IS-ENES3 Milestone M5.2

ESGF CMIP6 Summary

Reporting period: 01/01/2019 – 30/06/2020

Authors: Michael Lautenschlager, Stephan Kindermann,
Reviewer: Philip Kershaw, Sandro Fiore
Release date: July 2020

ABSTRACT

This report provides a summary of the status of the ESGF involvement from the ENES partners to establish the operational CMIP6 ESGF infrastructure.

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

Table of contents

1. Objectives

The IS-ENES partners have an essential role in the development and evolution of the international ESGF data infrastructure. WP5/NA4 will define the European priorities to be brought into the ESGF working groups as well as committees.

- Coordinate all IS-ENES/ESGF contributions
- ENES/ESGF project management (CNRS-IPSL, DKRZ, STFC)
- Replication strategy (CNRS-IPSL, DKRZ, STFC)
- User support organization (1st, second Level etc.), coordinate European contribution to overall ESGF support (DKRZ, KNMI, LiU, CERFACS)
- Metrics collection consolidation (CMCC).

2. Description of work: Methodology and Results

The coordination of IS-ENES/ESGF contributions is shared between all IS-ENES3 data work package partners. All partners are involved in development and maintenance of specific data infrastructure parts and services. Governance and coordination at the European level is performed in the ENES-DTF (Data Task Force) and the implementation is organized in the IS-ENES3 cross data work package meetings. At the international, trans-European level infrastructure governance, maintenance and implementation is coordinated in the ESGF-XC (Executive Committee). Additionally ENES partners are leading the ESGF operations team (CDNOT) which coordinates the worldwide operations of the ESGF infrastructure in support of CMIP6.

The IS-ENES3 RP1 was dominated at the international, trans-European level by an intensive discussion over the future architecture for ESGF and the specification of the roadmap with short-term (summer 2020) and medium-term goals (beyond summer 2020). The current ESGF architecture design is more than 10 years old and problems in maintenance and performance, security issues and the emergence of new technologies drive the needs for its revision.

The review of the architecture was first initiated in September 2018. A first document was prepared to gather input from members of the ESGF community involved with the technical operation and development of the system. This was presented at the Face-to-Face meeting in Washington DC, December 2018. Following from this, the Executive Committee put plans in place for a dedicated meeting in 2019 to take forward proposals for a new architecture. This was held at Milton Hill House, Steventon in the UK in November 2019. More details can be obtained from the parallel Milestone M5.1 "Draft Architecture Design".

The meeting was organised along these guiding principles: keep to a small technically focused meeting (around twenty invited representatives from the community participating); flexible agenda with no planned presentations in order to maximise time to discuss issues and make decisions; hold a series of pre-meeting telcos in advance to prepare the groundwork. These centered on four high-level topics:

- User experiences
- Data repository and management
- Compute on data
- Platforms and system administration

The discussion results in a high-level architecture diagram for the next generation of the ESGF data infrastructure.

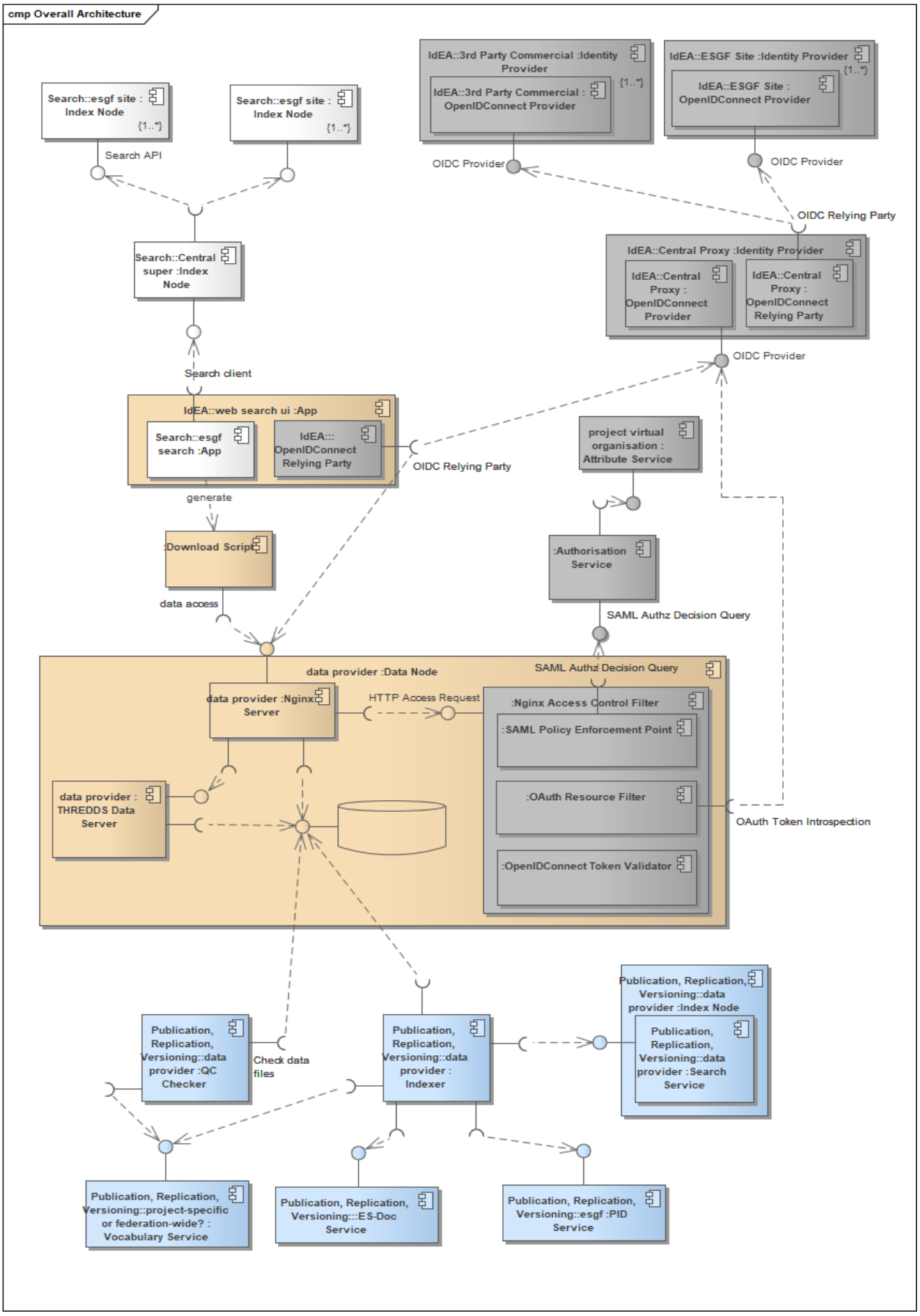
cmp Overall Architecture
IdEA::3rd Party Commercial :Identity Provider
IdEA::3rd Party Commercial : OpenIDConnect Provider
{1..*}
IdEA::ESGF Site :Identity Provider
IdEA::ESGF Site : OpenIDConnect Provider
{1..*}
Search::esgf site : Index Node
{1..*}
Search::esgf site : Index Node
{1..*}
Search API
OIDC Provider
OIDC Provider
OIDC Relying Party
Search::Central super :Index Node
IdEA::Central Proxy :Identity Provider
IdEA::Central Proxy : OpenIDConnect Provider
IdEA::Central Proxy : OpenIDConnect Relying Party
OIDC Provider
Search client
IdEA::web search ui :App
Search::esgf search :App
IdEA::: OpenIDConnect Relying Party
OIDC Relying Party
project virtual organisation : Attribute Service
generate
:Download Script
:Authorisation Service
data access
SAML Authz Decision Query
data provider :Data Node
SAML Authz Decision Query
data provider :Nginx Server
HTTP Access Request
:Nginx Access Control Filter
:SAML Policy Enforcement Point
data provider : THREDDS Data Server
:OAuth Resource Filter
OAuth Token Introspection
:OpenIDConnect Token Validator
Publication, Replication, Versioning::data provider :Index Node
Publication, Replication, Versioning::data provider :Search Service
Publication, Replication, Versioning::data provider :QC Checker
Check data files
Publication, Replication, Versioning::data provider : Indexer
Publication, Replication, Versioning::project-specific or federation-wide? : Vocabulary Service
Publication, Replication, Versioning:::ES-Doc Service
Publication, Replication, Versioning::esgf :PID Service

Fig. 1: ESGF high-level architecture showing components for identity management, data discovery, access and publication

At the European level, IS-ENES3 data work packages partners discussed and specified the ENES-CDI (Climate Data Infrastructure) as the European contribution to ESGF. Grey boxes show the ESGF services exploited in the ENES-CDI that are associated with collaborative development efforts carried out with partners outside Europe. White boxes correspond to components of the ENES-CDI developed by EU institutions exclusively. This specification is documented in the IS-ENES3 milestone M10.1 and more extensively in the deliverable D10.1 and it shows the ENES-CDI architecture (Fig.2):

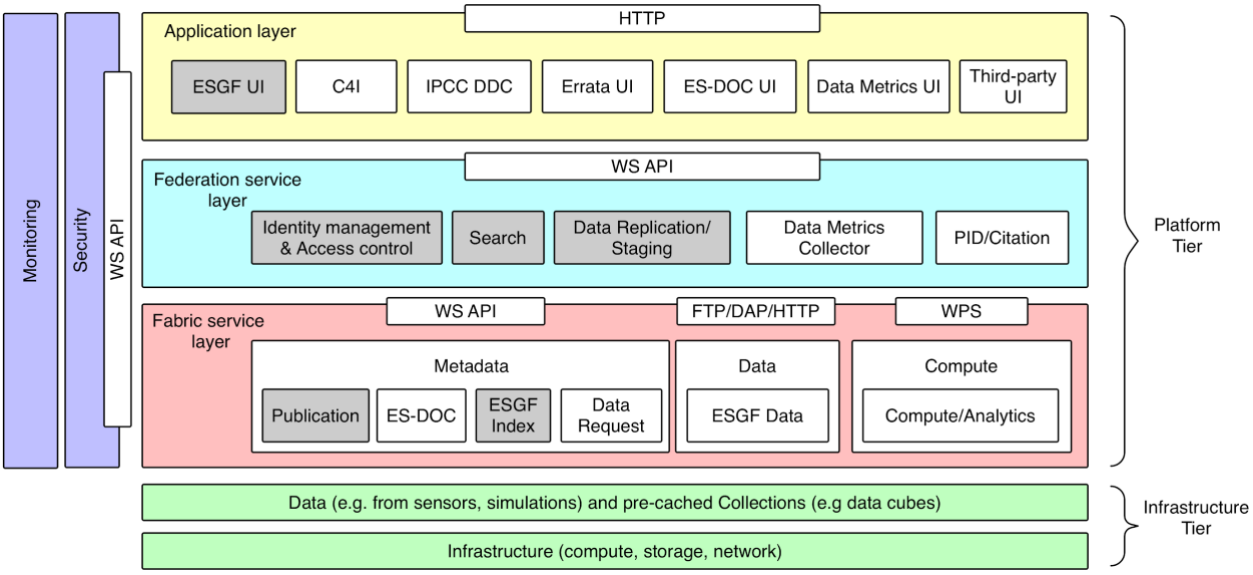

Fig. 2 ENES CDI software architecture

This diagram has been presented and discussed at the 1st IS-ENES3 GA and is used as one of the core diagrams in the IS-ENES sustainability specification process.

In addition the data infrastructure discussions and specifications IS-ENES3 has been active in CMIP6 data management and ESGF data integration. In close cooperation between ESGF-XC and the WIP (WCRP Infrastructure Panel) the CDNOT (CMIP Data Node Operations Team) has been established to oversee and to organise the CMIP6 data management in ESGF. Requirements from two major user classes have to be mapped: the climate model data providers and the climate model data consumers. The European coordination of CDNOT activities is discussed across the IS-ENES3 data work packages at the operational level and in the ENES-DTF at conceptual levels.

CDNOT has been chaired by an IS-ENES partner from the beginning an activity report was recently submitted for publication[1]

An ESGF-wide CMIP6 data replication strategy turned out to be more complicated than for CMIP5 because of the expected large amount of data volume and individual data entities (about 10 times larger than for CMIP5). The replication strategy needed to take into account different (sometimes conflicting) requirements and constraints:

- The available network bandwidth between tier1 data nodes (replica centers) as well as between tier1 and tier2 sites. Thus e.g. often data was first replicated by a single tier1 site and then replicated from there to other tier1 sites.
- The local storage capacities. The capacities of local data pools at tier1 sites needed to be dynamically adapted to the local needs of specific tier1 associated user communities as well as overall data replication requirements (for data access resilience and load distribution).
- Data consistency requirements across tier1 sites required coordinated activities to synchronize e.g. data de-publications.

Three ENES partners host and operate large CMIP6 data replica pools (tier1 replica sites) and coordinate their replication activities to satisfy and balance the needs of different user groups and requirements:

- local user communities (mostly from the climate modeling context) European and international climate data analysis research (e.g. supporting IPCC WGs and climate evaluation activities)
- international requirements to provide replicas for data access resilience and load balancing, especially providing replicas in Europe besides replicas hosted at the node at LLNL (USA).

User support for ESGF/CMIP6 is shared between IS-ENES3 and LLNL for the first level support. User requests arrive through the ESGF-users mailing list as for CMIP5. First level supporters pass the requests to ESGF working teams for technical questions and to the WIP for CMIP6 content related questions for more depth second level support. A shared support knowledge base providing FAQs and dedicated support instructions are hosted and maintained on Github. A system for automatic recording of user requests has not yet been installed.

The metrics collection consolidation has been performed at ESGF data nodes for CMIP6. The former ESGF dashboard has been replaced by the ESGF Data Statistics Service (http://esgf-ui.cmcc.it/esgf-dashboard-ui/index.html) developed and disseminated by CMCC.

[1] R. Petrie, S. Denvil, S. Ames, G. Levavasseur, S. Fiore, C. Allen, F. Antonio, K. Berger, P.-A. Bretonnière, L. Cinquini, E. Dart, P. Dwarakanath, K. Druken, B. Evans, L. Franchistéguy, S. Gardoll, E. Gerbier, M. Greenslade, D. Hassell, A. Iwi, M. Juckes, S. Kindermann, L. Lacinski, M. Mirto, A. Ben Nasser, P. Nassisi, E. Nienhouse, S. Nikonov, A. Nuzzo, C. Richards, S. Ridzwan, M. Rixen, K. Serradell, K. Snow, A. Stephens, M. Stockhause, H. Vahlenkamp, and R. Wagner, "Coordinating an operational data distribution network for CMIP6 data", Geoscientific Model Development (GMD) [submitted].

The new metrics system performs logs gathering and indexing from the distributed data nodes, jointly with a centralized big data processing pipeline set up at CMCC, which analyzes the collected logs and derives a wide set of statistics. Such statistics provide multiple and aggregated views about the data usage and publication across the federation, as well as insights about the most downloaded data by variable, experiments, sources, etc. on specific projects (i.e. CMIP5, CMIP6, CORDEX).

Description in details on how the milestone has been achieved:

To support CMIP6, the deployment, integration and operation of the following IS-ENES partner developed extensions was coordinated:

- Data statistics service: better understanding of the amount of downloaded data, the most downloaded/used datasets and the data published in the federation
- Data citation service: DOI assignment for data referencing in literature
- Data identification service: PID assignment to files and datasets for long term data reference persistence and versioning support
- Model documentation (ES-DOC): documentation of model configuration as well as data errata information
- Data replication: management of large parallel data transfers to and from ESGF replica centers

The IS-ENES/ESGF coordination has so far shown the following key contributions from IS-ENES partners in ESGF. These are critical for the future sustainability of the infrastructure:

- operational deployment of ESGF based on state of the art deployment technology (Docker, Kubernetes, Ansible, ..)
- ESGF replication software and organization
- “Core ESGF” extensions supporting data identification, citation and model documentation
- User support

3. Difficulties overcome

Due to the impact of the Corona virus pandemic, the organisation of discussions and coordination of activities is exclusively by telco and VC. No F2F meetings are planned, at least up to the end of 2020.

Specification and agreement of infrastructure funding opportunities incorporates three funding streams: European project money, national project money and institutional contributions. None of these streams is funded on a long-term (more than 3 years) perspective. With respect to the IS-

ENES sustainability discussion and the ability of the European partners to commit to ESGF a long-term funding confirmation is required.

4. Next steps

Next steps follow the IS-ENES3 DoW.

- Direct exchange between LLNL/PCMDI and IS-ENES3 will be considered in more detail and under the Coronavirus constraints.
- A priority list of involvement areas will be summarized during RP2 in context of the ESGF future architecture roadmap[2] and the IS-ENES3 sustainability process. The specification of involvement areas is seen as critical to the future sustainability of the ESGF effort and European ESGF infrastructure.

2 http://doi.org/10.5281/zenodo.3928223