IS-ENES3 Deliverable D5.4

IS-ENES3 involvement in ESGF

Reporting period: 01/07/2020 – 31/12/2021

Authors: Stephan Kindermann (DKRZ), Katharina Berger (DKRZ)
Guillaume Levavasseur (IPSL),
Paola Nassisi (CMCC),
Philip Kershaw (UKRI)

Reviewer(s): Michael Lautenschlager, Christian Pagé
Release date: 31/12/2021

ABSTRACT

IS-ENES partners play a central role in the international ESGF data federation. They are involved in ESGF activities at all levels: architecture, management, operations, future developments etc. This document will summarize the contributions from IS-ENES3 partners in the international ESGF effort. Additionally priorities of involvement agreed on in the ENES Data Task Force and communicated in ESGF are summarized. These priorities correspond directly to the priorities associated to the roadmap for the last year of IS-ENES3 contributing to the future ESGF roadmap.

Revision table			
Version	**Date**	**Name**	**Comments**
V.0.1	01/11/2020	Stephan Kindermann	Initial version: structure and contribution collection
V0.2	25/11/2020	Phil Kershaw, Michael Lautenschlager, Guillaume Levavasseur, Paola Nassini	completion of sections
V0.9	1/12/2020	Stephan Kindermann	draft submitted to reviewers
V1.0	10/12/2020	Stephan Kindermann, Michael Lautenschlager, Christian Page	final version

Dissemination Level		
PU	Public	X

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

Table of contents

Inhalt

Executive Summary

Overall ENES Climate Data Infrastructure (CDI) topics and issues are discussed and agreed in the ENES Data Task Force which has regular meetings once a month. This includes reviewing, defining and prioritizing the involvement of IS-ENES3 partners in the ESGF. The involvement in ESGF includes all levels and can be roughly structured in:

- Future vision and preparing for the future: the ESGF involvement not only needs to ensure the current stable operation of the infrastructure but must early on adapt to the new and emerging technological landscape.
- Coordination board memberships: ENES partners lead and contribute to different boards ensuring the close coordination of the internationally networked activities and agreeing on a shared understanding of issues and priorities as well as available funding streams.
- Core infrastructure development and operations: ENES partners contribute as part of different working teams to the stable operation of the existing ESGF infrastructure, currently concentrating on operationally supporting CMIP6 data distribution and refactoring the ESGF software stack as well as prepare for the provisioning of associated compute services.

Key aspects and the priorities with respect to the engagement levels characterized above are summarized in the following report. Because the ENES CDI is strongly relying on a close collaboration of well established large climate compute and data centers in Europe the engagement needs to take into account the individual institutional viewpoints as well as ongoing and future European data infrastructure efforts (EOSC and GAIA-X related, Digital Twin Earth and data lake as well as data space initiatives etc.). Thus an agreement on a shared architectural understanding and associated priorities and the inclusion of these in the overall ESGF roadmap is of key importance for the evolution of the ENES CDI.

1 Objectives

The ESGF data infrastructure is based on an international collaboration which strongly relies on a close cooperation of US partners and European partners. The European partners coordinate their priorities, plans and operations as part of IS-ENES3. This coordination is defining the roadmap and the developments towards a stable and robust as well as a technologically sustainable future ESGF data infrastructure in Europe, which respects the individual requirements of the large European climate data centers. As part of this the European ESGF roadmap is elaborated, it is defined and communicated as part of the IS-ENES Work Package WP5/NA4.

The involvement of IS-ENES3 in ESGF is organized on multiple levels (e.g. architectural, technical, organizational, operational) and performed as part of multiple working groups and different boards (e.g. the WIP, the ESGF Executive Committee (XC) and the ESGF Scientific Board (SC) at the international level, described in section 2 and the ENES Data Task Force at European level). This deliverable summarizes these different forms of engagement and thus builds on the earlier activities reported as part of Milestone M5.2 in the previous reporting period.

In the following the activities are structured and summarized along the following key areas:

- **Architectural design and future roadmap:**

  This aspect concentrates on work to define a common European roadmap with respect to their ESGF infrastructure and the associated architectural design of the infrastructure reflecting the European priorities and requirements.

- **Working groups, boards and collaborations:**

  This aspect concentrates on the involvement in the different working groups, boards and collaborations which was needed to bring forward and influence the existing and future ESGF infrastructure according to the European priorities and requirements.

- **Operations**

  This aspect summarizes all the work done to enable the stable operations of the different installations constituting the European part of the global ESGF data federation. As the operational work of this part is closely related to the overall ESGF operations, IS-ENES3 partners were centrally involved in coordination of overall ESGF operational aspects (e.g. user support as well as data node manager coordination).

2 Description of work: Methodology and Results

2.1 Architectural design and future roadmap

The architecture of current ENES CDI as outlined in the Deliverables D10.1[1] , D10.2[2] and upcoming D10.3[3] illustrates its dependency on an internationally agreed on and maintained ESGF software stack. Defining a new architectural design and an associated roadmap to implement this needs to take into account the fact that the current ESGF software stack is on the one hand in heavy operational use. Equally though, with ten years since the original establishment of ESGF, it was recognised that there was a strong need to modernize and restructure the SW stack. A major review was conducted in 2019[4] and the *Future Architecture* was initiated directing efforts towards the implementation of a more modular design to enable greater flexibility for deployment and ease of maintenance in operations. Additionally, the adoption of community standards for interfaces to enable ESGF to better interoperate with other infrastructures in the environmental sciences (e.g. infrastructures organized in ENVRI community[5]).

The core ESGF architecture as agreed on by the ENES CDI partners is illustrated in Figure 1. This architecture defines an evolutionary step of the current system, apart from the identity management system, core interfaces and service components remain unchanged. Instead the focus is on support of stable, sustainable service deployment possibilities based on containers and container orchestration tools.

[1] IS-ENES Deliverable D10.1 "Architectural Document of the ENES CDI software stack" https://zenodo.org/record/4309892#.YaDSHNDMJD-

[2] IS-ENES Deliverable D10.2 "First release of the ENES CDI software stack" https://zenodo.org/record/4450012#.YaDSc9DMJD-

[3] IS-ENES Deliverable D10.3 "Second release of the ENES CDI software stack" - draft, will be published alongside the other deliverables in Zenodo and the project repository https://is.enes.org/documents/deliverables

[4] https://doi.org/10.5281/zenodo.3928222

[5] ENVRI community: https://envri.eu/

Key changes to highlight are:

- the adoption of OpenID Connect[6]/OAuth 2.0[7] and OPA[8] in the identity management and access control system to build a future proof AAI layer based on open standards enabling wide interoperability.
- revisions to the ESG Publisher and THREDDS Data Cataloguing to simplify the data registration and file serving parts (e.g. using Nginx) of the ENES CDI infrastructure.

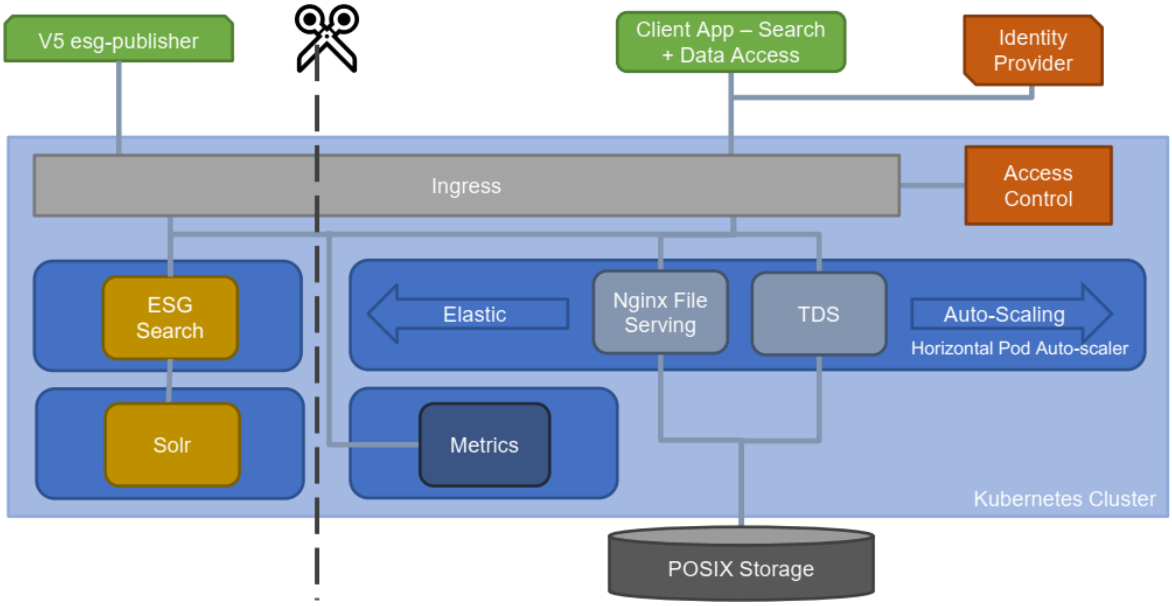

Figure 1: First evolutionary step of future ESGF architecture reflecting the requirement for modern service deployment methods (based on containers and container orchestration).

The vertical dashed line in Figure 1 provides an indicator for the next evolutionary step of the ESGF architecture illustrated in Figure 2. This involves a complete replacement of the current search system based on ESG Search and Apache Solr, replacing it with the STAC API[9] as the search interface and ElasticSearch for the backend search index. It is also important to note that the new search system is being written in such a way as to be able to catalogue data held on object-store as well as POSIX file systems. Object storage is seeing more widespread use alongside the increased uptake of cloud for service hosting.

A key requirement expressed in the Future Architecture meeting by technical representatives from the ESGF partners was a desire to move away from a model where many individual sites maintain

[6] https://openid.net/connect/
[7] https://oauth.net/2/
[8] https://www.openpolicyagent.org/
[9] The SpatioTemporal Asset Catalog API: https://stacspec.org/STAC-api.html

Index Nodes. The Index Node is the component in ESGF which provides search services. In Phase 2 therefore, a more centralised approach is proposed whereby search services are hosted at only a few locations in all likelihood using public cloud. The expectation is that few centralized well coordinated search service installations which are based on modern (e.g. kybernetes based) deployment methods will provide a more stable service delivery then the existing distributed installations based on less coordinated site specific deployment scenarios.

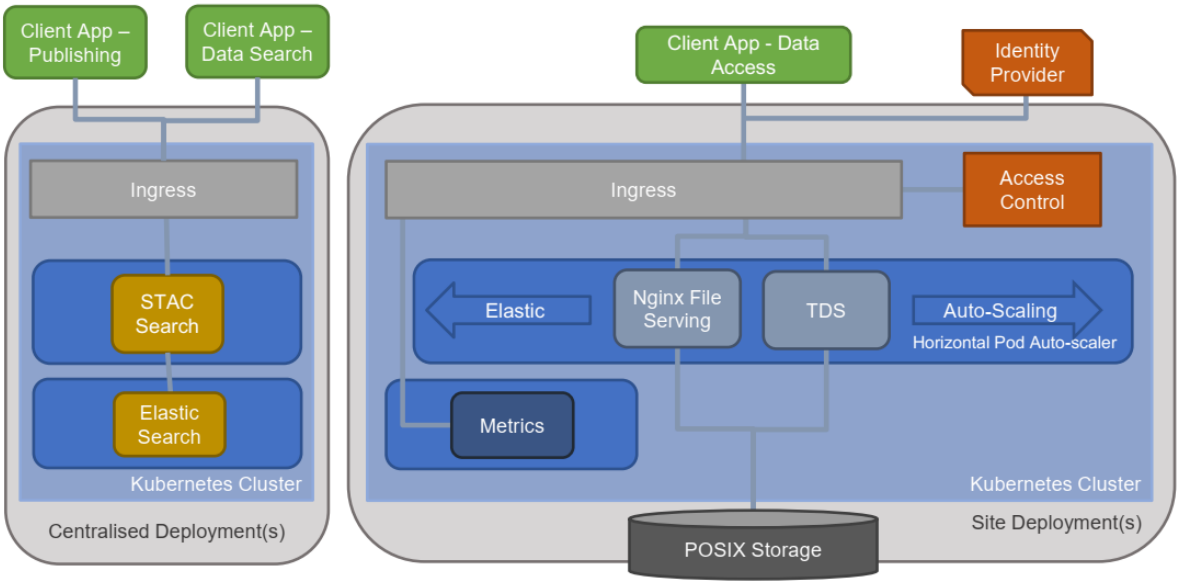

Figure 2. Final ESGF architecture supporting flexible deployment scenarios by separation of components and supporting standard interfaces.

Given the challenge mentioned earlier of updating an operational system based on internationally distributed and closely interdependent service deployments, the European partners agreed on the following roadmap towards realizing this new architecture:

Phase 1: Rollout future architecture 1 at European nodes

• Re-engineered version of ESGF – maintaining the same interfaces but with re-engineered and enhanced underlying infrastructure.
• Deployed on AWS (GFDL), CEDA (test), other European sites (beginning of year 2022).

Phase 2: (in parallel to Phase 1, STAC API integration testing with European partners)

• STAC Search replaces ESG Search (local test deployments).
• Integration testing with European partners next.
• Production STAC search service – end of year 2022.

This roadmap was communicated in the relevant ESGF boards (ESGF XC) and it was also part of the discussions ENES-DTF partners had with the different consortia preparing proposals for the new ESGF funding round in the US. IS-ENES3 will intensively collaborate with the new US ESGF partners to integrate these European developments into a future overall ESGF.
This roadmap also reflects the need to minimize the impact on operations and end user experience: the first phase does not change any end user facing interfaces, whereas the second phase concentrates on one central component of the infrastructure: the search interface - not influencing the core data distribution functions of the ENES CDI.

2.2 Boards and collaborations

In the following a summary is given with respect to the key engagements of IS-ENES CDI members in the different working groups, boards and collaborations.

CMIP data node operations team (CDNOT[10]): IS-ENES (DKRZ) currently chairs the ESGF CDNOT team together with LLNL as co-chair, which is described in detail in the "Coordinating the operational data distribution network" report published in GMD, 14, 629–644, 2021[11]. The primary goal of the CDNOT team was to support an operational ESGF infrastructure to support the CMIP6 data publication and distribution effort. It is planed to increase the coordination via CDNOT in the future to tackle the operational problems created by the need to modernize and update the ESGF infrastructure (see previous section). The focus of work in the reporting period was on supporting the CMIP6 data distribution based on the ESGF infrastructure, specifically:

- coordinating software updates at sites in close collaboration with the data node managers
- overseeing best practices of CMIP6 data publication
- resolving data search inconsistency problems (e.g. with respect to retracted data, which is still visible or inconsistent search results at different tier1 sites or with respect to data replicas)
- discussing operational data replication issues
- addressing end user facing data access problems at sites (oftenly at tier2 sites)
- addressing specific service problems e.g. with respect to the persistent identifier infrastructure
- coordinating data usage metrics collection across ESGF

[10] CDNOT terms of reference: https://wcrp-cmip.github.io/WGCM_Infrastructure_Panel/Papers/CDNOT_Terms_of_Reference.pdf

[11] Coordinating an operational data distribution network for CMIP6 data: Geosci. Model Dev., 14, 629–644, 2021 https://doi.org/10.5194/gmd-14-629-2021

- set up policies for ESGF Tier1 and Tier2 sites and make sure those policies are satisfied over time
- work on documentation for ESGF node admins

The WGCM Infrastructure Panel (WIP)[12]: The WIP infrastructure panel was established to serve the WGCM and, more broadly, the WCRP community. It is responsible for data request and infrastructure development coordination for WCRP projects. As part of this it established the CDNOT team and also oversees the CMIP6 Data Request (lead by an IS-ENES partner - UKRI) , which specifies the variables that should be archived from each of the over 300 CMIP6 experiments (and defines the metadata associated with each variable). Other parts of the infrastructure are ES-DOC and the Citation Service. The membership list for the WIP is maintained on the WGCM Climate pages[13] and currently includes IS-ENES3 members from DKRZ, NCAS, STFC and SMHI. Important activity areas in the reporting period included:

- Agreeing and discussing key architectural future changes in ESGF. One example is support for a centralized ESGF index/search solution hosted on cloud.
- Approach towards standardization of Data Request and file naming conventions over WCRP projects, first targeted project is the harmonization with CORDEX (new WIP member).
- Coordination of CMIP data protocols with data providers (modeling centers) for data publication in ESGF.
- License/Citation questions around CC-BY SA data usage restrictions.

The ESGF Executive Committee (ESGF XC): Currently the ESGF XC has three IS-ENES3 members (DKRZ, UKRI, CERFACS) and is led by UKRI. Important coordination tasks in the reporting period included:

- Communicating the IS-ENES3 roadmap and future architecture developments.
- Coordinating the ESGF relation to commercial cloud service offerings (starting, especially in context with the ESGF/Amazon collaboration on CMIP6 data and compute service provisioning).
- Contact and discussion with the different consortia bidding for the new US ESGF2.0 funding, making sure the European perspective is reflected in their proposals.

[12] The WGCM Infrastructure panel: https://wcrp-cmip.github.io/WGCM_Infrastructure_Panel/
[13] WGCM Infrastructure Panel (WIP): https://www.wcrp-climate.org/wgcm-cmip/wip

ESGF working groups

ENES CDI members are heavily involved in different ESGF working teams. The key activity area in the reporting period was user support and documentation. Recently UKRI and DKRZ also re-initiated the Compute Working Team with a specific focus on stepwise establishing the IS-ENES3 WPS compute service at larger ESGF sites. Developments for this European compute service solution is currently co-funded by IS-ENES and Copernicus and provides also the current interface solution between ESGF and the Copernicus Climate Data Store.

Another relevant working group coordinated by ENES CDI members (CMCC) and co-chaired by LLNL is the ESGF Data Statistics working team. The key activity in this area has been the coordination and support related to the data usage metrics collection in ESGF. This activity has been instrumental to set up a production-level service (ESGF Data Statistics) which has tracked so far more than 700 millions of downloads concerning 24 data nodes across the whole ESGF federation.

2.3 FAIR data services and standards

The core ESGF data distribution infrastructure is enriched and accompanied by services and activities supporting data standards and FAIR data management contributed mainly by IS-ENES3 partners. The services and activities are described in detail in the "ENES CDI software stack" deliverables (see "Architectural design and future roadmap" section above) as well as the service report of the first reporting period[14]. As these ENES CDI services (CF standard, data request, errata service, PID and data citation infrastructure, ES-DOC) are tightly integrated into the core ESGF data distribution system intense collaboration and coordination continued in the reporting period especially with the US partners to support the CMIP6 data delivery and especially supporting FAIR data concepts. Operational issues and implications were discussed as part of CDNOT and also in the ESGF XC and WIP panel.

[14] IS-ENES Deliverable D7.1 "First KPI and TA report for the ENES CDI services"
https://is.enes.org/documents/deliverables/d7-1-first-kpi-and-ta-report-for-enes-cdi-data-services/view

3 Conclusions and Recommendations

3.1 Changing technological and organizational background

The agreement on shared vision and associated roadmap for the ENES CDI is of high importance given the change of US funding sources for ESGF starting in 2022 with new partner institutes and associated new context of US priorities. ENES partners demonstrated, in close collaboration with US partners (e.g. GFDL/Amazon), steps leading towards the new ESGF architecture, supporting modular and scalable cloud deployments. The agreed European decision to prioritize the development and deployment of a new standard-conforming search API that is based on STAC was pushed forward and communicated to the ESGF. The ENES CDI partners agreed to take a lead in this and invest in the development of a pre-operational system during the last year of IS-ENES3.

3.2 Recommendations

The collaboration between ENES CDI partners and the ESGF is on a good track for the last year of IS-ENE3 funding, thanks to their engagement and coordination as part of the ENES-DTF. Future collaboration will strongly depend on the outcomes and results of the IS-ENES sustainability efforts and the associated decisions on sustained organisational structures. The outcome of the IS-ENES sustainability scoping phase can be obtained from Milestone 2.3[15]. Shared use of cloud infrastructures (public and private) and the use of cloud technology to manage data lakes and data infrastructures is becoming more and more important. Yet the institutional and funding context US and European partners are working in are very different and need to be considered early on in the emerging future ESGF infrastructure. As part of the ENES-DTF partners already discuss and coordinate their involvements and possible perspectives with the emerging European Open Science Cloud. IS-ENES partners deployed an ENES Data Space that delivers an open, scalable and cloud-enabled data science environment for climate data analysis on top of the EOSC Compute Platform. On the other hand work integrating institutional cloud infrastructures has already started by ENES partners (e.g at DKRZ and STFC). On the US side collaborations with public cloud providers have started (e.g. Google and Amazon), to which ENES-CDI partners (especially STFC) contributed. Yet a shared understanding and roadmap in ESGF for the exploitation of these new possibilities and required associated organizational regulations needs to be established.

[15] IS-ENES3 Milestone 2.3 "Sustainability Scoping Document"
https://is.enes.org/documents/milestones/m2-3-sustainability-scoping-report/view