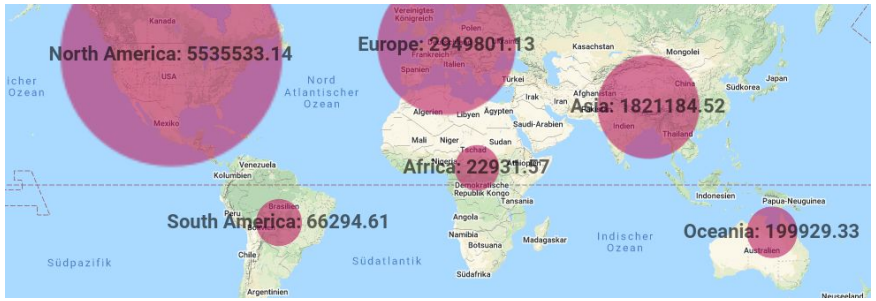
- **a) hosting of core components in commercial clouds**
- **b) hosting data and compute in commercial (and institutional) clouds**

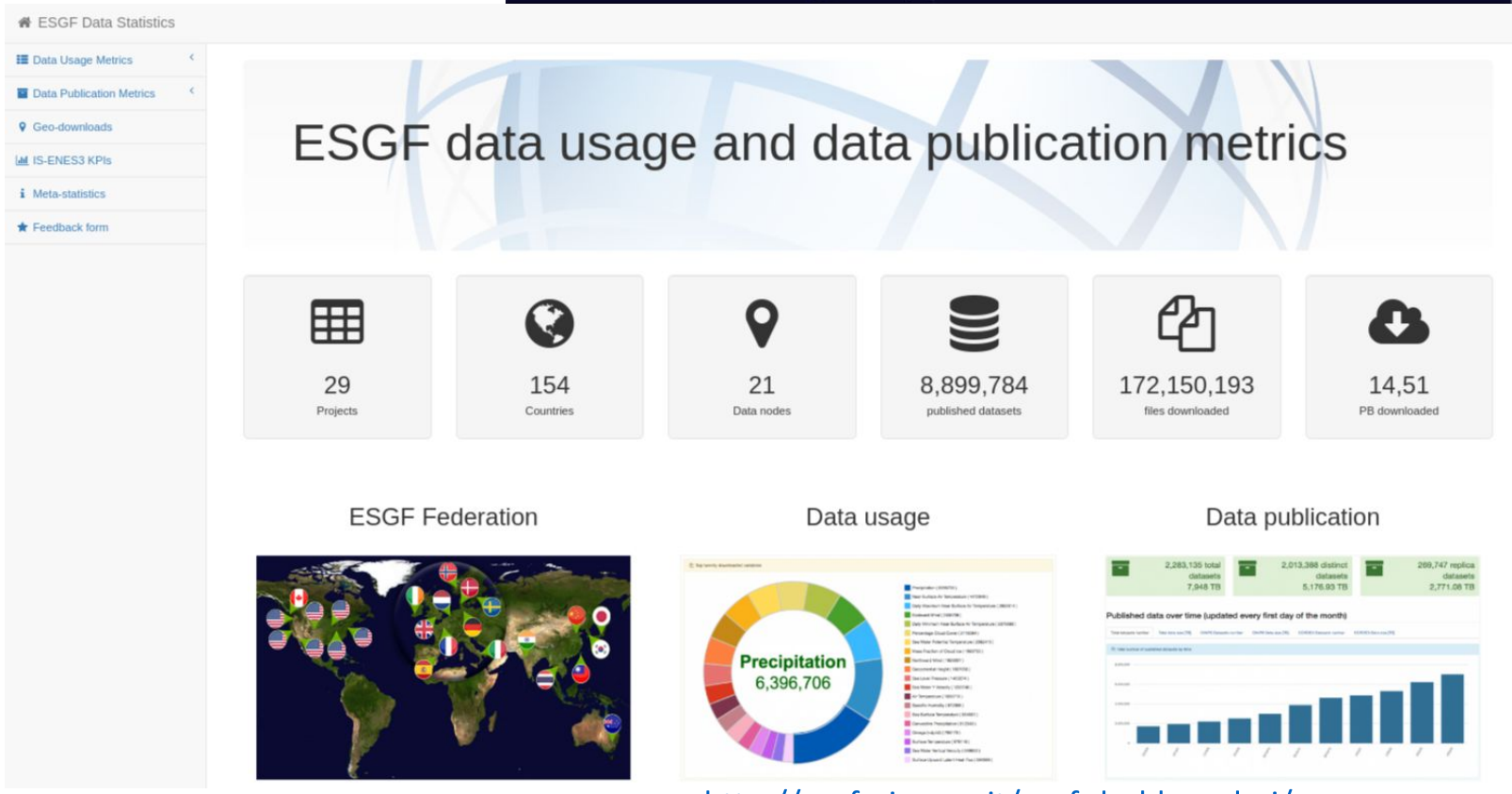
The ENES CDI: objectives

Support data distribution for internationally coordinated climate model intercomparison projects (CMIP, currently CMIP6)

Example: ENES CDI supporting CMIP6:

- 4 ESGF portals (world: 10)
- 11 ESGF data nodes (world: 29)
  - France, Germany, UK (tier1)
  - Ireland, Italy, Norway, Spain, Sweden
- Models: 46 (world: 103)
- Modeling Institutes: 19 (world: 41)



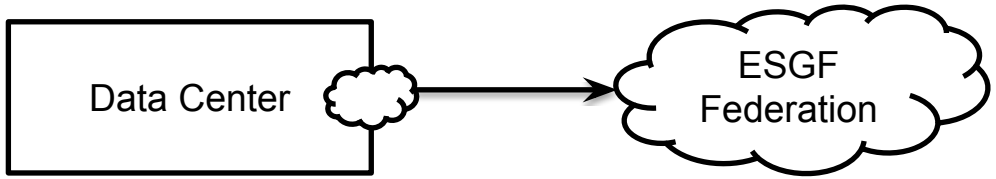


http://esgf-ui.cmcc.it/esgf-dashboard-ui/

ENES CDI: The ESGF software stack

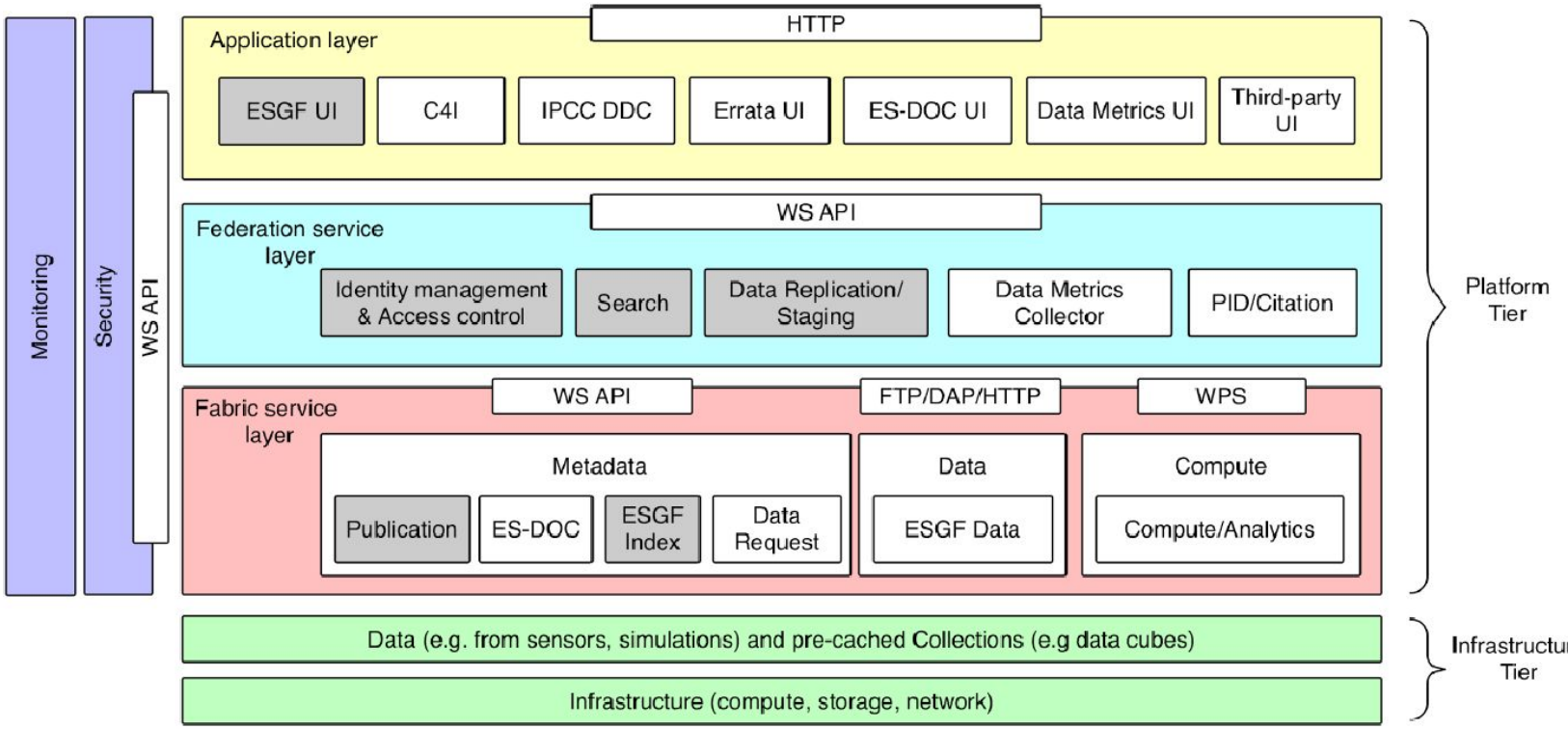
The ESGF SW stack evolved over the past > 10 years as part of an international collaborative effort supported by national/institutional funding

- IS-ENES3 project is currently coordinating the European contributions (white boxes in picture)
- github hosted and open source



Current challenges:

- Modularization, orchestrated deployment support (docker, ansible, kybernetes)
- Worldwide operating system, best effort funding → sustainability
  - EU: discussed as part of IS-ENES3
  - Worldwide: ESGF XC, WCRP, .. discussions



ENES CDI: example services beyond basic data distribution

Beyond data search and access services in support of CMIP6:

- Persistent data identification service (PID service)
- Errata service
- Data citation service (DOI service)

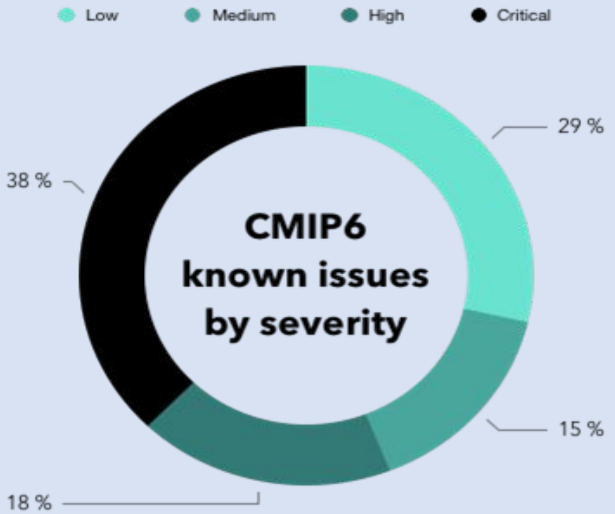
} Supporting key aspects of FAIR data principles

PID service:

- Operational integration based on worldwide distributed message queue
- overall 16.5 * 10^6 PIDs assigned (including retracted etc.)
- << 0.1% registration problems
- automatic curation scripts in place
- Cooperation with EOSC, EPIC, EUDAT, RDA for sustainability

Errata service:

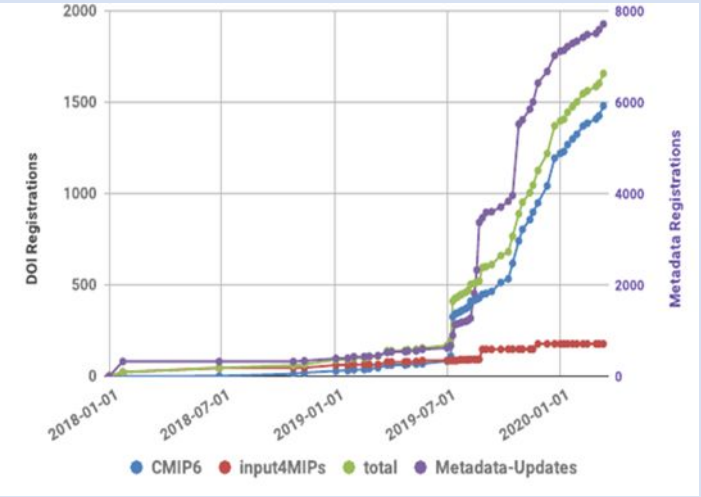
- > 175 issues registered
- > 90.000 data sets affected
  3% of CMIP6 archive



Data citation service:

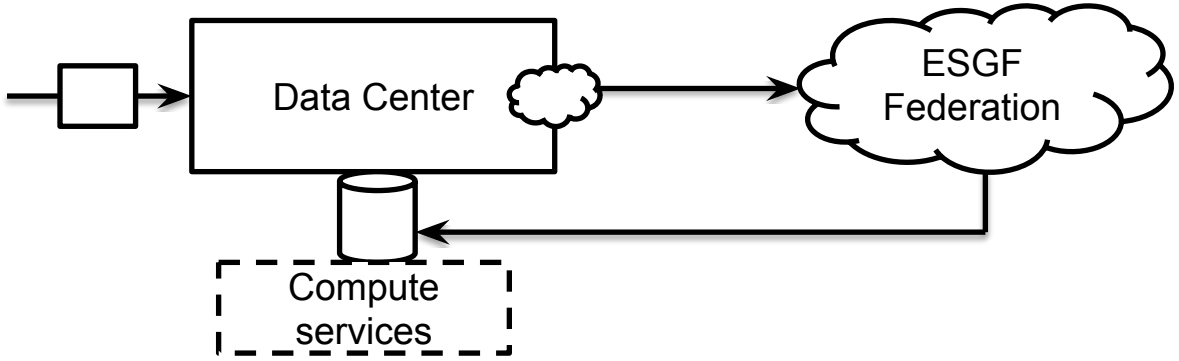
- > 6000 DOI registrations for CMIP6
- > 8000 citation metadata updates

DataCite DOI Registration for CMIP6



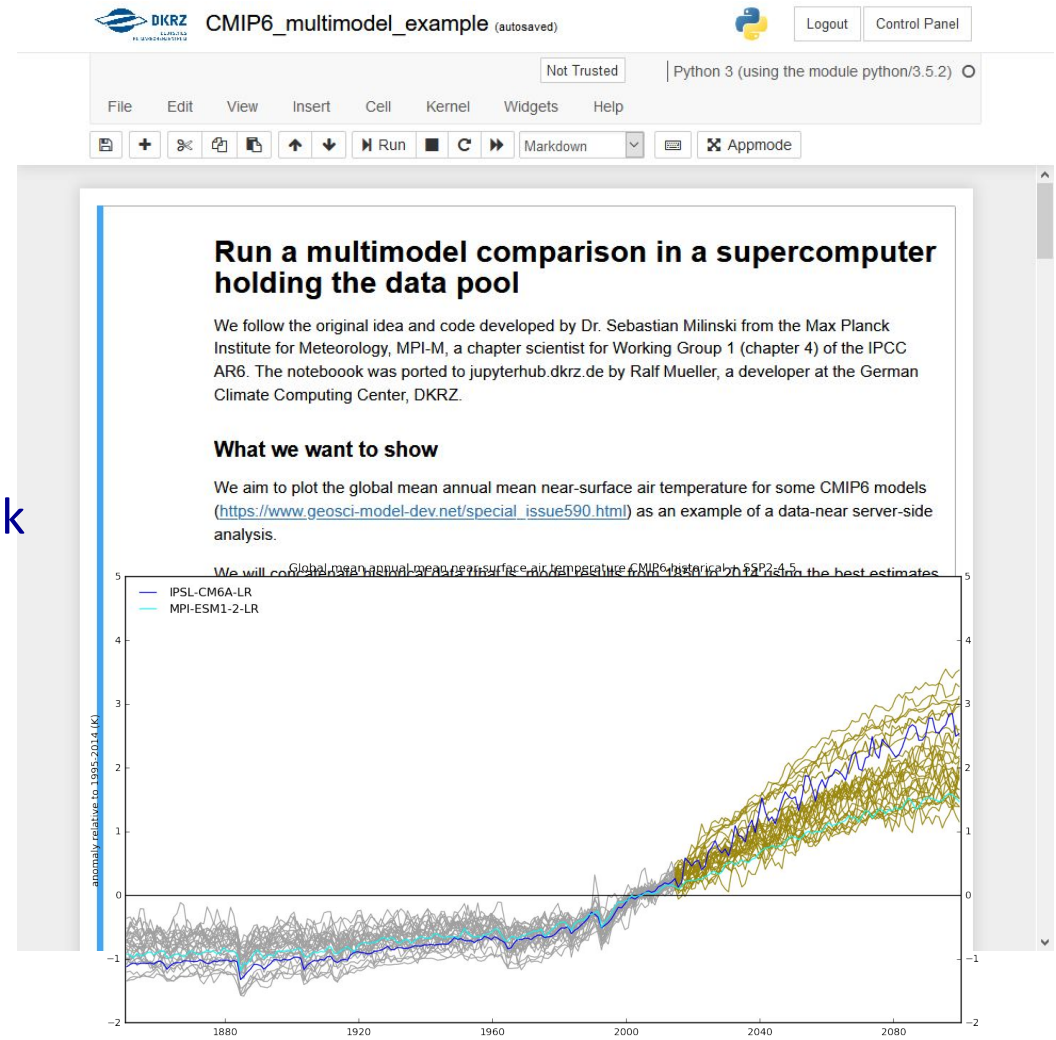
stats: http://bit.ly/CMIP6_DOI_Statistics

ENES CDI: processing services



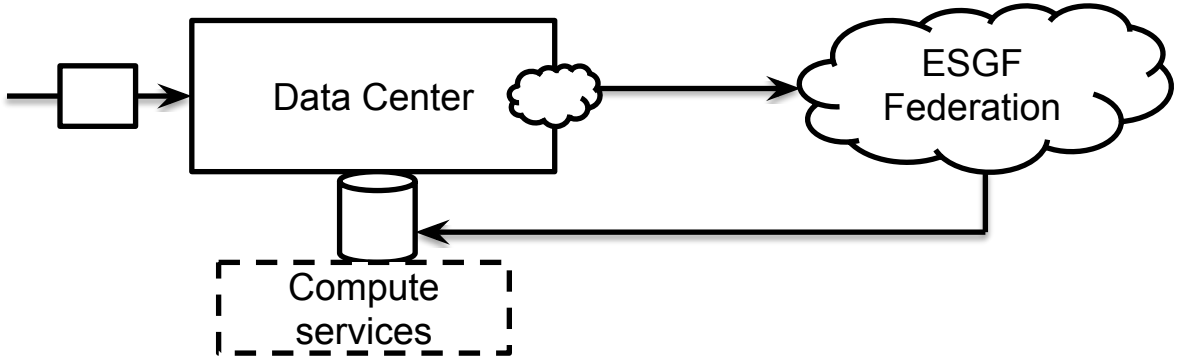
Supporting data near processing

- Prerequisit: Replication of CMIP6 collections into Multi-PByte disk pools (coordinated replication strategy)
- Different processing approaches:
  - Hosting of data near VMs
  - Jupyter-hub installations
  - OGC Web Processing Services (WPS)
    - also for interaction with Copernicus CDS
  - Direct access to virtual workspaces



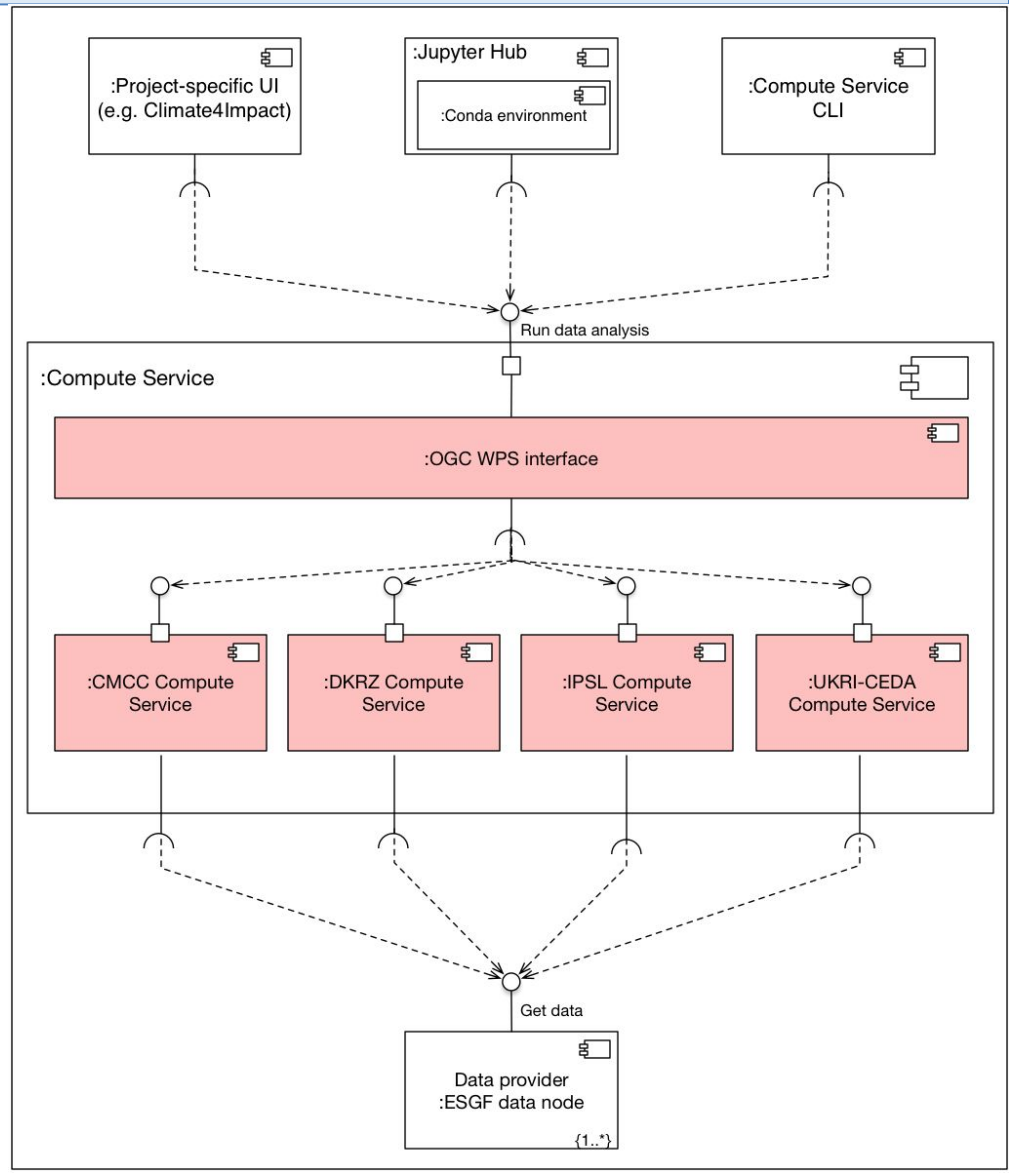
The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

ENES CDI: processing services



Supporting the diverse data near processing requirements

- data center organized data pools
- basic data reduction and transformation services as web processing services
- jupyter-hub installations
- dedicated portals/services for e.g. climate impact community (C4I portal)
- Coordinated service activities at European level (e.g. IS-ENES3, Copernicus)
- different, flexible compute backends (xarray, dask, ophidia, ..)
- ..

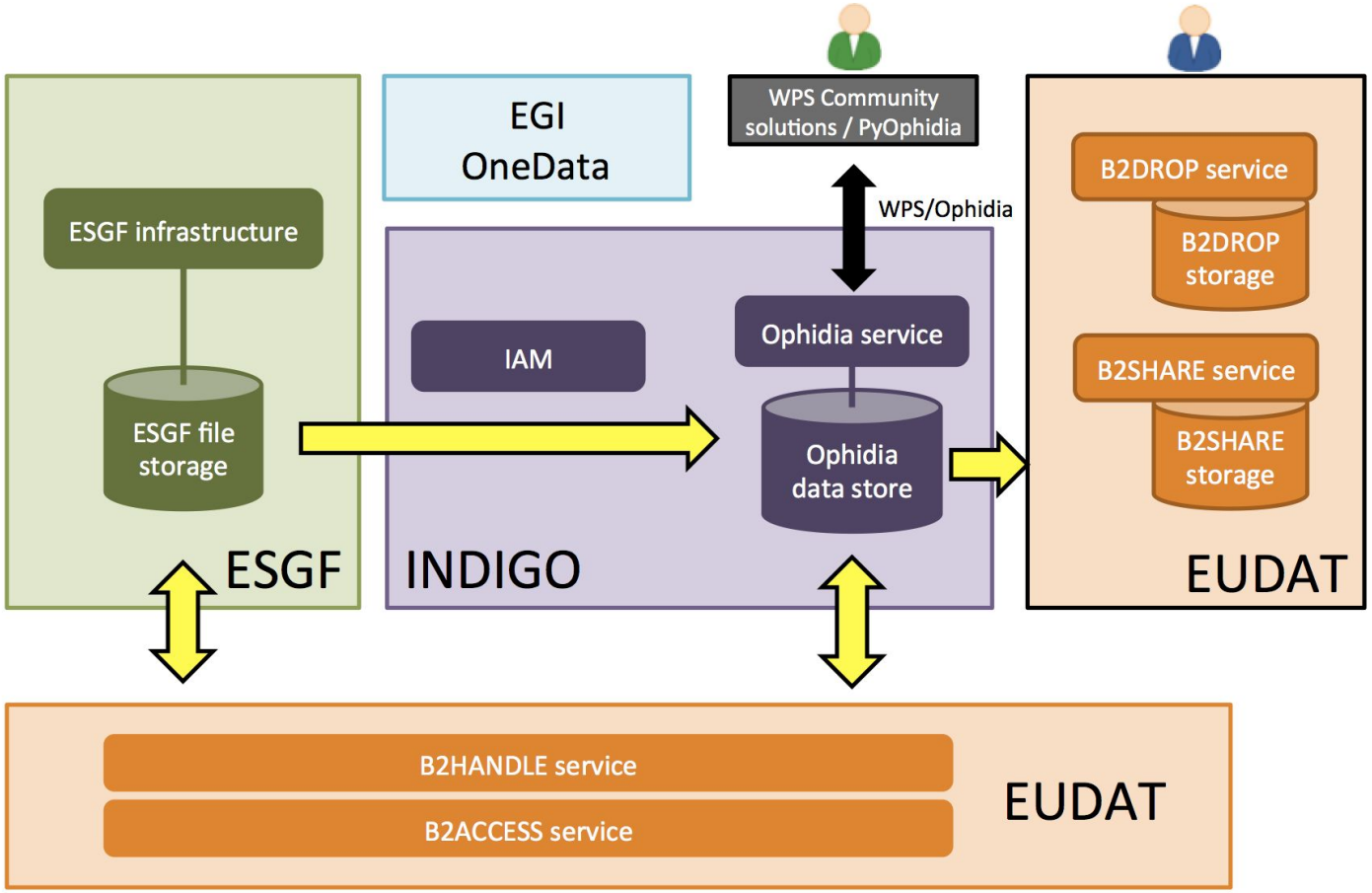


ENES CDI: ECAS service example

The ENES Climate Analytics Service (ECAS), proposed by CMCC & DKRZ in the EU H2020 EOSC-Hub project, supports climate data analysis experiments with a strong focus on data intensive analysis, provenance management, and server-side approaches

It is one of the EOSC-Hub Thematic Services as well as a Compute Service in the IS-ENES3 project

ECAS consists of multiple integrated components from INDIGO-DataCloud, EUDAT, ESGF and EGI, centered around the Ophidia HPDA framework



EOSC-hub receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 777536

ENES CDI: Future Perspectives

Moving towards cloud based solutions:

A) data hosting on the cloud

- google cloud hosting of CMIP6 subset (Pangeo community)
- AWS cloud prototype (US initiated, partners from IS-ENES contributing)
- institutional pilots
- data storage based on Zarr → "analysis ready data" → "analytics optimized data store"

→ future funding model(s) unclear

B) Individual infrastructural component hosting

- centralized search index
- centralized portal and Identity Proxy
- load balancing components (used e.g. to provide >99% availability services for Copernicus)
- …

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

THE CONSORTIUM

Coordinated by **CNRS-IPSL**, the IS-ENES3 project gathers **22 partners** in **11 countries**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°824084

Our website
https://is.enes.org/

Follow us on Twitter !
@ISENES_RI

Contact us at
is-enes@ipsl.fr

YouTube
Find our videos on our channel !

zenodo
Join our ZENODO community !

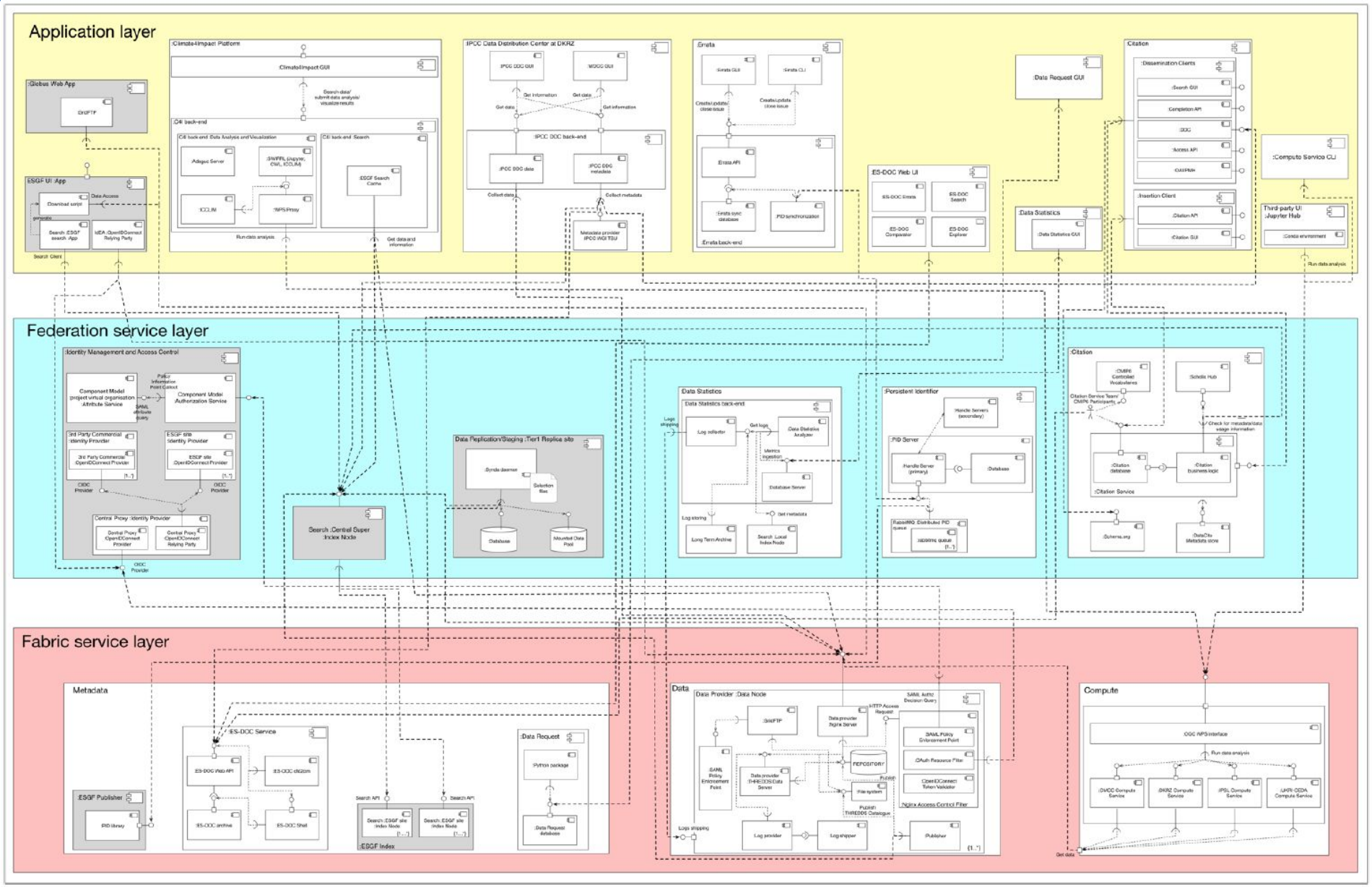
Additional slides

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

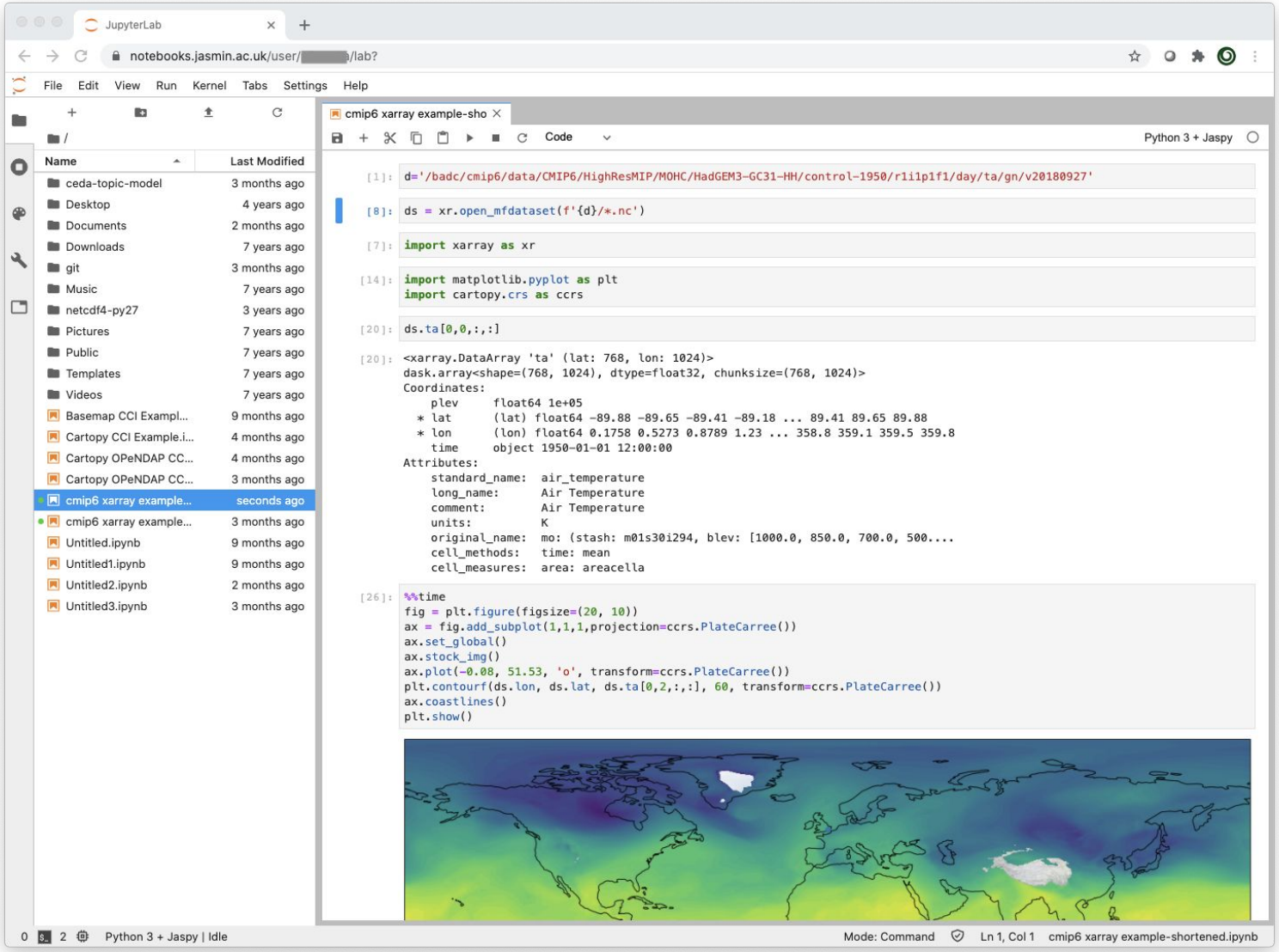
ENES CDI

Current architecture:

- complex dependencies
- intervoven partially maintained components



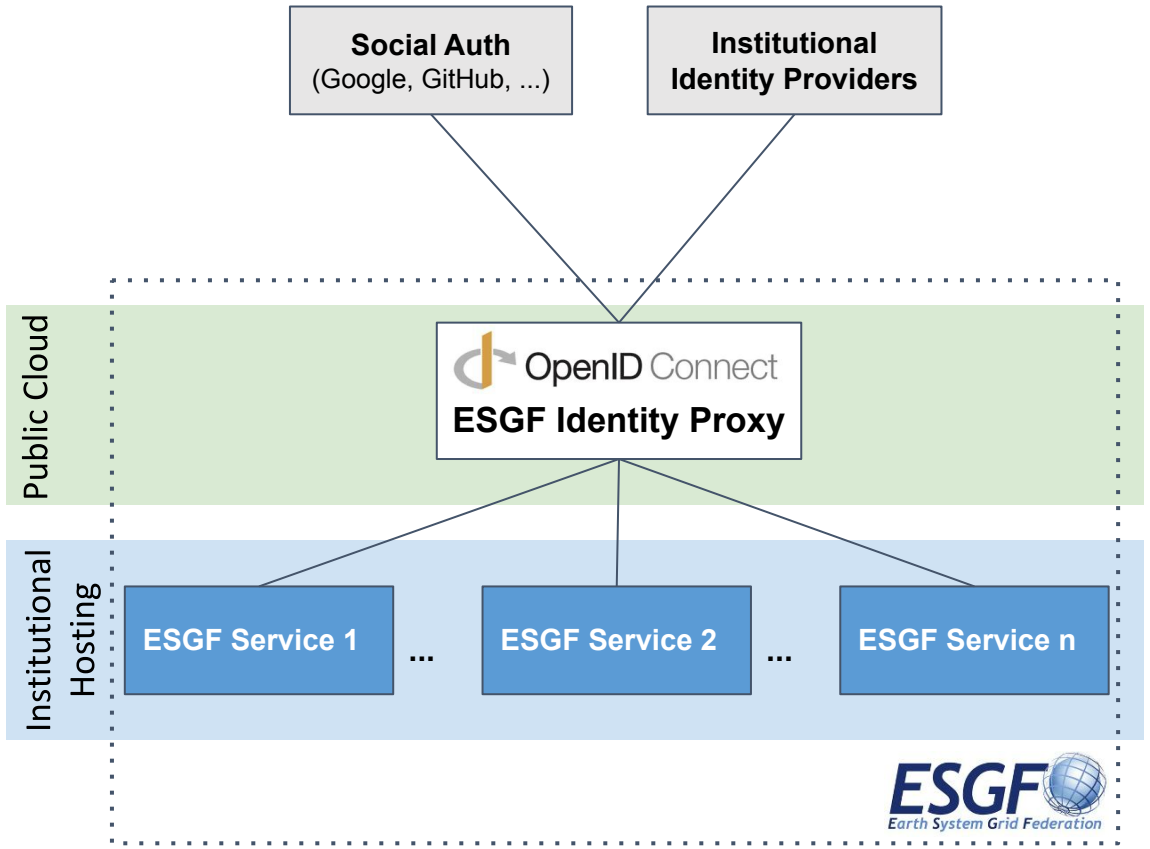
Jupyter Notebooks



- Jupyter Notebooks enable users to perform analysis of large data holdings at centres like DKRZ and CEDA (JASMIN facility)
- Avoids the need for “download and process at home” model
- Also provides potential as a training tool
- Notebooks could be a shared resource for the community demonstrating different analyses of data

ESGF Future Architecture - Examples of Planned Improvements: Identity Management

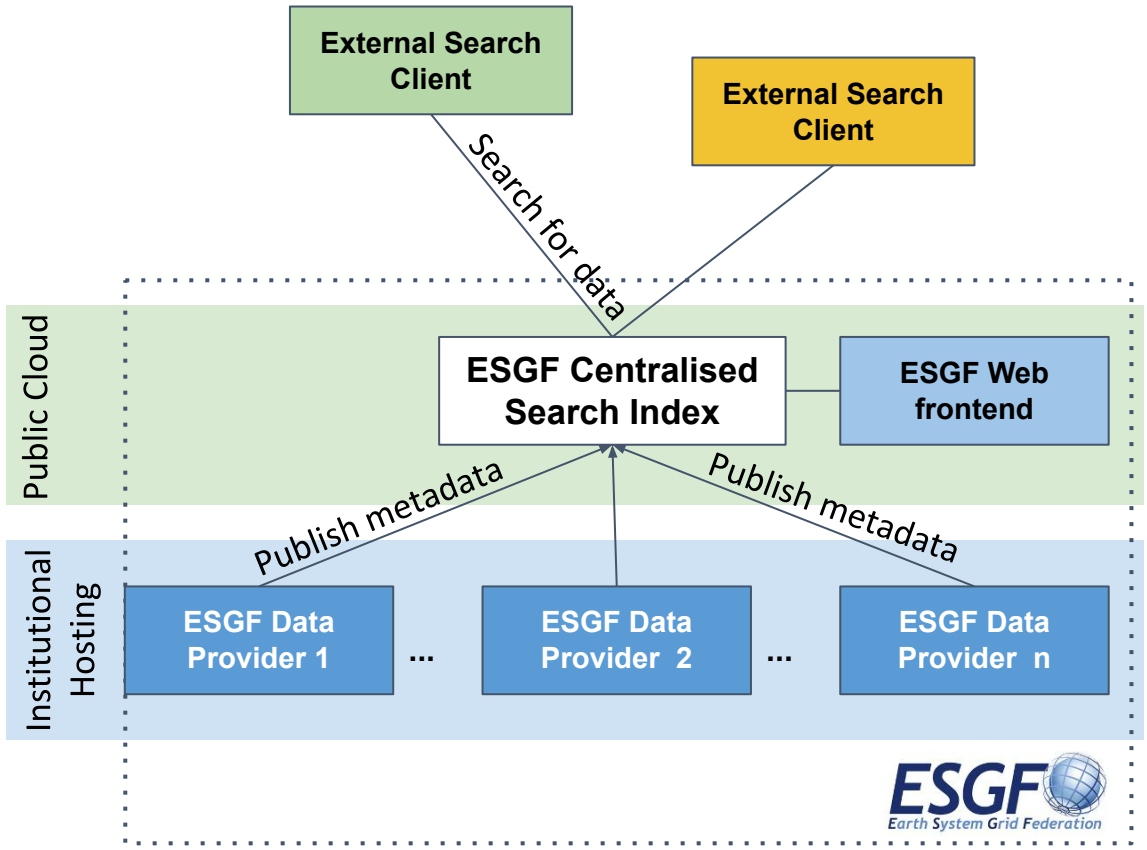
Identity Proxy Model



- Identity management and access control is needed in ESGF to secure access to restricted datasets and other resources
- The current system uses a many-to-many relationship between identity services and those services that are dependent on them
- The European AARC2 Project provides an alternative model
- External identity providers are proxied through a centralised Identity Proxy
- The Identity Proxy enables the presentation of a common interface to ESGF services
- It simplifies the interface between ESGF services and external identity providers

ESGF Future Architecture - Examples of Planned Improvements: Search Services

Centralised Search Service



- ESGF Search services provide users with the ability to find the model data they are interested in accessing
- The existing architecture for ESGF uses a system of distributed search services at hosting institutions in the federation
- This is flexible but complex to maintain
- A new proposal is for a centralised search service (index) for ESGF
- This would be hosted on public cloud to make it resilient
- Access is simplified for applications because they can go to the one overall service for search queries