



# ENES Climate Data Infrastructure: Data and Network – Status / Requirements / Challenges / Roadmap

ENES Data Infrastructure Providers:  
BADC/CEDA, CERFACS, CMCC, DKRZ, IPSL, LIU, KNMI, SMHI, UiB, ...

Frank Toussaint, Michael Lautenschlager  
On behalf of and with contributions from ENES Data Task Force  
<https://redmine.dkrz.de/projects/enes-data-task-force>

Revised Version: 17-02-16

# Content

- IS-ENES2 Achievements
- ENES CDI
  - Components
  - Requirements / Challenges
  - Institutional Plans
  - Inter-institutional Plans
- Development Examples
  - ESGF CMIP6 Services
  - Climate for Impacts Portal
  - Ophidia
- Future Perspectives

## IS-ENES2 Achievements

A major outcome of the IS-ENES2 project with respect to data is the establishment of the ENES CDI (Climate Data Infrastructure). The ENES CDI

- was implemented for CMIP5 under IS-ENES2 funding.
- is operationally continued for CMIP6 after IS-ENES2 with institutional and national funding.
- is coordinated by the ENES DTF (Data Task Force)
- has strong weight in the international ESGF development, maintenance and operation as well as in the CMIP6 data management.
- future development is presently planned in a number of EU proposals in the EOSC (European Open Science Cloud) framework.

# ENES CDI: Components

## Technical Components:

- Data Nodes
  - Associated (large) storage pools (disk + tape) , long term archival facilities
- Portals
  - Metadata, Search, User-Management, Documentation
  - Higher level services including specific processing services (Climate4Impact)
- Network Infrastructure
- Processing Resources (currently not exposed/connected in an infrastructure, yet need to do so)

## Software Components:

- ESGF software stack (data publication, data node, portal)
- Data preparation and data quality control (“cmorization”, CMIP / CF compliance, ..)
- Data replication and data cache maintenance (“synda”)
- Model / experiment related metadata tools (“es-doc”)
- Data related metadata tools (“errata”,..)
- Data processing tools and toolboxes (“cdo”, “downscaling”, “icclim”, “WPS/WMS”, ..)
- Workflow-enabled big data analytics frameworks (“Ophidia”)
- Data identification and data citation infrastructure and tools (“PIDs”, “DOIs”,..)
- Processing support frameworks

# ENES CDI: Components

## Non Technical Components:

- ENES Data Task Force
- ENES / ESGF partnership
  - Substantial ESGF infrastructure development done by European ENES partners
  - Representation in ESGF Executive Committee
- ENES / WCRP WIP participation
  - Substantial WIP work done by ENES partners
- Cooperation with / involvement in other data infrastructure projects
  - EUDAT
  - Envri+
  - ICNWG (international network infrastructure team)
  - RDA
  - INDIGO-DataCloud
  - Copernicus
- Funding streams
  - Future European data infrastructure plans (H2020, open science cloud, EINFRA-12, ..)
  - Future national/institutional plans

# ENES CDI: Requirements / Challenges

## A) Hosting future high volume data projects (CMIP6 ... )

- To support data analysis large amounts of data have to be collected and made accessible centrally for data processing at „tier1“ sites (e.g. national data caches)
  - Future overall data collection requirements (e.g. to support multi model evaluations) will exceed local storage capacities
- 
- ENES wide data replication and data collection strategies
  - Sharing data storage responsibilities
  - High bandwidth interconnections between ENES tier1 sites (and to international tier1 sites)
  - ENES wide data management plans and policies
  - New tools and services to support cross-institutional data management (overall research object (collection) management versus POSIX file/dir + MD management)
  - Follow current / future data object storage and storage cloud approaches

# ENES CDI: Requirements / Challenges

## B) Data processing and data evaluation support

- Especially tier1 sites need to support „data near processing“
  - (Large) compute cluster co-located to large data caches
  - Hosting of compute services alongside data services  
(use by specific portals like climate4impact or Copernicus or by end users directly)
- 
- ENES national compute islands versus a distributed compute infrastructure  
(grid experiences → many non technical obstacles .. !!)
  - New code packaging and deployment and hosting solutions (virtual machines on demand, docker, cloud services – new things for HPC centers ..)
  - Code management besides data management:  
(code quality assurance, packaging, plugin interfaces, .. )
  - Sustainable compute infrastructure ???  
ENES / ESGF strong enough for this ? → collaboration etc. with other efforts !?

# ENES CDI: Requirements / Challenges

## C) Research object management

- ENES data used over long periods of time and also in interdisciplinary contexts (→ underlying technology changes, volatile http references, metadata ! ..)
  - Foundation needed for sustainable data management services (e.g. persistent data identification independent of file systems, http references, ..)
  - Data / metadata / collections → research objects and RO management services
  - Support for RO object versioning, replication, provenance, etc. needed
- ENES / DKRZ driven PID infrastructure for CMIP6 is only the first (small) step
- PID management services integration into current data management practices at data centers needed
  - multi faceted challenge to establish sustainable RO management infrastructure in ENES and beyond
- RDA Europe Call Application (CMCC) for Collaboration Project on RDA PID recommendation adoption (collaboration with DKRZ)
- Ongoing collaborations in context of EPIC, EUDAT, RDA → H2020, EINFRA-12, ... ?!



# ENES CDI: Requirements / Challenges

## D) Data and Metadata: data infrastructure ingest

- Data homogenization („cmorization“) process not only for model data
  - Flexible data quality control tools
  - Metadata collection about the data generation context (model, experiment) („es-doc“,..)
  - Metadata and DRS standards extensions to climate indices/indicators, other observation data types, and level-2+ processed products
- Flexible community toolset to support groups in data homogenization needed (status: individual cmor or cmor like local adaptations – no real community effort .. many groups e.g. CORDEX start from scratch )
- Flexible community toolset to support groups in data quality control (work started, integration of cf-check, cmor-check, mip-checks ..)
- Es-doc toolset sustainability

# ENES CDI: Institutional Plans

## DKRZ:

- PID, LTA, Data citation, EUDAT, RDA, Replication, ....

## BADC:

- Security infrastructure (ESGF AAI), data replication (ICNWG), ....

## CERFACS:

- Build on C4I/KNMI portal/services: EUDAT/EGI/RDA (data analytics, e.g. cloud computing + interfacing APIs between infras, workflows, data life cycle), still seeking H2020 appropriate consortium/funding to pursue

## IPSL:

- Multi Model Analysis, replication (synda based), analysis workflow, ES-DOC, Errata (PID based), EOSC

## KNMI:

- Build on and support C4I with CERFACS/SMHI/WUR/CMCC/...; Host ESGF data node; PerfSonar; ESGC CWT; Work on provenance and indicator metadata;

## CMCC:

- CMCC Provides a Tier2 site. It is also committed to develop and maintain the ESGF modules related to the download metrics (IS-ENES2 funded activity). Since 2010, CMCC has been also developing a framework for big data analytics (Ophidia) in close synergy with ESGF. A dedicated cluster will be made soon available for the ESGF-CWT activity. Plans relate to apply to H2020 and Copernicus calls on data-infrastructure topics.

## LIU:

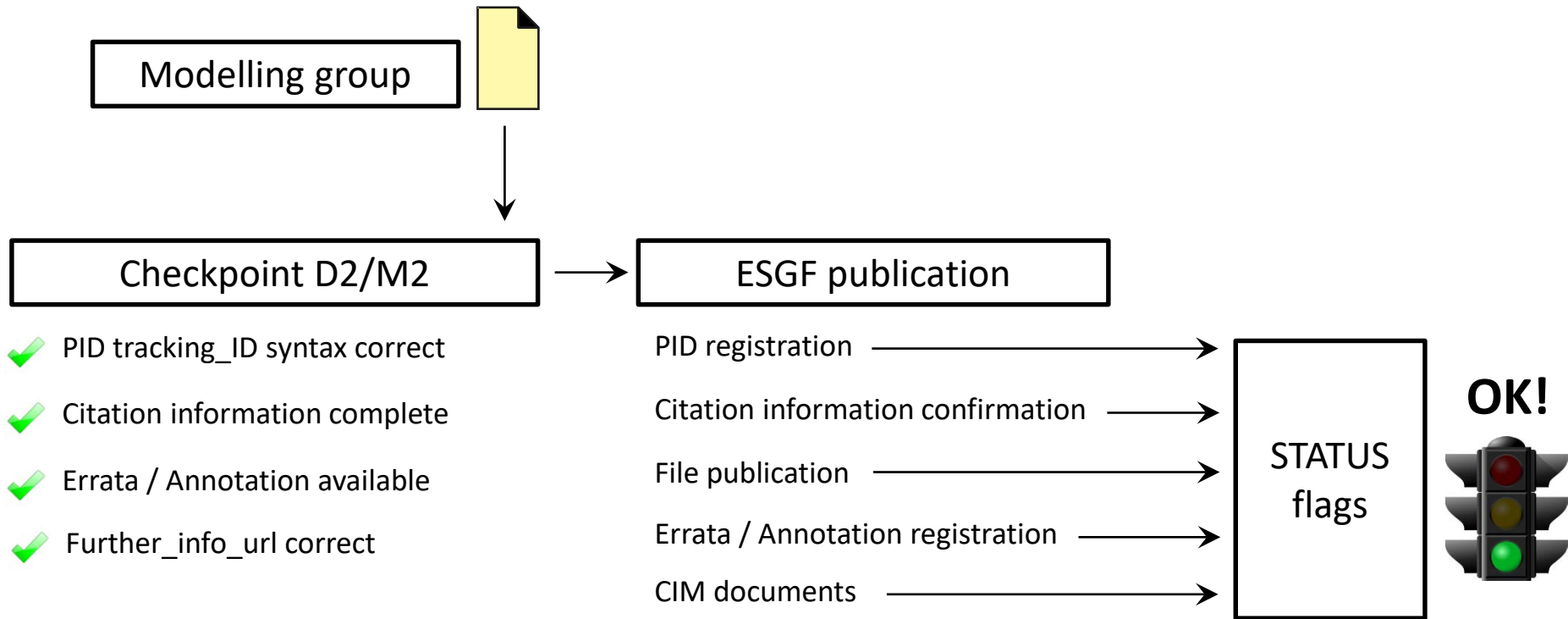
- Operate and develop ESGF and MARS infrastructure. Nordic support for ESGF. Operate and deploying iRODS, hosting PID framework services. Involved in EGI, NeIC and ECDS

# ENES CDI: Inter-Institutional Plans

CMIP6 ENES cross institutional infrastructure developments in e.g. ESGF

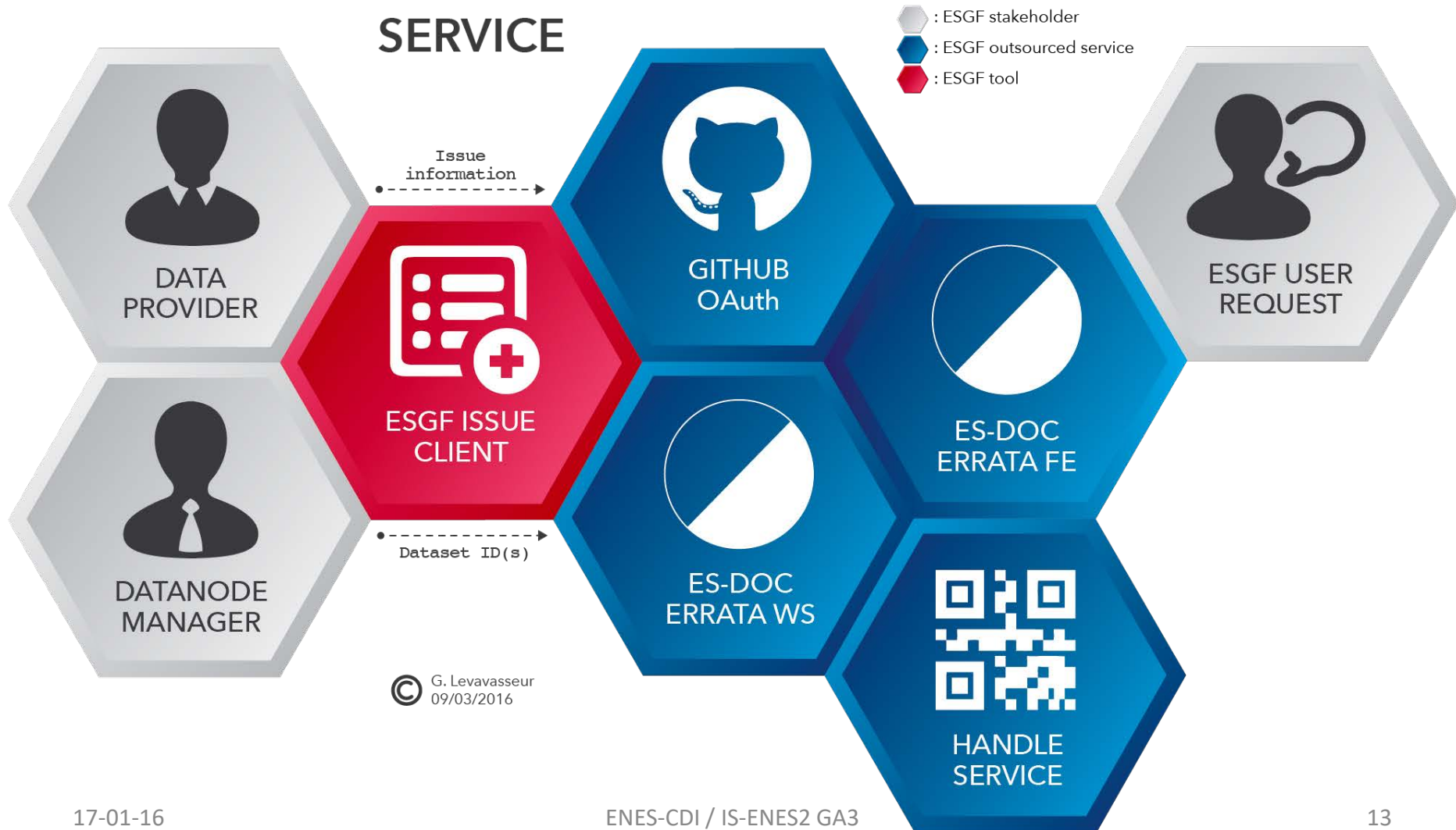
- Replication
- Data network bandwidth
- PID infrastructure
- (Early) data citation
- Errata management
- ES-DOC/CIM and emerging new Metadata & DRS for level-2+ products
- Data provision and processing: ENES & Copernicus; adapt current platform (C4I) and services to new datasets and experiments
- User Support: after IS-ENES2 ??

# ESGF CMIP6 Data Publication

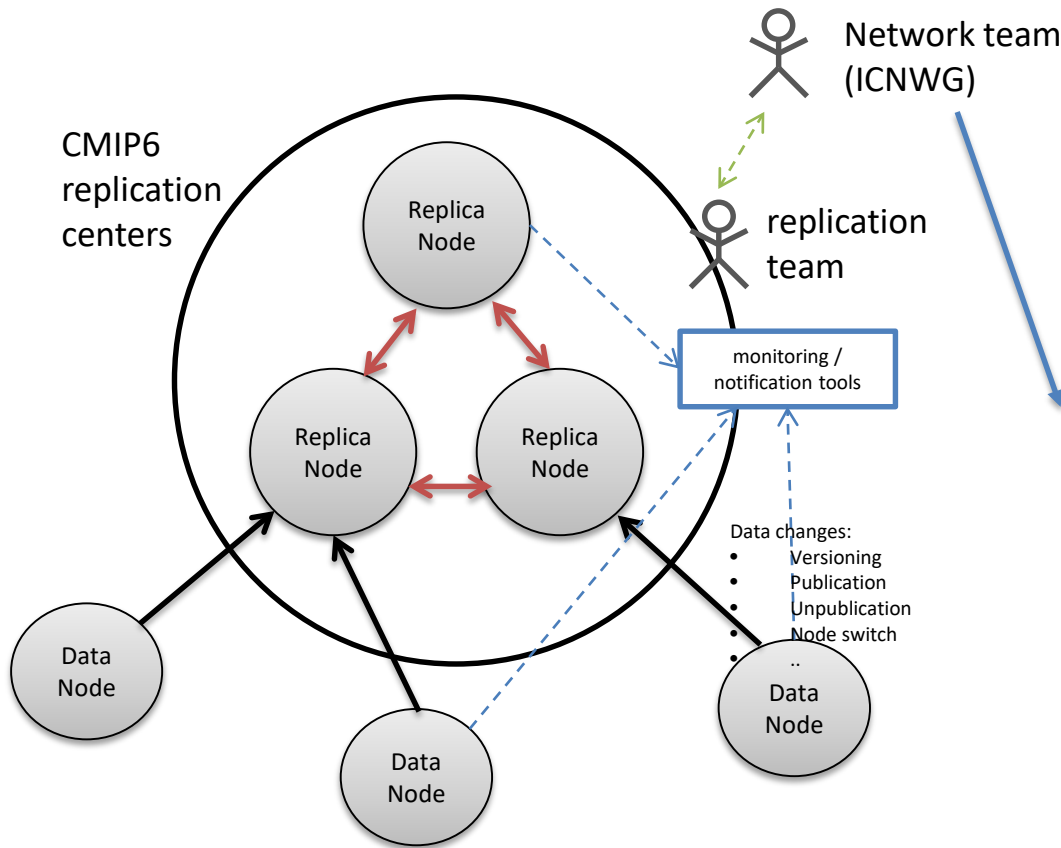


# ESGF CMIP6 ES-DOC Errata Service

## CMIP6 ERRATA SERVICE



# ESGF CMIP6 Data Replication



## CMIP6 data volume

- 20 – 30 PB compressed NetCDF-4
- WIP suggestion: 2 PB core data for replication

## Network bandwidth (ICNWG)

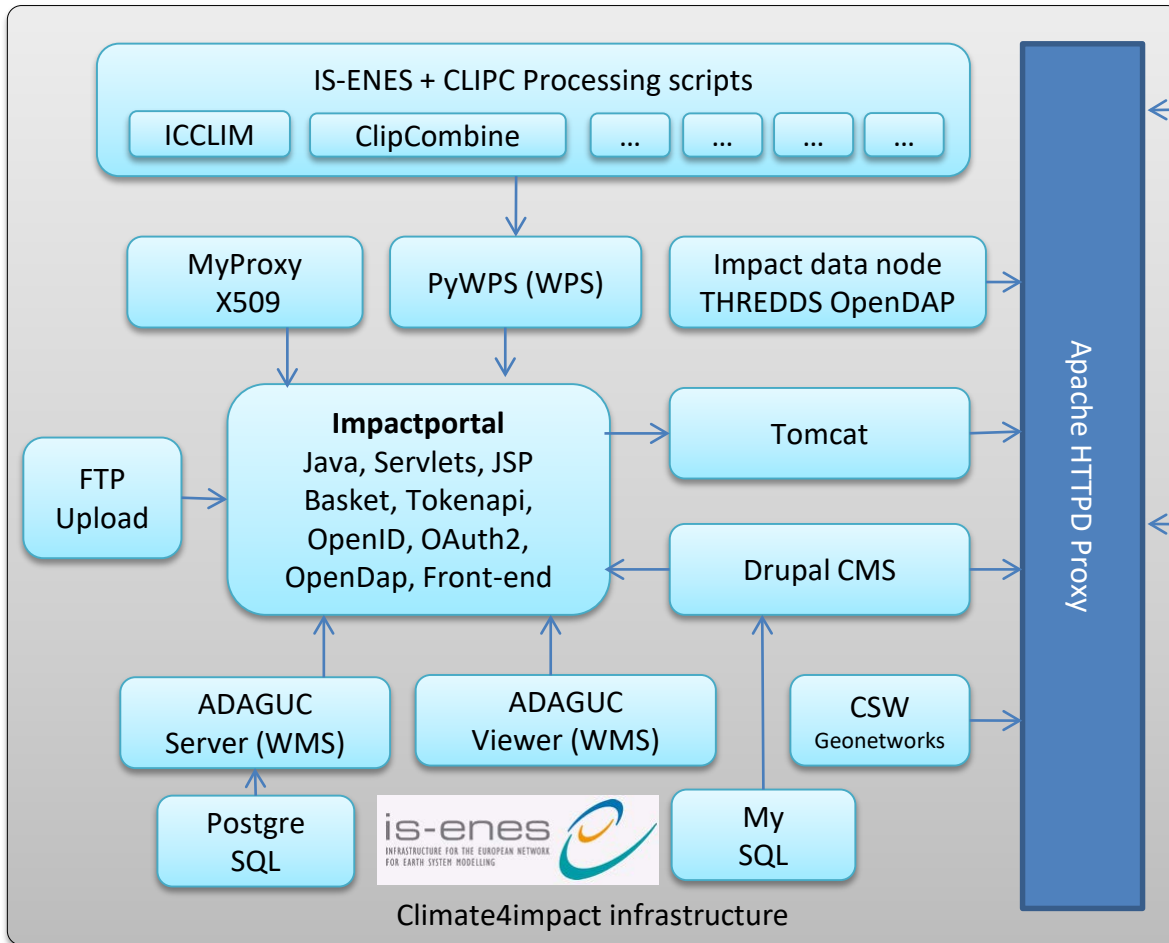
- Theoretical: 10 Gbit/s
- Real: about 1 Gbit/s (local – local)
- Projection: 3 – 5 Gbit/s

## Full replication of 2 PB core data

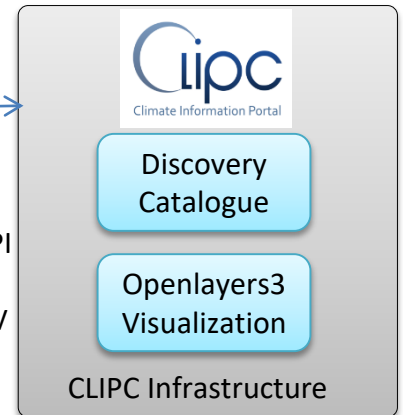
- Real: 26 weeks
- Projected: 9 – 5 weeks

# Climate for Impacts Portal (C4I)

## Climate4impact backend services

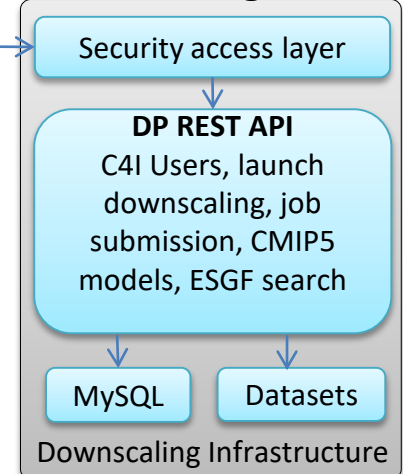


## CLIPC frontend



WMS  
Access token  
WCS  
WPS  
CSW  
OpenDAP  
Storage API (basket)  
W3C PROV

## Downscaling Portal

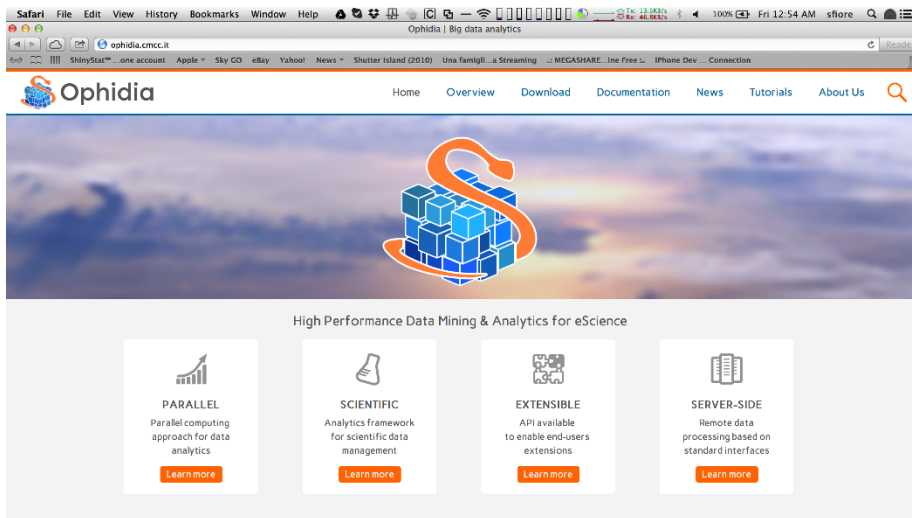


Access Token

# Workflow-enabled big data analytics – Ophidia

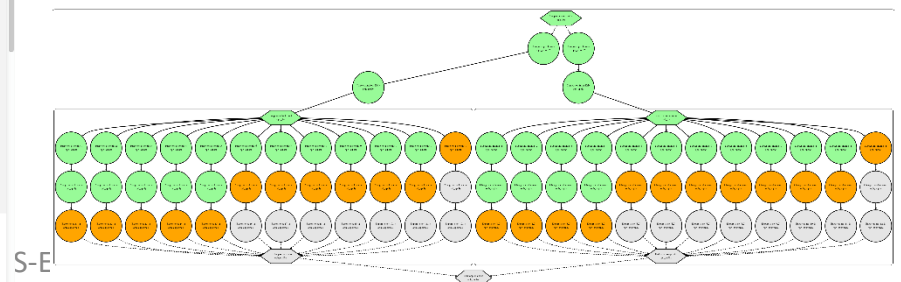
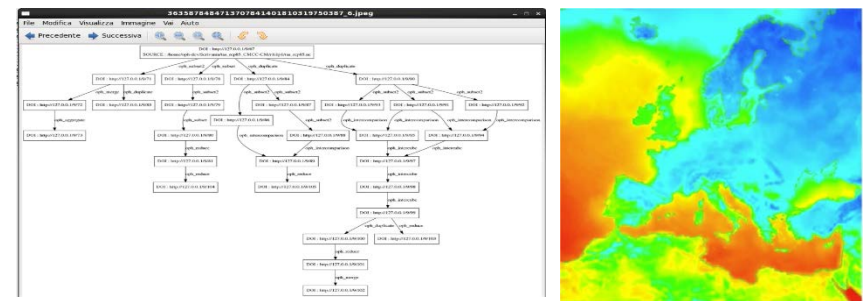
Ophidia is a research project addressing big data challenges for eScience

- It provides support for declarative, parallel, server-side data analysis exploiting parallel computing techniques and database approaches
- It supports end-to-end analytics workflows for eScience
- It has been exploited in **CLIP-C** to process climate indicators in some big data scenarios and in the **INDIGO-DataCloud** to implement distributed, multi-model analytics experiments



The screenshot shows the Ophidia website interface. At the top, there's a navigation bar with links for Home, Overview, Download, Documentation, News, Tutorials, and About Us. Below this is a large banner image with the Ophidia logo, which consists of a stylized 'S' shape made of blue cubes. Underneath the banner, the text reads "High Performance Data Mining & Analytics for eScience". There are four main feature cards: "PARALLEL" (Parallel computing approach for data analytics), "SCIENTIFIC" (Analytics framework for scientific data management), "EXTENSIBLE" (API available to enable end-users extensions), and "SERVER-SIDE" (Remote data processing based on standard interfaces). Each card has a "Learn more" button.

Ophidia is a CMCC Foundation research project addressing big data challenges for eScience. It provides support for data-intensive analysis exploiting advanced parallel computing techniques and smart data distribution methods. It exploits an array-based storage model and a hierarchical storage organisation to partition and distribute multidimensional scientific datasets over multiple nodes. The Ophidia analytics framework can be exploited in different scientific domains (e.g. Climate Change, Earth Sciences, Life Sciences) and with very heterogeneous sets of data.





## ENES CDI: Future Perspectives

Even after end of the IS-ENES2 project the ENES CDI will be coordinated by the ENES DTF in order to:

- Strengthen ENES data infrastructure “mission statement” and “identity” which is needed to define contour in context of emerging science cloud etc. trends
- Integration into EOSC (European Open Science Cloud)
  - EU pushing the integration of EUDAT, EGI and INDIGO to aim for an operational EOSC by 2020
  - ENES is represented in EUDAT by CERFACS, DKRZ, MPI-M
  - ENES is represented in INDIGO-DataCloud by CMCC
  - Direct integration of ENES partners at institutional level in COPERNICUS, E-INFRA 12A and E-INFRA 21
- Preserve Europe’s leading role in ESGF development, maintenance and operation
- Develop strategies to handle huge data volumes (CMIP6: 20 – 30 PB) together with huge numbers of data entities (CMIP6: 50 – 150 Mio)
  - Sharing of data storage responsibilities between European data nodes
  - Federated data processing at storage locations
    - Orchestration of data analytic workflows using data processing near storage, cloud computing (EGI, ...) with local temporary storage (EGI DataHub, ...), using standard interfaces (ESGF API, EUDAT GEF API, ...)
  - Higher level services including specific processing services and intelligent search facilities (Climate4Impact)
  - Network bandwidth on demand
  - Replacement of classical file system by cloud storage of digital objects
  - Alternative data storage and evaluation strategies to adapt data production to IT-prospects