IS-ENES3 Deliverable D10.5

Final release of the ENES CDI software stack

Reporting period: 01/01/2022 - 31/03/2023

Authors: C. Pagé (CERFACS), G. Levavasseur (CNRS-IPSL), P. Nassisi (CMCC), A. Ben Nasser (CNRS-IPSL), K. Berger (DKRZ), M. Burman (DKRZ), D. Hassell (UREAD-NCAS), M. Juckes (UKRI), P. Kershaw (UKRI), S. Kindermann (DKRZ), A. Nuzzo (CMCC), A. Stephens (UKRI), A. Spinuso (KNMI), M. Stockhause (DKRZ), L. Bärring (SMHI)

Reviewers: S. Kindermann (DKRZ), S. Joussaume (CNRS-IPSL)

Release date: 20/04/2023

ABSTRACT

This deliverable illustrates the final release of the ENES Climate Data Infrastructure (CDI) software stack (software repositories, licensing information, change logs, links to technical documentation). We report on the update of the implementation of the ENES CDI services in regards to the requirements collected within the milestone M10.1 and the previous release (D10.3). This document describes the final version including the latest developments of the core data distribution services, climate4impact, ES-DOC, compute services, data request schema and tools for MIPs, and file metadata specifications.

Dissemination Level	
PU	Public

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

Revision table			
Version	**Date**	**Name**	**Comments**
Document Structure and Contributors	01/02/2023	Christian Pagé	Preliminary Structure
Contributions from partners	01/02/2023-09/03/2023	All authors	
Document formatting and last updates	13-18/03/2023	Christian Pagé, Stephan Kindermann, Alessandro Spinuso, David Hassell, Philip Kershaw, Alessandra Nuzzo, Ag Stephens, Paola Nassisi, Guillaume Levavasseur	Content contributions
Version to review	18/03/2023	Christian Pagé	Last inputs & final formatting
Submitted version	19/04/2023	Stephan Kindermann + Sylvie Joussaume + Christian Pagé	Final review

Table of contents

List Of Images

List Of Tables

Executive Summary

The ENES Climate Data Infrastructure (CDI) is an achievement of the IS-ENES project to support the climate modeling community, climate impact community as well as interdisciplinary research domains. The ENES CDI consists of a collection of stable and consistent services, software and metadata specifications, to sustain access, evaluation and analysis of high volume of climate model data from the international Coupled Model Intercomparison Project (CMIP) and the COordinated Regional Downscaling Experiments (CORDEX) simulations.

In D10.1 [1] we have provided the general architecture of the envisaged infrastructure, with technical expectations for the mid to long term implementation. These are made concrete in D10.2 [2] report, which provides the details on the progress made in the realisation of the architectural principles. The final architecture is described in D10.3 [6] report, which provides a final version of the software stack.

As an update, the document follows a similar organisation as the D10.3 [6] report. After the general introduction, 5 main sections depict the CDI. Section 2 provides the overview of the second release, providing updates of each component/software of the CDI architecture covering the final Reporting Period (RP3). All the components are illustrated in the following four sections, respectively addressing (i) core data services, (ii) metadata schemas and reference tools, (iii) gateways for dissemination and access to computational facilities, (iv) solutions for authentication and authorisation. In the final conclusions, we particularly highlight the main achievement that consolidated the core data services in RP3, that will need to be sustained after the end of IS-ENES3, especially regarding core data archival and replication. We also focus on the necessary deployment of the new Earth System Grid Federation (ESGF) stack release, including the recent technical choices for a federated identity and access entitlement (IdEA).

1 Introduction

The ENES CDI architectural design was presented through the D10.1 report "Architectural document of the ENES CDI software stack" [1] according to the overall IS-ENES project objective #3, which fosters the support for the exploitation of model data by both the Earth system science community and the climate change impacts community.

This document provides a final description of the CDI achieved with the final implementation of the requirements and architecture presented in D10.3 "Second release of the ENES CDI software stack" [6]. The document addresses the deliverable D10.5 "Final release of the ENES CDI software stack" of the IS-ENES3 project, within the WP10/JRA3 "ENES Climate Data Infrastructure software stack developments".

The work presented in the report is packaged in a comprehensive official release, which addresses the different capabilities of the ENES infrastructure, from Data and Metadata, to Computation and Dissemination services. Each component is described in respect to the progress made, the issues encountered, how these have been solved and what is the status at the end of the IS-ENES3 project, and how it will evolve. It shows what each component offers in the current release and how it connects and exploits the other capabilities of the infrastructure. We specify the means of access (eg. Available as a service / Software Package) and provide references to technical documentation and official repositories. Where applicable, deviations from D10.1 [1] or delays from D10.2 [2] expectations are highlighted together with appropriate justification.

2 CDI Release Overview

We provide here the updates of the ENES CDI Architecture and a summary of the software components which are available in this final release. Sections are broken down by component and where appropriate details of the development of new software and services are also described. As all of these components are under constant development the version in operations and in public releases may be different than the ones mentioned here.

2.1 Updates to the Architecture

The ENES CDI is organized into multiple tiers and layers, through which the distributed components of the architecture interact with each other to provide the ENES community with a comprehensive set of services related to data and metadata access, and tools for dissemination and computation capabilities. As described in the document D10.2 "First release of ENES CDI software stack" [2], the ENES CDI architecture is and will be continuously updated during the project lifetime, allowing for the new requirements gathered in WP5/NA4 and coming from the IS-ENES community, also including external initiatives at both European (e.g. EOSC, EGI/EUDAT, Copernicus, etc.) and International levels (e.g. ESGF).

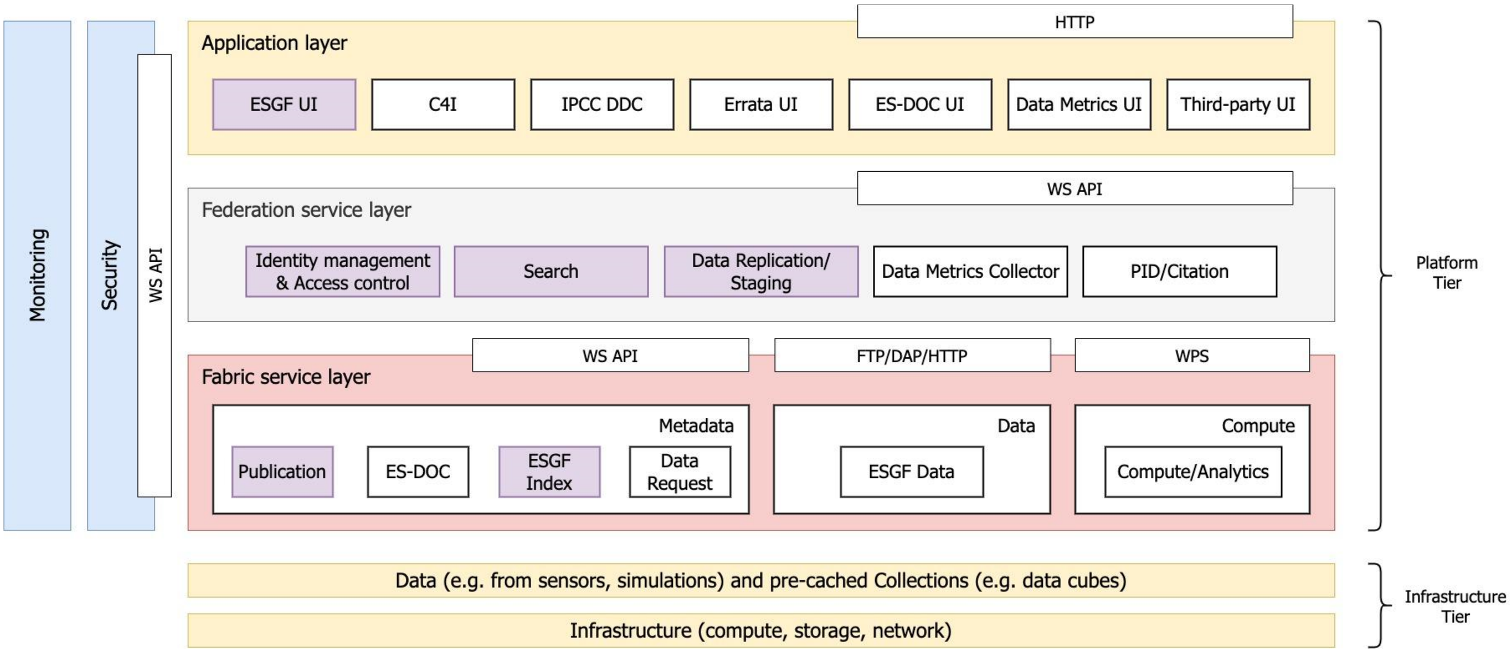


Figure 1. ENES CDI software stack architecture (purple boxes are ESGF services exploited in the ENES-CDI that are associated with collaborative development efforts carried out with partners outside Europe)

Figure 1 above depicts the ENES CDI layers. The document D10.2 [2] fully describes each layer and components, with changes being described in the following sections.

There have been no major updates to the architecture with respect to the second release of the ENES CDI software stack. Figure 2 below shows an updated version of the UML (Unified Modeling Language) component diagram of the ENES CDI software stack. There have been no major updates to the architecture with respect to the first version presented in document D10.2 [2]; except for the Identity Management and Access Control components according to interactions changes described in section of this document.

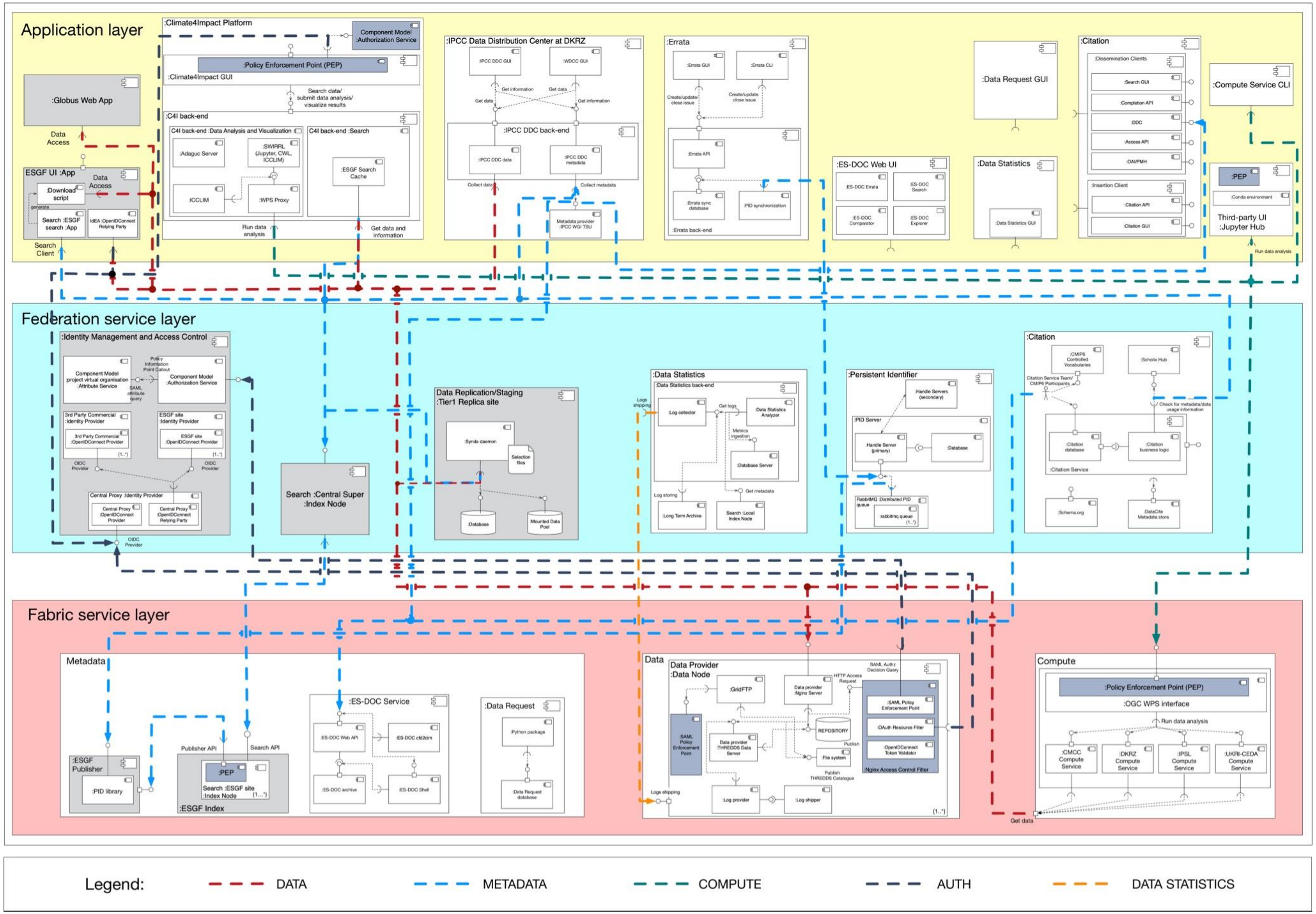


Figure 2. Updated component diagram of the ENES CDI architecture.

2.2 Integrated Available Software and Services

We present here an overview of the software available in the current release. We include access URLs (where the component is deployed and available as a service), its source code repository (if public) and the current version tags associated with the release, if these have been produced.

ENES CDI Service		Software components				
Name	**URL**	**Name**	**Description**	**Documentation**	**Repository**	**Release version (tag, branch)**
ESGF Data	-	esg-publisher	Python library to publish dataset on the ESGF.	https://esgf.github.io/esg-publisher/index.html	https://github.com/ESGF/esg-publisher	4.0.0-beta2
		esgf prepare	Python library to prepare data on ESGF publication.	http://esgf.github.io/esgf-prepare/	https://github.com/ESGF/esgf-prepare	2.9.2006
		esgf-pyclient	Python libray to request ESGF Search API.	https://esgf-pyclient.readthedocs.io/en/latest/	https://github.com/ESGF/esgf-pyclient	0.2.2
		CoG	ESGF Search UI.	https://esgf.github.io/COG/	https://github.com/EarthSystemCoG/COG	master branch
Data Citation	At DKRZ: http://bit.ly/CMIP6_Citation_Search At CMIP: https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id_citation.html	CMIP6 Data Citation Service	Maintenance of citation metadata+registration of DOIs and service provider for providers+users with a database backend and GUI and API interfaces	http://cmip6cite.wdc-climate.de	-	(restricted access)
Persistent Identifier (PID)	-	ESGF PID publisher	Python library to publish ESGF PID.	https://doc.redmine.dkrz.de/esgfpid/html/	https://github.com/IS-ENES-Data/esgf-pid	0.8.0
		RabbitMQ federation	PID messaging software.	https://acme-climate.atlassian.net/wiki/spaces/ESGF/pages/107708573/PID+Services+Working+Team+esgf-pidwt	https://www.rabbitmq.com/	(restricted access)
		PID consumer	backend where all PID registrations are processed	Internal documentation	https://gitlab.dkrz.de/esgf/handlequeue consumer	(restricted access)
IPCC Data Distribution Centre	At DKRZ: http://ipcc.wdc-climate.de At IPCC: http://www.ipcc-data.org	DDC	Long term curated archive of IPCC-relevant CMIP6 datasets	-	-	-

Errata	https://errata.es-doc.org/	Web-service	Errata Web-service.	https://technical.es-doc.org/	https://github.com/ES-DOC/esdoc-errata-ws	**master**
		Front-end	Errata front-end and forms.	https://technical.es-doc.org/	https://github.com/ES-DOC/esdoc-errata-fe	**master**
		CLI	Errata CLI to manage issue life-cycle.	https://es-doc.github.io/esdoc-errata-client/	https://github.com/ES-DOC/esdoc-errata-client	2.3.1
Data Statistics	http://esgf-ui.cmcc.it	esgf-dashboard	ESGF statistics dashboard.	https://acme-climate.atlassian.net/wiki/spaces/ESGF/pages/1054113816/Proposed+ESGF+Usage+of+Filebeat+and+Logstash	https://github.com/ESGF/esgf-dashboard	**master**
		esgf-dashboard-ui	ESGF dashboard front-end.	https://acme-climate.atlassian.net/wiki/spaces/ESGF/pages/1043464194/Federated+data+usage+statistics+ESGF+Dashboard	https://github.com/ESGF/esgf-dashboard-ui	**master**
Data Replication		synda	Python library to manager ESGF download.	http://prodiguer.github.io/synda/	https://github.com/Prodiguer/synda	**3.15**
Compute	https://ecaslab.cmcc.it/jupyter/hub/login	ECAS	ENES Climate Analytics Service instance	https://ecaslab.cmcc.it/web/home.html	https://github.com/ECAS-Lab	**master**
		Ophidia	CMCC projects for High Performance Data Mining & Analytics for eScience	http://ophidia.cmcc.it/	https://github.com/OphidiaBigData	**master**
		Birdhouse WPS framework	Python project related to WPS to support climate data analysis.	https://birdhouse.readthedocs.io/en/latest/	https://github.com/bird-house	master
		Twitcher (security proxy)	WPS security proxy	https://twitcher.readthedocs.io/en/latest/	https://github.com/bird-house/twitcher	0.9.0
		Roocs	ESGF-specific WPS	https://roocs.github.io/	https://github.com/roocs	**0.9.2**
ES-DOC	http://es-doc.org	CMIP6 content	-	-	https://github.com/ES-DOC-INSTITUTIONAL	-
		CIM schema	CIM document for Earth system documentation model	https://technical.es-doc.org/	https://github.com/ES-DOC/esdoc-cim-v2-schema	2.2
		pyesdoc	Python client for ES-DOC	https://technical.es-doc.org/	https://github.com/ES-DOC/esdoc-py-client	0.14.2.0

		pyessv	Python library to manager controlled vocabulary for ES-DOC	https://technical.es-doc.org/	https://github.com/ES-DOC/pyessv	0.8.4.3
		cf2cim	Python library to publish simulation CIM document for ESGF published data.	https://technical.es-doc.org/	https://github.com/ES-DOC/esdoc-cdf2cim	1.0.3.0
Metadata and schema service	http://cfconventions.org/	CF-convention	Climate and Forecast Convention	http://cfconventions.org/	https://github.com/cf-convention/	
		cfdm	Python reference implementation of the CF data model.	https://ncas-cms.github.io/cfdm/	https://pypi.org/project/cfdm/	**1.9.0.1**
		cf-checker	NetCDF Climate Forecast Conventions compliance checker	http://cfconventions.org/compliance-checker.html	https://pypi.org/project/cfchecker/	**4.1.0**
		cf-python	CF-compliant earth science data analysis library	https://ncas-cms.github.io/cf-python/	https://pypi.org/project/cf-python/	**3.11.0**
Identity Management and Access Entitlement		esgf-slcs-server	OAuth 2.0 and Short-lived Credential Service	-	https://github.com/ESGF/esgf-slcs-server	0.1.0
Climate4Impact	PROD: https://www.climate4impact.eu/	c4i-backend	C4I v2 back-end	-	https://gitlab.com/is-enes-cdi-c4i/c4i-backend	**0.1.0**
		c4i-frontend	C4I v2 front-end	-	https://gitlab.com/is-enes-cdi-c4i/c4i-frontend	**0.2.3**
		c4i-compose		-	https://gitlab.com/is-enes-cdi-c4i/is-enes3/c4i-compose	
		c4i-storybook		-	https://gitlab.com/is-enes-cdi-c4i/is-enes3/c4i-storybook	
		c4i-openid-relay		-	https://gitlab.com/is-enes-cdi-c4i/is-enes3/c4i-openid-relay	
		c4i-nginx-esgfsearch		-	https://gitlab.com/is-enes-cdi-c4i/is-enes3/c4i-nginx-esgfsearch	
		c4i-notebooks-service	Collection of notebooks using icclim	-	https://gitlab.com/is-enes-cdi-c4i/notebooks	**master**

		SWIRRL API	Software for Interactive Reproducible Research Labs	https://www.climate4impact.eu/c4i-frontend/helpSwirrl	https://gitlab.com/KNMI-OSS/swirrl/swirrl-api	master
		SWIRRL API for Jupyter Notebooks	Jupyte-Lab extensions for SWIRRL	-	https://gitlab.com/KNMI-OSS/swirrl/jupyterswirrlui	master
		icclim	Index Calculation CLIMate library	https://icclim.readthedocs.io/en/latest/	https://github.com/cerfacs-globc/icclim	**6.2.0**

Table 1. Final ENES-CDI Release, Software and Services overview (version in bold red have been updated in comparison with D10.2)

3 Data Services

As a reminder, the core ENES-CDI data services provide the capabilities needed to meet the functional requirements mentioned in Table 2 of D10.1 [1], Section 3.2.1. More specifically, those concerning data, citation, Persistent IDentifiers, the Distributed Data Center (DDC), errata, statistics and replication services are labelled as [DATAFR#-], [CITFR#-], [PIDFR#-], [DDCFR#-], [ERRFR#-], [STATSFR#-], [REPLICFR#-].

3.1 ESGF Data

The core components to make data accessible via ESGF are data preparation/standardization and quality check, data publication and data search, which is integrated in the ESGF portal component. Data delivery is supported by the standard protocols of the individual ESGF data nodes (HTTP as well as Globus/GridFTP for some larger nodes).

3.1.1 Brief reminder of related tools

The `esgf-prepare`[1] module helps data providers to create a project-related standardized directory structure for better organization of data files. It also provides a command to iterate over all files and create text files, aka *mapfiles*, listing all netCDF files to publish on the EGSF. The `esg-publisher`[2] can be used to publish the data to all components related to data distribution in the ESGF: a Postgres database, a THREDDS[3] data server and a Solr Index. It takes the *mapfiles* as input and reads all related files to extract the required metadata. The publisher component is also related to the PID server and creates the dataset PIDs and sends all PID information to the RabbitMQ[4] servers to register the PIDs.

The `esg-search`[5] component is integrated in the ESGF CoG[6] frontend so users can easily search and download data using different search facets.

3.1.2 Final release details

These services are in heavy operational use and were adapted to better cope with problem situations and to improve stability since mid-2020. The following adaptations are worth highlighting:

- improvement of the `esg-publisher` to prevent PID registration problems;
- tool development to monitor the consistency of the search index (e.g. with respect to replicas as well as PID registrations).
- new release of the ESGF data node installation based on docker containers
- replacement of current user interface (CoG) with Metagrid in Beta testing

[1] https://github.com/ESGF/esgf-prepare
[2] https://github.com/ESGF/esg-publisher
[3] https://www.unidata.ucar.edu/software/tds/
[4] https://www.rabbitmq.com/
[5] https://github.com/ESGF/esg-search
[6] https://github.com/ESGF/COG

3.1.3 Future search UI replacing CoG component

A new web frontend "METAGRID"[7] has been developed to replace the legacy CoG web application. This new search interface is developed in collaboration with US ESGF partners (LLNL). LLNL has integrated the new ESGF identity system developed by European partners into METAGRID. Work is now underway by US partners to support Globus[8] download. The future ESGF release will introduce a new implementation of the search system using the Spatio-Temporal Assets Catalogs (STAC[9]) specification for the web API and ElasticSearch for the underlying database technology. A prototype has been developed through funding from the IS-ENES3 project. Since the commencement of the new US ESGF2 project, US partners are developing a search solution based on Globus. However, collaboration work is underway between US and European partners to develop a common solution. This is likely to use STAC as the common interface for the web service API but use differing underlying technologies for storing search content (Globus uses OpenSearch on AWS, European implementation uses ElasticSearch).

3.2 Data Citation

Data Citation has become an integral part of scholarly publications. Initiatives, like COPDESS, ESIP, FORCE11 or Scholix, work on standardizations and guidelines for data citations. The Intergovernmental Panel on Climate Change (IPCC) Working Group I (WGI) has integrated data citations in the Sixth Assessment Report (AR6) to improve the traceability and transparency of the key findings of the climate assessment. In order to enable the citation of CMIP6 data, the data has to be provided for humans as well as for machine-readable access. Another key consideration is to disseminate the information about CMIP6 data references outside the project context (http://cmip6cite.wdc-climate.de).

3.2.1 Brief reminder of service functionalities

As described in [2], the service provides three functions:

1. Gathering of information via a GUI and an API from CMIP6 participants including user support;
2. Automated processing of DOI registrations and metadata updates, monitoring, and semi-automated curation;
3. Providing citation information for human and machine access using project-specific and standardized interfaces (schema.org, XMLs on OAI server).

3.2.2 Service stabilization

After the literature and data cut-off date of the IPCC WGI AR6, the data references should remain stable and author lists should not change any longer. Thus, the citation services had to

[7] https://github.com/aims-group/metagrid

[8] extension of the File Transfer Protocol (FTP) for grid computing

[9] https://stacspec.org/

be adjusted. It was decided to focus on the GUI and retire the Citation Service API. To meet this requirement, the functionality of the GUI was extended and changed: a new functionality enables modeling centers to control the visibility of their entries including experiment entries and thus enables them to focus on those citation entries with remaining tasks.

The dissemination of data citation information was enhanced by adding an API for machines with the same functionality as the existing Citation Search GUI. This was based on user feedback in the CMIP6 survey. The API has been used by the WGCM Infrastructure Panel to provide data citation tables on the web pages[10] and by the Errata Service.

The automated processing, that monitors the ESGF index about non-referenced datasets, has been made asynchronous with the DOI registration process. The ESGF index requests for the time-critical DOI registration process were rewritten. The performance and stability of the service components remain satisfactory and stable. The DOI registration process remains running hourly.

3.2.3 Further tasks and future work beyond IS-ENES3

The support of the Citation Service for input4MIPs (including boundary conditions and forcing datasets for model intercomparison projects) was extended to the interim activity “CMIP6plus” upon request by the project management. The gap between PIDs and DOIs has been closed on the PID side by adding DataCite DOI references to the Handle metadata. The publication of these links via Scholix remains an action item under discussion with OpenAire. In order to prepare the next phase of CMIP, CMIP7, a Task Team Data Citation has been established to recommend a few options for a sustainable Data Citation Service for CMIP7 and beyond, as the in-kind funding at DKRZ is not sufficient to provide the service for additional MIPs.

3.3 Persistent Identifiers

The persistent identifier (PID) service consists of permanently registered tracking identifiers during the ESGF publication. A PID is attached to each CMIP6 file and dataset and provides a landing page as a documentation hub. Thus, the persistent identifier service and its associated infrastructure (publication/registration client, message transfer via rabbitmq, server side components deployed at DKRZ) proved to be very helpful to maintain stable references to published CMIP6 data, track data replicas and versioning history as well as interlink errata information.

3.3.1 Brief reminder of the service components

The PID service consists in multiple interacting independently deployed components with clear APIs and interfaces which are integrated into the ESGF ENES CDI infrastructure:

- A distributed message transport layer
- ENES CDI specific message server components deployed at DKRZ

[10] https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id_citation.html

- Handle system[11] backend components for handling storage and generic PID CRUD (create, read, update and delete) operations.
- PID publication client tools (integrated into the ESGF publication software and interacting with the distributed message transport layer)
- PID curation tools to correct missing or erroneous PID registrations

3.3.1 Final release details

In RP3 work concentrated on improving the operational stability of the system with respect to future deployment scenarios (virtualization and dockerization) as well as monitoring and curation tools. Especially the extension of curation tools to correct and extend existing PID entries can be seen as a major improvement. Thus e.g. errors in the PID registration process can be corrected without the need for re-publication on the data provider side. Also PID related information can be updated later on to improve future FAIR data use cases, especially with respect to interlinking DOIs with PIDs.

3.3.2 Future work beyond IS-ENES3

In the future, besides continuous efforts to correct PIDs where flawed publication/un-publication process has led to incomplete information, it is planned to interlink the CMIP6 PIDs with DOIs of those datasets that were long-term archived in the World Data Centre for Climate (WDCC). This work has been started but is not finished yet. In a longer term, updating the PID profile to attain compliance with the Research Data Alliance (RDA) recommendations and to reach a higher machine-actionability is considered, but this would be a major effort and needs to be well prepared in cooperation with the relevant working groups from RDA and the Fair Digital Objects Group.

3.4 IPCC Data Distribution Centre at DKRZ

The IPCC DDC supports the authors in the writing process, especially in the analysis of data to derive key findings by providing Virtual Workspaces. The CMIP6 data subset underlying the AR6 will be long-term archived in the IPCC DDC AR6 Reference Data Archive as part of the traceability of AR6 outcomes and as well as for data re-use. The quality requirements for the IPCC DDC data and metadata are high, complying to the TRUST principles (Transparency, Responsibility, User Focus, Sustainability, Technology) as implemented in e.g. the Core Trust Seal.

3.4.1 Long-Term Data Archival to build the IPCC AR6 Reference Data Archive

WGI Technical Support Unit (TSU) has to provide the list of CMIP6 datasets used as input data for the IPCC WGI AR6, which they collect from the chapter authors. This dataset list was supposed to be available by March 2021. Due to delays at WGI TSU, the concept for CMIP6 data archival in the IPCC AR6 Reference Data Archive was altered: In the first step the datasets requested by authors for the data pool get archived. These are also disseminated within the Copernicus Climate Data Store for CMIP6. In a second step the CMIP6 datasets from the WGI

[11] http://www.handle.net/

TSU list not already archived will be added to the long-term archive together with information on usage in the AR6 (chapters and figures). The quality of the provided information required several cumbersome steps of quality checks and corrections, which further delayed the data curation and archival. Some datasets could not be identified or were no longer available in the ESGF. These corrections of the original dataset list from the TSU are carefully documented. The AR6 chapter and final dataset references have been added to the metadata allowing data users and AR6 readers to navigate between the different AR6 outcomes. The CMIP6 input data archival at DKRZ is still ongoing.

For AR6, DKRZ also long-term preserves intermediate datasets created by the IPCC authors during their assessment, which were selected by the TSU as datasets with a high reuse potential.

3.4.2 Developments towards consolidation and standardization

IPCC DDC sets up a joint catalog of the data holdings at the DDC Partners. A metadata profile of the World Wide Web Consortium (W3C) Data Catalog Vocabulary (DCAT) standard was developed. The DDC Partner DKRZ has provided the metadata in the agreed form, which required metadata export, mapping and an interface for metadata provision. Other work within the IPCC DDC are a redesign and restructuring of the DDC web pages and a new help desk, to which all DDC Partners contribute. Feedback for the IPCC FAIR TG-Data recommendations for AR7 based on experiences from all partners are formulated and are about to be sent to the IPCC Bureau. The unsecure DDC funding issue was targeted in a DDC Options paper, which was sent to the IPCC Plenary as part of the regular TG-Data report. Discussions on sustainable DDC funding are ongoing within IPCC and IPCC TG-Data.

3.4.3 Future work beyond IS-ENES3

The finalization of the long-term archival of the CMIP6 data subset underpinning the AR6 is the main remaining task; the long-term archival of the intermediate datasets is about to be finalized.

3.5 Errata

The ES-DOC errata system is a proposed platform to document and archive the reasons that would motivate the publication of a newer dataset version.

The main goal of providing this system is to improve data quality, if the system is used properly i.e. timely issue reporting, accurate description and comprehensive updates.

This puts considerable additional work in errata officers' hands (who are frequently also the data managers responsible for the publication of the data), encountering a bottleneck in the workflow that compromises the proper implementation of the system. After 4 years service production phase, the effect became more and more noticeable on errata provision. Leading to imagining the development and subsequent deployment of version 2.0 of the errata system.

3.5.1 What’s changed

The provision part of the first iteration of the errata system relies exclusively on a short list of identified errata officers to create and update errata entries.

This group is authenticated and authorized via the github OAuth system, and they are the only users allowed to contribute to the errata database.

However, after years of service production phase, users and errata officers noticed the pipeline was broken or overloaded between user notification to data managers and errata creation.

This led to a rethink of the provision part of the errata system. The change is mainly regarding the authorization part of the workflow.

In the version 2.0 of the errata service, unauthenticated users (therefore previously unauthorized) are now able to suggest errata entries. This will delegate the burden of providing quality errata information on the larger user base [fig.3]. This community effort will help feed the database with far more information than relying on a select few users.

However a main concern remains to verify and check the quality and integrity of the information suggested by users before ingestion, and this leads to the second change that occurred in this iteration, the moderation aspect. There will always remain a need for a group of identified power-users that will be able to validate, modify or reject user entries. These will be, at first at least, the same errata officers we rely on today to do the entire process of errata lifecycle. Hopefully the change will alleviate the pressure of providing exhaustive errata information upon user request, by only providing knowledgeable input into the validation process [fig.4].

A third required gear into this new workflow is a notification system that would serve as an update provider to both users and the moderation group.

We have opted not to host the potential discussion between users and moderators within the errata system because we deemed it out of scope, this will have to take place if necessary through email.

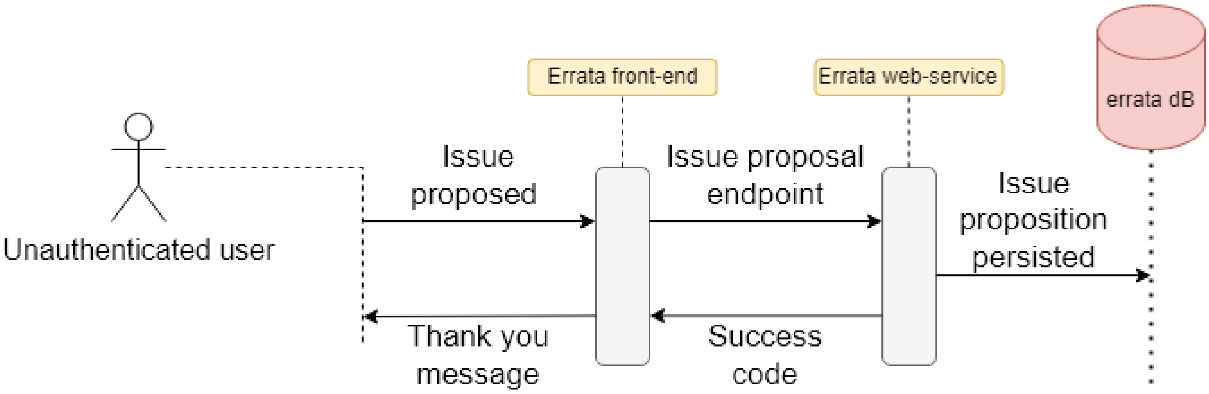


Figure 3. Unauthenticated user workflow

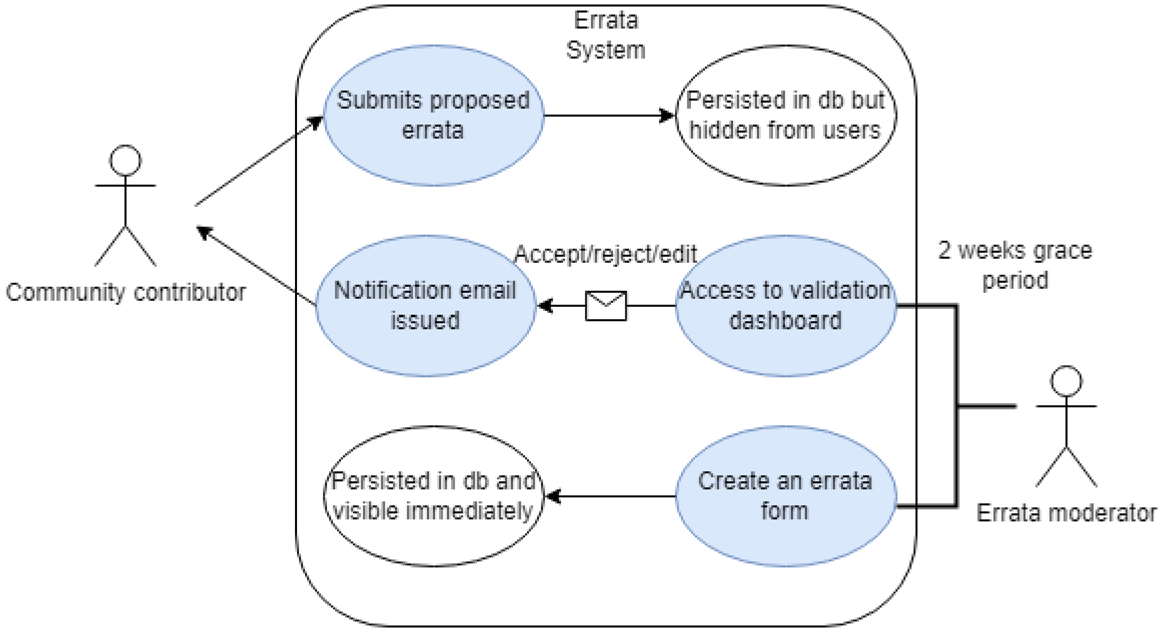


Figure 4. Moderator and user interaction

3.5.2 Deployment

We currently rely on webfaction web services provider to host our ES-DOC VMs (Virtual Machines). Following the same process we have used for ES-DOC services, there has been a test deployment performed on the dedicated test virtual machine. The aim of this deployment is to perform alpha phase testing, by running a set of predefined user scenarios and ironing out the bugs that could potentially appear.

3.5.3 Future work after IS-ENES3

While the main work for designing and developing this newer version of the service has been done within IS-ENES3, a second potential beta phase could take place prior to the official production release of the system that will see the deployment migrate to the production VM and exposed on the official endpoints of the current system.

A migration of the current database will be necessary to maintain the integrity of the existing archive, and database scripts have been also prepared and tested to undertake the task.

3.6 ESGF Data Statistics

The ESGF Data Statistics service, through its distributed and scalable architecture, captures, analyses and provides data usage and data publication metrics at a single data node level, within ENES and at the scope of the whole ESGF.

3.6.1 Brief reminder of the general architecture

The ESGF Data Statistics is located under the Federation service layer of the ENES CDI and directly interacts with i) the connected data nodes and ii) a dedicated local index node to retrieve all the metadata information about the data downloaded by the nodes (see Figure 5).

It was initially developed during the previous phase of the project (IS-ENES2) and significantly refactored in IS-ENES3 to better address the functional and non-functional requirements listed in the Milestone M10.1 “Technical requirements on the software stack”.

From a high level perspective, it

- collects and stores a high volume of heterogeneous metrics, covering general and project-specific measures;
- aggregates such metrics through an ad-hoc Extract-Transform-Load (ETL) system;
- stores them into a dedicated data warehouse;
- and provides a rich set of charts and reports through a web interface, allowing users and system managers to visualise the status of the IS-ENES/ESGF infrastructure through a set of smart and attractive web gadgets.

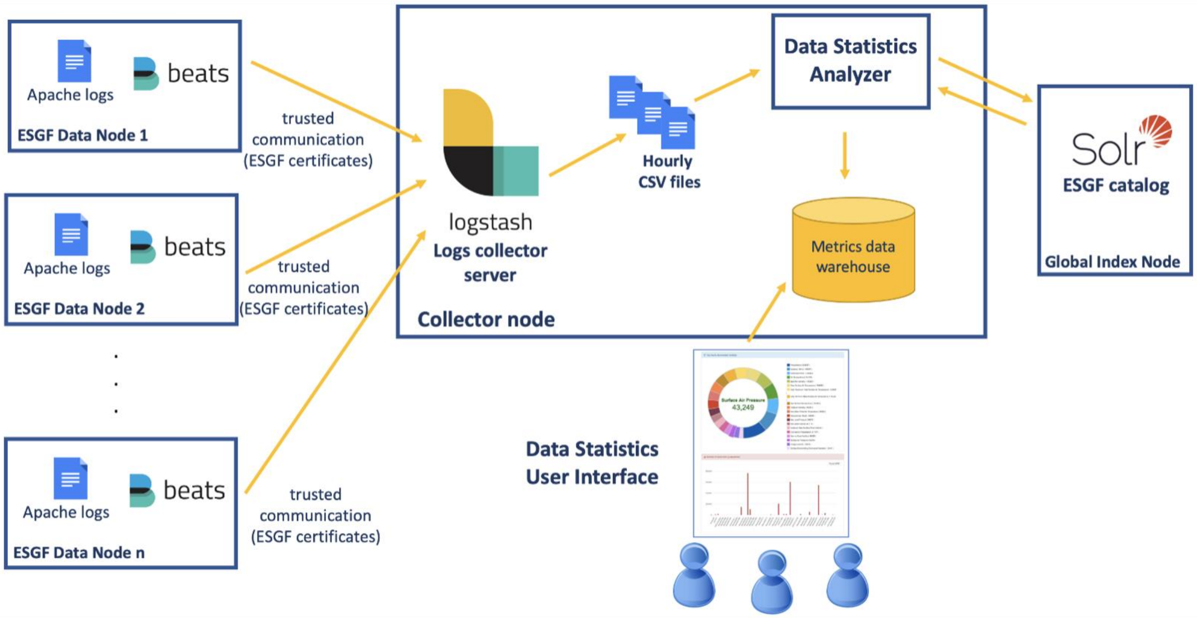


Figure 5. The ESGF Data Statistics architecture

3.6.2 Final release details

The service is properly working and no particular issues have been detected. Periodical data usage and publication metrics are provided and the IS-ENES3 Key Performance Indicators are regularly delivered. Moreover, as a proof-of–concept selected metrics are now regularly harvested and presented as graphical and tabular summaries at the CORDEX Project central website[12].

Besides the usual maintenance activities, listed below:

- integration of new data nodes into the architecture, consisting of i) interaction and support to node administrators, ii) sending of instructions to configure the local instance of the log shipper (Filebeat), iii) extension of the service configuration to accept the logs from the new node, iv) testing activities to check the log collection for the new node on the collector side v) and the processing of historical downloads and production of the related statistics,
- monthly backup activities of the statistics catalog, bug fixing, performance improving, periodic Logstash certificate update,
- continuous optimisation of the operational chain and its extension to include new functionalities and the improving of the existing ones,
- provision of additional and more explanatory graphical widgets within the ESGF Data Statistics user interface to second the user requests in terms of better understanding of the downloads and data publication trends,

During the last year, a new version of the service has been released to support the compliance to new data node technologies such as containers and Kubernetes. Specifically, to allow the new version of the data node, based on Kubernetes, to send logs to Logstash on the collector node, two main changes have been implemented in collaboration with CEDA:

- on the data node, Filebeat has been replaced by Logstash and, to send logs to the collector node, an *output* plugin called Lumberjack has been configured;
- on the collector side, a Lumberjack *input* plugin has been configured on Logstash.

The communication between the two instances has been correctly tested and the file format sent to the collector node is correctly reproduced also on the new Kubernetes version.

Also, investigation and test activities have been performed to support STAC, the future generations of the ESGF discovery service.

The ESGF Data Statistics service currently relies on the Solr instance to retrieve the metadata associated with each downloaded file. The new STAC query reproduces the old Solr query format, so the current metadata retrieval mechanism is not significantly affected in terms of querying the index node to collect metadata information.

Regarding the collection of download logs, the new data download service is able to send a log entry to the collector node with all the necessary information for the processing (file name, hostname, timestamp, etc.), so the operational chain has been properly adapted to process the new download log format.

[12] www.cordex.org, direct link to statistics webpage: https://cordex.org/statistics/

3.6.3 Future work after IS-ENES3

After the end of IS-ENES3, the service will be maintained with in-kind contribution from CMCC, in terms of service and infrastructure maintenance, integration of new data nodes and user support. The inclusion of the incoming CMIP7 project into the service will be evaluated time by time and, based on possible new institutional fundings, CMIP7 will be included into the service and the related data usage statistics will be delivered. Also, the deployment of the new service version, compliant with the new search service (STAC) will be put in place, when the new search will be operative on the data nodes.

New metrics will be included according to the requirements that came out during the latest ESGF F2F meeting in January 2023.

3.7 Data Replication

ESGF data are replicated across ESGF sites. Those replicas (i) improve data transfer rate around the world and (ii) ensure data recovery in the case of disk failure at some sites. Data replication service relies on a common strategy defined within the ESGF Data Replication team. Each site replicates a core subset of CMIP and CORDEX data and additional on-demand data, depending on storage capacities.

3.7.1 Brief reminder of the service architecture

Data replication involves the following architectural components:

- ESGF data nodes ("Tier 2") providing access to the originally published data collections from individual modelling centres;
- ESGF replica nodes ("Tier 1" - STFC-CEDA, DKRZ, IPSL, LiU) providing access to original data as well as replicated datasets. These replica nodes are associated to larger data pools hosting replicated datasets and also to high performance data transfer nodes supporting Globus-based data transfer;
- a replication management software component ("Synda") hosted at replica nodes, which triggers and manages parallel data replication streams involving different data nodes and different transfer protocols;
- a site specific data ingest and publication workflow integrating the replica datasets in the local data pools and publishing these datasets via ESGF.

3.7.2 Final release details

The replication software Synda is now distributed in the 3.4 version. The software is packaged through CNRS-IPSL conda channel making it extremely user-friendly to set up in a dedicated conda environment. To this day, there have been 553 separate downloads of the software, that is not only a versatile tool for power-users, but also provides an alternative to normal end users to interact and search and download data from the datastores.

The latest release is an important step towards the major overhaul we have started and continue working on aiming to modernize the tool, further optimize downloads, enhance transparency and error tracking, while maintaining the key features that made Synda the go-to replication tool for the community. By relying on the ***asyncio*** python module we now moved away from using system daemons to perform parallel downloads asynchronously. Synda today implements a task master that, through a pool of workers, assigns download tasks and maintains a watchful eye on the performance and status of each subtask (see Figure 6). We have opted to keep the old user-interface language for the time being to ease the transition for historical users.

At CNRS-IPSL, we know firsthand the potential of the tool since we are the first consumers, it is often our own requirements that guide development but we aim at also serving the community with a faster release cycle and a more responsive support and guidance

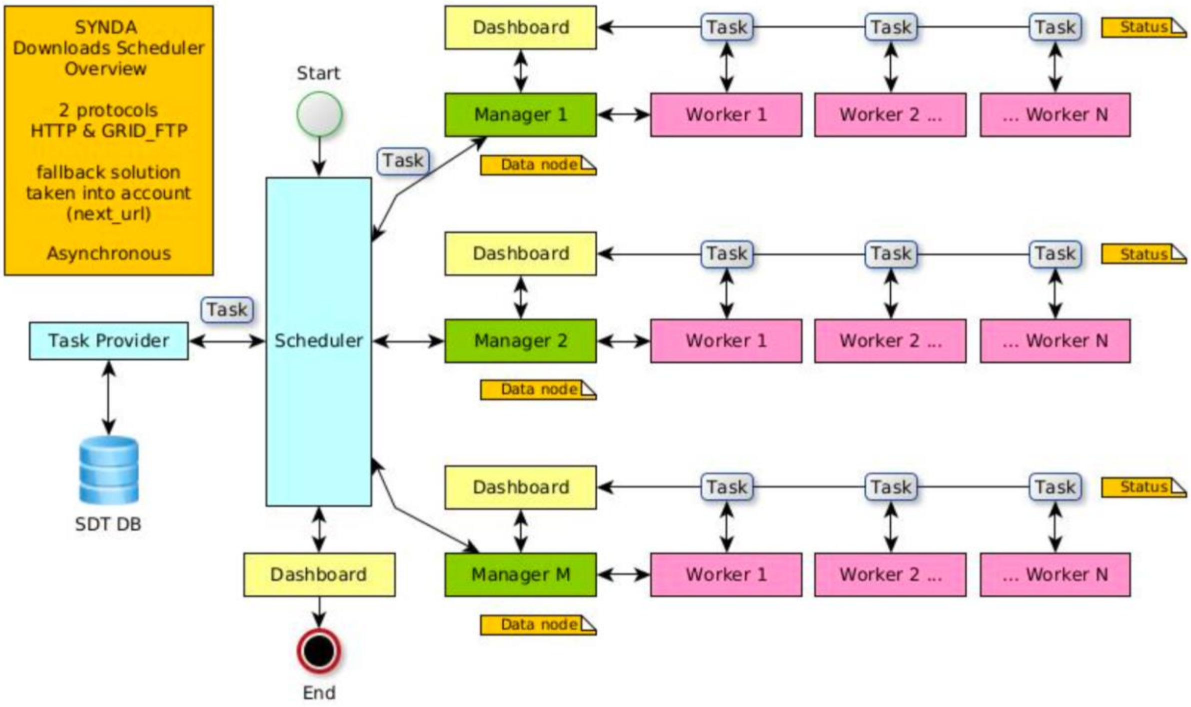


Figure 6. Synda scheduler overview

3.7.3 Future work beyond IS-ENES3

The highlights of the latest developments include the externalization of the configuration, and the implementation of a newer download manager via python's asyncio model that would enable much better download performance in case of important replication tasks. This enables the tool a far larger optimization in the download handling and parallelization.

Furthermore, we are now looking into taking the time to review our database model that has always been the core of the synda tool. The main motivation behind this review is to provide further optimization by implementing safer parallel access, and faster database transactions.

A key feature that is under development, is a much needed dashboard that provides through a handy UI a global review of the tasks performed, underway, waiting and in error if any. This should provide key performance indices that could guide users to further optimizing their synda usage.

At CNRS-IPSL we believe Synda is a tool that is far too important to let go and we wish to make it accessible to more users by removing needless complexities and offering more optimizations for our core power users, and we hope through the shorter release cycle the tool manages to answer expectations.

4 Metadata Schema and Services

The capability of efficiently handling metadata and data request schema is at the foundation to build the ENES CDI system in such a way to comply with the basic FAIR principles [4]. It addresses functional and non-functional requirements of the infrastructure. The former [CFFR#-] are mostly concerned with guaranteeing the provisioning of understandable standards to enable findability (F) and access to the data (A), while the latter [NFR#11] enables those mechanisms that, through interoperability (I) foster data reuse (R) and, to some extent, its reproducibility. This final release includes the specification of the Metadata for Climate Indices.

4.1 Climate and Forecast Convention

Substantial efforts have been conducted to develop software that validates compliance with the agreed CF (Climate and Forecast) standards[13], which aim at providing a description of the physical meaning of data and of their spatial and temporal properties. The main contributions are listed below with the updates which have taken place in 2021 through to 2023, which are incremental improvements:

Software Description	RP2 changes
cfdm: a Python reference implementation of the CF data model (version 1.10.0.1, March 2023) https://pypi.org/project/cfdm/	Implements the new features introduced in CF-1.9 (bar lossy compression by coordinate subsampling) and improves performance during reading of datasets.
cfchecker: the NetCDF Climate Forecast Conventions compliance checker (version 4.1.0, May 2021) https://pypi.org/project/cfchecker/	Implements the new features introduced in CF-1.8 (bar netCDF hierarchical groups).

[13] http://cfconventions.org/

cf-python: a CF-compliant earth science data analysis library (version 3.14.1, March 2023) https://pypi.org/project/cf-python/	Implements the new features introduced in CF-1.9 (bar lossy compression by coordinate subsampling), improves performance during reading of datasets and improves performance during regridding of datasets.

Table 2. Software modules compliant to the Climate and Forecast Convention

The tools have been developed by taking into account the official specification of the standards. Documentation web pages and discussion repositories, which led to the current definition of the vocabularies are listed below (these are updated as part of the service delivery reported on through WP7/SA2).

- Document: CF Convention Version 1.9 Document: http://cfconventions.org/Daa/cf-conventions/cf-conventions-1.8/cf-conventions.html
- Document: CF Standard Names Version 768http://cfconventions.org/Data/cf-standard-names/78/build/cf-standard-name-table.html
- Discussion: Conventions: https://github.com/cf-convention/cf-conventions/issues
- Discussion: Standard Names: https://github.com/cf-convention/discuss

4.2 CMIP Data Request

4.2.1 Current Status

There have been no new releases in this period (the current release is 1.2.0, Nov. 2022; this is identical in content to 1.0.33 published in Nov. 2020, but has some python library updates). The focus of activity has been on community discussions to guide the implementation of version 2, based on the framework defined in M10.2 - CMIP Data Request Schema 2.0[14]. A series of meetings was held, and are reported on in D3.3 Standards Synthesis.

Plans for CMIP7 discussed at the Working Group on Coupled Models annual meeting (WGCM-24, Dec. 2021) focussed on a community consultation run in 2022. In 2023 a number of CMIP Task Teams were established to oversee technical details of CMIP7 delivery, including a Data Request Task Team (DRTT) Reference and Objectives of the DRTT are under discussion and likely to be confirmed in early summer 2023. The timeline of CMIP7 remains unclear and will be strongly influenced by timelines for the 7th IPCC Assessment Cycle when the latter are announced (probably in late 2023).

4.2.2 Future work beyond IS-ENES3

The priority is to maintain flexibility, to deal with expected emerging requirements, and improve transparency by simplifying the structure. The structure will encourage MIPs to simplify their data requests, rather than offering a complex range of options which gave greater

[14] https://is.enes.org/documents/milestones/m10-2-cmip-data-request-schema-2.0/view

freedom but led to some confusion and loss of transparency. To compensate for this, there will be defined routes for data import and for viewing the content which will give some flexibility.

Further sources of simplification will be:

- Remove volume estimation from the core library (because of too many external dependencies);
- Standardise the way in which vocabularies are imported from CF, ES-DOC and CMIP CVs (Controlled Vocabularies).
- Restrict XML structure to map easily to SQL database structure.
- Clarify distinction between structural vocabularies which need to be fixed early and content which can evolve as the scientific focus clarifies.
- Clarify relation between import, export and internal structures to support greater flexibility in import and export while keeping a clean internal structure.
- Clean treatment of experiments as an imported vocabulary;
- Separation of ownership and provenance information (which will be dealt with using an approach based on ISO 11179 Metadata Registries) from content.

An illustration of the connectivity is given in Figure 7 below. Simplifying features include:

- Removal of central nodes to simplify the structure;
- Rigorous typing of links -- dotted paths seen in the 1.0 version are no longer supported;
- Consistent ontological structure for "Variable Packs" and "Experiment Packs".

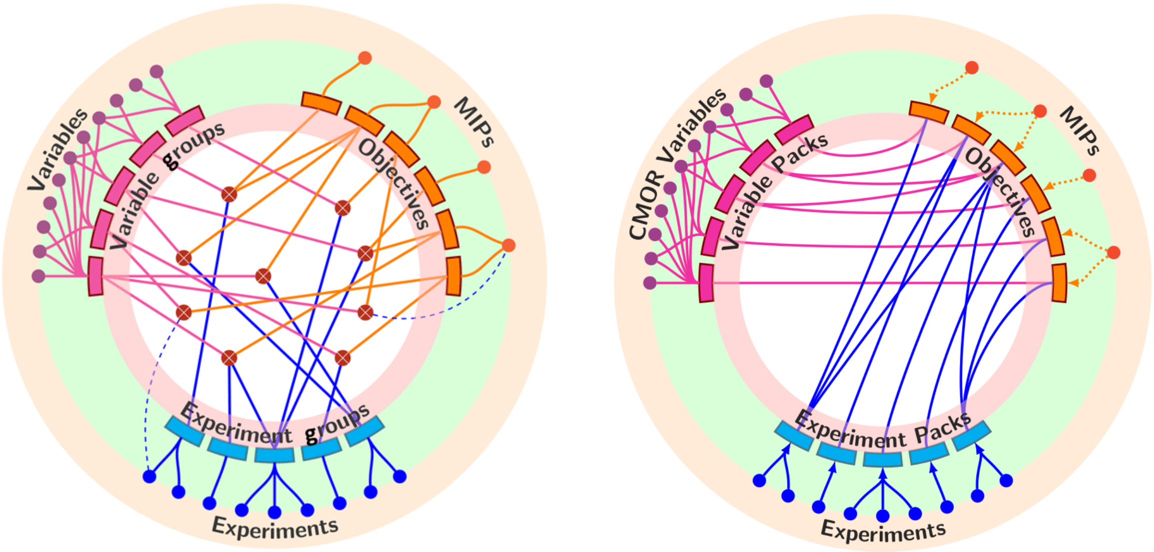


Figure 7. Connectivity diagram for the Data Request schema versions 1.0 and 2.0. (left) Data Request 1.0: nodes link to triples of Objectives, Variable Groups and Experiment Groups. (right) Data Request 2.0: Objectives link to Variable Packs and Experiment Packs

The next steps will be further meetings to review outcomes, followed by implementation and testing of the code. There has been progress in porting the request database into a python Django application which will enhance sustainability of the web service. Although progress on confirming technical objectives has been slower than expected. The creation of a Data Request

Task Team, on the other hand, has brought together a broad range of expertise which will greatly facilitate progress once CMIP7 is fully under way.

4.2.3 Summary of resources

Tools for contributing content to the CMIP Data Request

- ○ Forms: XLS Templates (https://w3id.org/cmip6dr)
- ○ Discussion: Github issues:
  - ■ Variables (https://github.com/cmip6dr/CMIP6_DataRequest_VariableDefinitions/issues)
  - ■ Request (https://github.com/cmip6dr/Request/issues)

- Tools for user-access
  - ○ Software: dreqPy (https://pypi.org/search/?q=dreqPy)
  - ○ Database: XML document within the dreqPy software package'
  - ○ Service: Data Request browser (https://w3id.org/cmip6dr/browse.html)

4.3 Metadata for Climate Indices

Climate indices encompass a wide variety of statistical summaries of climate data oriented towards specific user categories or application areas. With the enhanced support for advanced workflows provided by the ENES CDI there is a need to describe such climate indices in a unified and consistent way. That is, to equip the datafiles with metadata that describes the index data according to a unified set of terms and attributes. Essentially, this is achieved by collecting and reviewing well-established climate indices, e.g. from ETCCDI[15], ET-SCI[16], ECA&D[17] into a consistent format. This work was initiated in IS-ENES-2, but organising the information into a coherent form in the CLIX-META github repository[18] has been carried out in IS-ENES3. Further details are given in M10.3 and updated in M10.4.

5 Dissemination and Computational Services

In respect to the Requirements Overview (Table 2 of D10.1 [1]), in this section we address the Compute & Analytics functional requirements [COMPFR#-]. Non functional aspects (Table 3 of D10.1 [1]) will address mostly [NFR#8] and [NFR#11], concerning flexibility and interoperability of the services, respectively.

[15] ETCCDI: WMO-CCl/WCRP/JCOMM Expert Team on Climate Change Detection and Indices, https://www.wcrp-climate.org/etccdi, legacy webpage: http://etccdi.pacificclimate.org/list_27_indices.shtml

[16] ET-SCI: WMO/CCl Expert Team on Sector_specific Indices (discontinued), activities partly continues within WMO ET-CID Expert Team on Climate Information for Decision-Making, https://community.wmo.int/en/governance/commission-membership/sercom-management-group/standing-committee-climate-services/expert-team-climate-information-decision-making

[17] ECA&D: European Climate Assessment & Dataset, https://www.ecad.eu/

[18] https://github.com/clix-meta/clix-meta

As described in the first refease [2], the services illustrated in this section are improving the capabilities of the CDI for the provision of datasets and documentation, and the allocation of computational workspaces. Innovative technologies were adopted and further developed to better address the way researchers conduct their analyses, with built in mechanisms for FAIRness [3] and reproducibility [4].

5.1 Climate4Impact v2

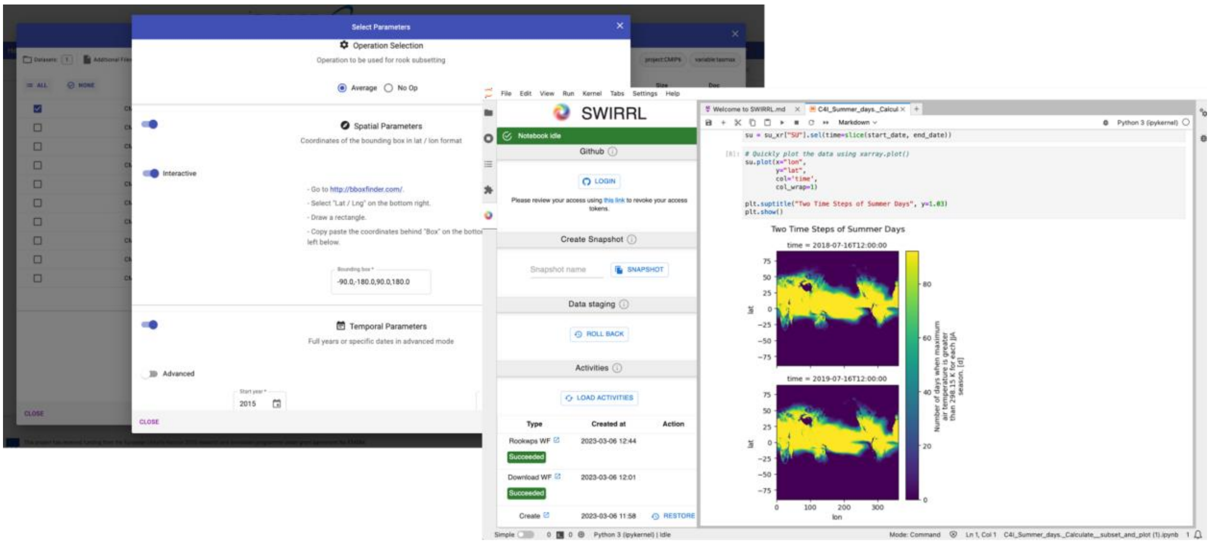
Climate4Impact (C4I) is a portal that enhances the discovery of climate research data and enables experimentation within impact analysis-ready workspaces. In the last phase of the project we opened to the public Climate4Impact v2[19], which is the new official release of the system.

The software technologies running the new portal better integrate with the components of the ENES-CDI. We improved the robustness of the integration of the various services, such as the federated identity and SSO (Single Sign-On) via the new IdeA IdP (Identity, Entitlement and Access Management Identity Provider); support for subsetting services (WPS: Web Processing Service), that will be incrementally available at more remote nodes (DKRZ and soon also CMCC are supported); provision and management of data-driven workspace via the SWIRRL API [5][20].

The portal and workspaces operate in an AWS (Amazon Web Services) environment, which is provisioned via a particular account managed by the KNMI. However, this is not prescriptive. The two components are built with cloud agnostic technologies, thereby they may be decoupled and hosted by different providers. This is important especially taking into account the sustainability of the service in the future, which may rely on more development teams and take advantage of a diversified procurement policy for the front-end and analysis workspaces.

[19] https://www.climate4impact.eu

[20] https://gitlab.com/KNMI-OSS/swirrl/



- Go to http://bboxfinder.com/.
- Select "Lat / Lng" on the bottom right.
- Draw a rectangle.
- Copy paste the coordinates behind "Box" on the bottom left below.

Bounding box *: -90.0,-180.0,90.0,180.0

Start year *: 2015

Please review your access using this link to revoke your access tokens.

Type	Created at	Action
Rookwps WF Succeeded	2023-03-06 12:44	
Download WF Succeeded	2023-03-06 12:01	
Create	2023-03-06 11:58	RESTORE

```
su = su_xr["SU"].sel(time=slice(start_date, end_date))

[8]: # Quickly plot the data using xarray.plot()
     su.plot(x="lon",
             y="lat",
             col='time',
             col_wrap=1)

     plt.suptitle("Two Time Steps of Summer Days", y=1.03)
     plt.show()
```

Figure 8. Remote subsetting in C4I. Interface for the parameterization of remote subsetting requests to C4I Workspaces. The Jupyter SWIRRL extension shows the execution of the WPS Roocs workflow in the activity list. The extension provides direct access to provenance information about the workflow.

To support users on processing and analyzing climate data, a python-based tool to calculate climate indices was further developed and completely revamped: ***icclim***. This tool has been completely rewritten and redesigned in IS-ENES3 in order to improve stability, sustainability, robustness, performance increase, using only python libraries as a backend and implementing all standards and guidance in open-source software code. There have been 2 major releases in IS-ENES3: v5 and v6. Current version at the end of RP3 is *6.2.0*[21]. It implements[22] most of the ECA&D climate indices[23], provides a powerful interface for users to define their own climate indices[24] (even complex ones), and can take advantage of parallel computing with *dask*. It is also used in the Jupyter Notebooks collection that is available in C4I[25] to help users in processing climate data for climate change impact analysis.

5.1.3 Future work beyond IS-ENES3

Further activities on C4I v2 should address the consolidation of the authorisation mechanisms on CORDEX data. This is conducted in cooperation with the data-nodes that will have to support the new ESGF IDentity Provider (IdP). Moreover, because of the inconsistent reliability of the implementation of the OpenDAP protocol at various nodes, we decided to support subsetting exclusively on providers offering a WPS implementation. Being this robust and in line with our intent to implement services taking into account FAIRness of data and operations. The WPS collects and disseminates provenance information about the subsetting operation, which C4I can ingest and manage seamlessly to its own provenance. This is not

[21] https://github.com/cerfacs-globc/icclim
[22] https://icclim.readthedocs.io/en/latest/references/ecad_functions_api.html
[23] https://www.ecad.eu/indicesextremes/indicesdictionary.php
[24] https://icclim.readthedocs.io/en/latest/references/generic_functions_api.html
[25] https://gitlab.com/is-enes-cdi-c4i/notebooks

possible with the OpenDAP endpoint, which is affected by poor performance and lack of a consistent implementation across nodes. For this reason its support by C4I workflows is dismissed. Thanks to feedback obtained during training, we acknowledge a demand for the implementation of collaborative data-pools, especially to run impact training in university courses more efficiently. Although this organizational structure is supported by the SWIRRL API natively, the C4I front-end has not been designed for collaborative purposes. We believe this would be a valuable capability to implement as future work. Finally, the support for future generations of ESGF discovery services, for instance based on STAC, should be addressed with dedicated resourcing for the development of the front-end adaptations. The team has followed progress on experimentation. However, the allocation of efforts to implement the change will require planning and resourcing.

5.2 ES-DOC

The ES-DOC (Earth System Documentation) software ecosystem facilitates both the provision and the consumption of documentation of the CMIP6 workflow and, where possible, automates the various and often complex stages involved.

5.2.1 Brief reminder of the service architecture

The processes being provided by the ES-DOC software stack are depicted in the workflow diagram of Figure 38 of report D10.1 [1] and are detailed in report D10.2 [2]. In summary, these are:

- A website hosted with WordPress, supported by dedicated servers and databases on the Opalstack commercial cloud hosting.
- A software ecosystem and archive contained in GitHub repositories under the ‘ES-DOC’ and ‘ES-DOC-INSTITUTIONAL’ organisations enable content pushed by modelling institutes to be processed and made available on the Wordpress website.
- Python-based utility libraries to automate the creation and publication of standardized documents and controlled vocabularies, and the storage of documents in repositories on GitHub.
- A shell-script library to facilitate development and maintenance.
- Python web services to manage documentation and errata stored in the Opalstack databases.
- Web applications written in JavaScript and as Vue.js components that support the viewing, searching, and comparing of the published documentation, as well as serving and displaying other relevant content.

All elements of this generic workflow have been implemented and are fully available as services, software packages and specification documents available from the ES-DOC organisation GitHub repositories at https://github.com/ES-DOC (see [2] for a more detailed description of repositories).

5.2.2 Second release progress and unexpected tasks

During 2022, progress has been a mixture of tasks that were planned for the year:

- More document types have been made available yet for creation (by the CMIP6 modelling groups) and consumption (by users of the CMIP6 outputs), namely Machine and Performance documentation. These documents describe the machines on which CMIP6 simulations were run, and the performance characteristics of those simulations (such as the number of model days that were simulated per real day).
- The experiment documents have been enhanced by the addition of a document versioning framework, and the addition of a new extended description field that allows the documents to be more interoperable with the Copernicus Climate Change Service (C3S) project, and portals with a non-expert user base.
- The ES-DOC comparator for comparing model descriptions has been updated to work for the CMIP6 models, as well as the existing CMIP5 model functionality.
- Documentation support for the CORDEX project has been implemented. This currently comprises the ability to document and publish regional climate model descriptions, which is the primary use case. Most CORDEX experiments are identical to CMIP6 experiments, which are already documented; an extension to document other CORDEX experiments may be required in the future.

A small amount of progress has been made on extending ES-DOC to the obs4mips project, providing observational datasets for model intercomparison projects. This has not resulted in a functional service, but the needs of the obs4mips community have been discussed. The intention was to make a subset of the ES-DOC functionality available during 2022, but this was not possible due to prioritization of the other CMIP6 requirements. However, the documentation of obs4mips is being considered as part of the CMIP7 design phase, and so this preparatory work has not been wasted.

ES-DOC experiment documentation was written for the Covid-MIP addition to CMIP6. Covid-MIP experiments investigate the climate implications of different economic recovery scenarios from the COVID-19 pandemic. The ES-DOC information was in place ahead of the deadline for inclusion in the IPCC 6th Assessment Report.

At the end of 2020, the Webfaction cloud hosting service withdrew some of the features that ES-DOC service relies on, prompting the need to find a new hosting service. Transferring to the new Opalstack service highlighted areas where more resilience was needed and also provided a documentation and training opportunity in 2022 so that other members of the team (or anyone else) could learn how to deploy the ES-DOC services.

5.2.3 Final release, future work beyond IS-ENES3 and progress on tasks

The final release of the ES-DOC software stack in late 2022 sees the inclusion of the final parts required for CMIP6:

- Machine and Performance documentation
- Conformance to protocols documentation
- Simulation and Ensemble documentation

The infrastructure for these components is now in place, and some CMIP modelling groups have engaged with it to create some documentation content.

The documentation of the ES-DOC stack itself has been completed (https://technical.es-doc.org/), ensuring resilience for the service.

Finally four ES-DOC team members are sitting on the new CMIP7 documentation task team. The experiences learned from CMIP6 (and CMIP5) will be brought to bear with the aim of delivering a truly sustainable and useful (if possibly pared back) documentation provision for CMIP7.

5.3 Institutional compute service deployments at ENES CDI sites

The ENES CDI aims to foster the “data near processing capabilities” paradigm. In this final release it has been achieved in close collaboration with the activities performed in WP5/NA4, to provide an integrated compute service for the ENES CDI. Thus beyond the different implementations of the core analytics services developed at each site, addressing institutional and national requirements, the main goal is to move towards a sustainable and integrated data analytics and processing layer for CMIP6 and CORDEX data, to efficiently support end-user needs (see the component diagram, Figure 8 of D10.2 [2]) . To this aim, three common aspects, that each compute service should implement during the project lifetime, have been defined and reported below:

- an interoperable and flexible server front-end based on the OGC-WPS interface [COMPFR#7][NFR#11][NFR#8];
- a programmatic client interface [COMPFR#6] with a Python binding;
- a security infrastructure based on the work and roadmap defined with the ESGF IdEA WG activity [COMPFR#5].

The progress made in the different institutional deployments of the compute service are reported below.

5.3.1 Compute Service at CMCC

5.3.1.1 Brief reminder of the general architecture

Details about the architecture of the CMCC compute service have been described in the Deliverable D10.1 (see Figure 9) and the Deliverables D10.2 and D10.3 presented information about the first and the second releases respectively.

Compared with the initial design, there were no deviations during the various stages of the project.

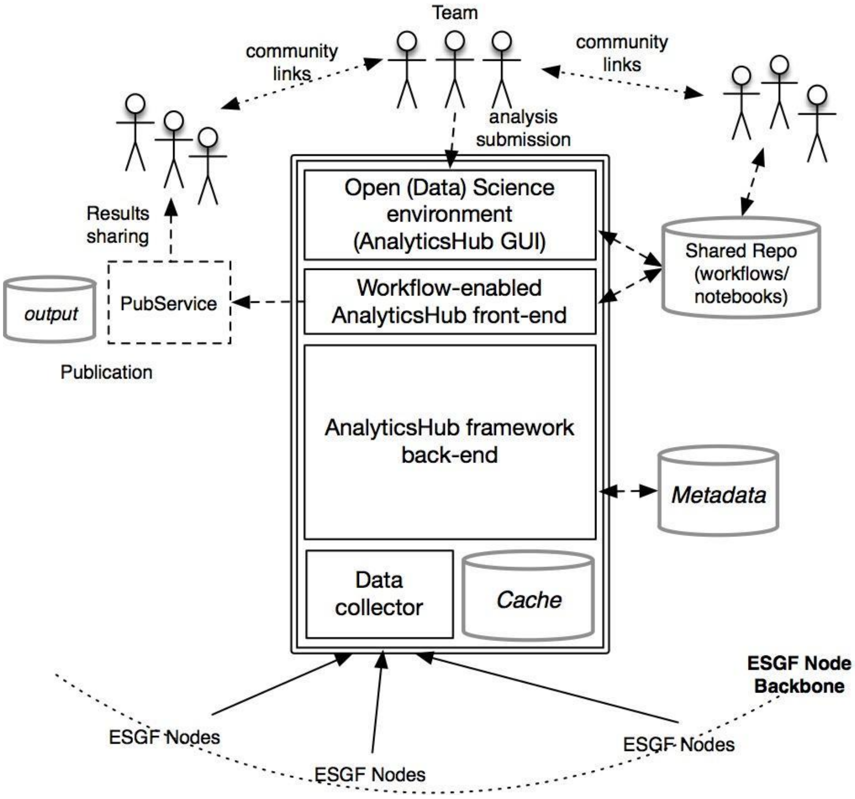


Figure 9. CMCC Analytics-Hub architecture

5.3.1. Final release details

During the last year of the project, the compute service at CMCC has been further improved to accomplish the requirement of providing common aspects to the different compute services. According to the initial design, described in the Deliverable D10.1, a new OGC Web Processing Service, based on the “Roocs” project, has been deployed, allowing the connection with the Climate4Impact portal.

Specifically, using the ESGF search, a user connected to the Climate4Impact portal can ask for a CMIP6 dataset available on the CMCC premises, which currently hosts more than 200.000 CMIP6 datasets corresponding to about 240TB. The Rook subsetting service, by avoiding unnecessary data downloads, performs temporal, vertical and/or horizontal subset operations and provides the portion of the dataset actually needed by the user. Simple averaging operations over time, vertical and horizontal domains over a user-selected range are also available via Rook in a pre-production stage.

5.3.1.3 Future work beyond IS-ENES3

An improved version of the CMCC Analytics Hub has been already deployed on the EGI infrastructure, providing a domain-specific implementation of the data space concept targeting the needs of climate scientists.

The ENES Data Space, developed in the context of the EGI-ACE project (Grant Agreement No. 101017567), leverages the compute service activities developed in IS-ENES to deliver a high-level and EOSC-enabled analytics environment for climate scientists. It was opened to end users at the end of 2021, offering a single integrated environment with ready-to-use data and programmatic capabilities for the development of data science applications. In particular, this solution integrates into a single environment: (i) Python libraries and frameworks for data analytics and visualization, together with (ii) a large data collection from key community experiments, like the Coupled Model Intercomparison Project (CMIP) and the Coordinated Regional Climate Downscaling Experiment (CORDEX), and (iii) scalable computing resources that can be deployed on demand on top of the EGI federated cloud infrastructure.

CMCC will further invest in this activity, in the future, improving the environment in terms of robustness, scalability and richness of the software offering, with the aim of democratizing the analyses and fostering collaboration in the climate domain.

5.3.2 Compute Service at UKRI

The UKRI compute service outline, and its architecture, was provided in deliverable D10.2 [2]. There has been little deviation from this during the latest period. The updates to the functionality have included features to select subsets of time and level by specifying a sequence of datetimes, specific years, months, days or values of levels.

5.3.2.1 Brief reminder of the infrastructure

The compute service at UKRI CEDA includes the development and deployment of the data sub-setting service "roocs" WPS stack[26], which has been developed in close collaboration with DKRZ, and is described in detail in section 5.3.3.2. Additionally, the JASMIN Notebook Service allows registered users to access both the archived CMIP and CORDEX data, as well as the CMIP6 object store holdings now stored on JASMIN (over 200TB are available in Zarr format). In recent months, UKRI CEDA has also been developing access methods using the nascent Kerchunk[27] library, which provides a JSON facade over files (such as NetCDF) to make them look like Zarr format. This has significant potential for ESGF Node Managers if combined with an S3-interface deployed over data stored in existing POSIX file systems - essentially it could enable cloud-native access to institutional data stores without copying any bytes of the actual data arrays themselves.

5.3.2.2 Second release details

The recent work on the "roocs" WPS has included:

- subsetting-by-value: allowing the query to include a sequence of discrete datetime and/or level values; this extends the original functionality allowing subsetting by interval;

[26] Roos: Remote operations on climate simulations: https://github.com/roocs

[27] https://fsspec.github.io/kerchunk

- support for CMIP6 Decadal data using a concatenation ("concat") operation of the “realization” dimension.
- improvements to the monitoring tools applied in the production system: tracking outputs and routinely checking that the scheduler and storage systems are functioning correctly;

The JASMIN Notebook Service was updated to a new software environment to support a greater array of updated open-source data analysis packages.

5.3.2.3 Future work beyond IS-ENES3

The "roocs" stack has been adopted by the ESGF Compute Working Team (CWT) as the WPS implementation for the ESGF Compute Node. The next steps for the "roocs" stack will include:

- spatial averaging: on regular lat/lon grid
- regrid operation for production use
- support for IPCC ATLAS data
- container based deployment with docker to be used for the next ESGF compute stack

Additionally, access to the existing holdings of CMIP6, CMIP5 and CORDEX data will be enhanced by Intake catalogs pointing to both the POSIX (NetCDF) version of the data as well as the object store (Zarr) version of CMIP6 (and some CMIP5 data).

CEDA will continue to investigate the potential use of Kerchunk as a possible solution for providing "cloud-native" access to existing data stores. This would be an important alternative to converting data to Zarr, with the following benefits:

- No data duplication: avoiding excess energy usage
- No software/processing costs to maintain conversion pipelines
- No requirement to track data changes (such as new versions)

5.3.3 Compute Service at DKRZ

The generic outline of the compute service and its architecture at DKRZ was provided as part of deliverable D10.2 [2] and stayed stable. The jupyterhub infrastructure was migrated to support the new HPC deployment at DKRZ and now includes additional support for GPU based computations. The other features of the jupyterhub environment (pre-defined compute kernels as well as support of user defined kernels) are also supported on the new system. Additionally specific kernel support for ESMValTool was added.

5.3.3.1 Brief reminder of the infrastructure

The core components are the HPC backend with attached large CMIP data pool which are made accessible via different interfaces: A jupyterhub deployment, direct access via frontend machines as well as OGC standardized processing service interfaces supporting basic data reduction and manipulation operations on data in the data pool (e.g. temporal and spatial subsetting).

5.3.3.2 Rook subsetting Service

CNRS-IPSL and DKRZ are working together on a Copernicus project to provide data access to climate projections like CMIP6 and CORDEX to the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/). CEDA is continuing to closely work together with CNRS-IPSL and DKRZ on WPS services, yet is no longer part of the Copernicus contract. The data access is provided using exclusive and distributed ESGF data nodes. These data nodes allow downloading of whole netCDF files of chosen datasets. To reduce the amount of data that gets transferred, we have in addition to the data nodes a subsetting service called Rook (developed as part of the Roocs project https://roocs.github.io/). The subsetting service allows to specify time and area ranges for datasets and performs the subsetting operation on the data pool site (see Figure 10). The result of the subsetting operation is provided for download to the requesting client (Climate Data Store).

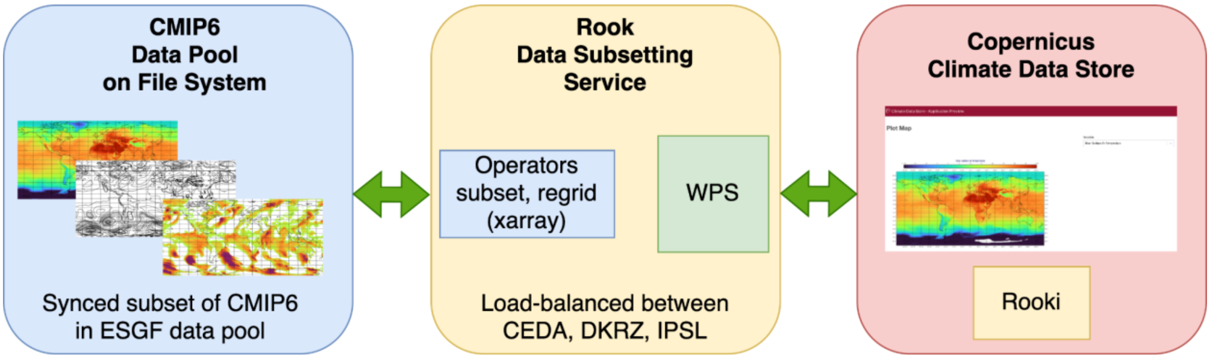


Figure 10. Rook service implementation

In addition to the subset operator there is also a temporal average (month, year) available. We currently work on the spatial average and regrid operator. The latest release supports CMIP6-Decadal data using a concat operation over the “realization” dimension. The Rook service is using the OGC Web Processing Service Standard[28] which allows the extension of operators on the service API level.

In IS-ENES, we use the same software stack to provide a subsetting service on the ESGF site to the Climate4Impact portal. Using the ESGF search, the C4I service asks for a subset of a CMIP6/CMIP5 dataset available at a specific ESGF data node. The Rook subsetting service avoids unnecessary data downloads and replaces the previously used OpenDAP implementation of Thredds. The OpenDAP implementation used in ESGF is not reliable and will not be available in future ESGF releases. Rook is potentially replacing OpenDAP in future ESGF installations. Unlike in OpenDAP, the operators can be extended and Rook will also provide averaging and regridding. Rook operators also produce provenance information using the W3C-prov standard (https://www.w3.org/TR/prov-overview/). This information is integrated into the provenance documentation of the C4I portal. The access to the Rook subsetting service in ESGF is protected using OAuth access tokens. These tokens are used by

[28] https://ogcapi.ogc.org/processes/

the C4I portal and provided by an ESGF Keycloak (https://www.keycloak.org/) instance at CEDA/STFC.

Currently only one site (DKRZ) is providing the Rook subsetting service for ESGF. In Future more sites can be added (CEDA, IPSL, etc.).

5.3.3.3 Final release details

The updates and improvements in the current release concentrated on the following components and aspects:

- Improvement of the CMIP6 data pool in the compute service infrastructure. An automatic procedure was established to regularly create and update intake catalogs for the CMIP6 data collections, which can be directly used in notebooks running in the jupyterhub deployment at DKRZ as well as part of batch jobs for the HPC system. The CMIP6 catalogs were additionally extended by additional catalogs for CMIP5, CORDEX as well as ERA5 data collections.

- Making available additional ready to use compute environments and associated notebook kernels. Different ready to use environments are accessible which now also include e.g. pre-established ESMValTool kernels for the Jupyterhub installation at DKRZ.

- Improvement of documentation based on jupyter notebooks. For this automatic continuous integration tests were established to automatically check the correctness of the demo notebooks in the context of the continuously changing compute and data environment.

- The compute environment was extended by a cloud storage component holding CMIP6 subsets based on "cloud native" storage formats (namely ZARR). Also for these subsets intake catalogs are available and are also hosted on the cloud. Based on this smaller scale data analysis hosted at end-users notebooks or hosted as part of Virtual Research Environments (like the D4Science EOSC infrastructure) can directly work on these data.

- Operationalization and extension of the functionality of the web processing service deployment, which is based on the "remote operations on climate simulations (roocs)[29]" developments. Functionalities now include temporal/spatial subsetting, averaging by time (year, month) as well as regridding (testing phase). The latest version supports CMIP6-Decadal data using a concat operation over the "realization" dimension.

[29] https://github.com/roocs

5.3.3.4 Future work

Starting in 2022 a new HPC system has been deployed at DKRZ and thus migration has taken place with respect to the compute service to the new platform. The core components of the compute service will stay unchanged also for the new system.

As part of the new system also a new tape backend has been deployed. To be able to flexibly stage data from tape for exploitation in the processing environment will be a major goal. This also includes the improvement of the previously mentioned "analysis ready data" provisioning on cloud storage to be able to stage data hosted on tape on cloud storage based on the cloud native storage format ZARR.

5.3.4 Compute Service at CNRS-IPSL

The generic outline of the compute service at CNRS-IPSL was provided as part of deliverable D10.2 [2], it has been reinforced in D10.3 [6] and finally consolidated in this release.

5.3.4.1 Brief reminder of the infrastructure

The last tranche of the IPSL computing capacity has been renewed. The new IPSL cluster now counts 2500 CPU cores together with 8.5TB of RAM, three to four times faster than the previous generation. The IPSL computing center deployed 4 GPU cores together with 256GB of RAM to answer IA computing facilities and climate services needs.

The storage spaces dedicated to users have also been renewed and now support Lustre 2.10. In addition, the IPSL computing centre provides 50TB shared storage for data analysis (Lustre), temporary and final results, alongside of a 4Po of specific CMIP and CORDEX and observational datasets (Reanalysis, Obs4MIPs, input4MIPS, etc.) with centralized access (including the whole French climate modelling production from IPSL and CNRM). The data access now benefits of "intake-esm"[30] catalogs available for all data archives (CMIP5/6, CORDEX, C3S, and related tied projects like PMIP3 or LUCID).

5.3.4.2 Final release details

The compute service design at CNRS-IPSL still mainly relies on Virtual Access (VA) with generic and "on-demand" remote access (i.e., through SSH) to dedicated login nodes. Pre-configured virtual environments have been installed in order to mutualize useful tools for data quality check and analysis for all users on the computing center:

- "climaf" virtual environment includes the CliMAF library for climate model evaluation
- "cdms2" environment provides the latest version of the "cdms2" library (Climate Data Management System).

[30] intake-esm: https://github.com/intake/intake-esm

- “analyze” provides a base environment for data analyzing including the well-known Xarray[31] and Dask[32] libraries that provide user-friendly I/O and parallel tasking.
- “esmvaltool” provides a pre-configured instance of the last version of the ESMValTool.
- “cmor” provides a pre-configured instance of CMOR tool to standardized CMIP and CORDEX data.
- “icclim” provides a base environment as used in Climate4Impact portal to compute useful indicators.
- “pcmdi-metrics” provides a pre-configured instance of the PCMDI Metrics package for model evaluation.

All these environments can be loaded by any users on the computing center through the “module” command-line.

In early 2022, IPSL computing center deployed into production a Virtual Access to its computing facilities through a JupyterHub with a Pangeo-like suite of useful libraries. The JupyterHub relies on the “Dask-jobqueue”[33] plugin to interface Dask with the usual IPSL PBS (Portable Batch System) manager (ciclad-web.ipsl.jussieu.fr).

The Web Processing Service is still up and running on a 8CPU machine and dedicated to Copernicus needs. This WPS is based on the “Roocs”[34] project led by CEDA and DKRZ partners.

A Kubernetes instance has been finally deployed into production at IPSL. Due to missing staff allocated to the infrastructure itself the Kubernetes is for educational use only for the time being.

5.3.4.3 Future work beyond IS-ENES3

In 2023, the JupyterHub will be deployed on top of a new Kubernetes instance to scale up the IPSL VA. Training and documentation about available compute services and research environments will be also improved to better accompany the users and reduce pressure on the user’s support.

The 4PB of referenced data will be renewed and mutualized in 2024 with other French national infrastructure. This will make CMIP and CORDEX data more interoperable with observation and some biodiversity data sets.

Finally, the WPS facilities will be enforced in the coming year and opened to IPSL computing centre users and extended to the ENES community (not only in the Copernicus context), with:

[31] Xarray: http://xarray.pydata.org/en/stable/
[32] Dask: https://dask.org/
[33] Dask-jobqueue: https://jobqueue.dask.org/en/latest/
[34] Rooc project: https://github.com/roocs/

- A new engineer at IPSL to support DKRZ in the development of the “Roocs” project.
- Additional dedicated computing resources due to the upgrade of the IPSL virtualisation platform.

6 Identity Management and Access Entitlement

Identity management and access entitlement encompasses functionality to enable the authentication (asserting the identity of a human or service *actor*) and authorisation (management of access rights to restricted resources for that actor). Historically, for ESGF and the ENES RI restricted resources have concerned data such as CORDEX and CMIP5 datasets. With the move towards open access data and data-proximate computing, needs have shifted towards the ability to support access restrictions for computing resources. A good example of the latter is sign-in to a Jupyter notebook environment.

Under the ESGF Future Architecture initiative, a new implementation of the identity and access entitlement (IdEA) system was developed starting in 2019 and continued as part of the IS-ENES3 project. Since reporting for D10.2 all components have been completed. In summary there is:

- a complete new system for user authentication including single sign-on and user delegation
- a new system for user authorisation enabling simpler integration with web services

Integration testing has been conducted between partners trialing the new authentication and authorisation technologies. A complete end to end test system has been deployed for the Climate4Impact portal. This demonstrates access to secured CORDEX datasets held at the CEDA data node using the new system.

6.1 Current Operational Status for Authentication and Authorisation with ENES CDI

This can be described as follows:

- Systems requiring authentication and authorization use the existing legacy ESGF system based on OpenID 2.0 and short-lived user X.509 certificates for authentication and SAML interfaces for authorisation.
- Some services, notably the Climate4Impact Portal take advantage of OpenID Connect single sign-on and OAuth 2.0 for delegation of authentication. These services are in
- In some cases, where simple authentication is required, GitHub’s OAuth 2.0 service is used.

6.2 Implementation status for ESGF Future Architecture Components

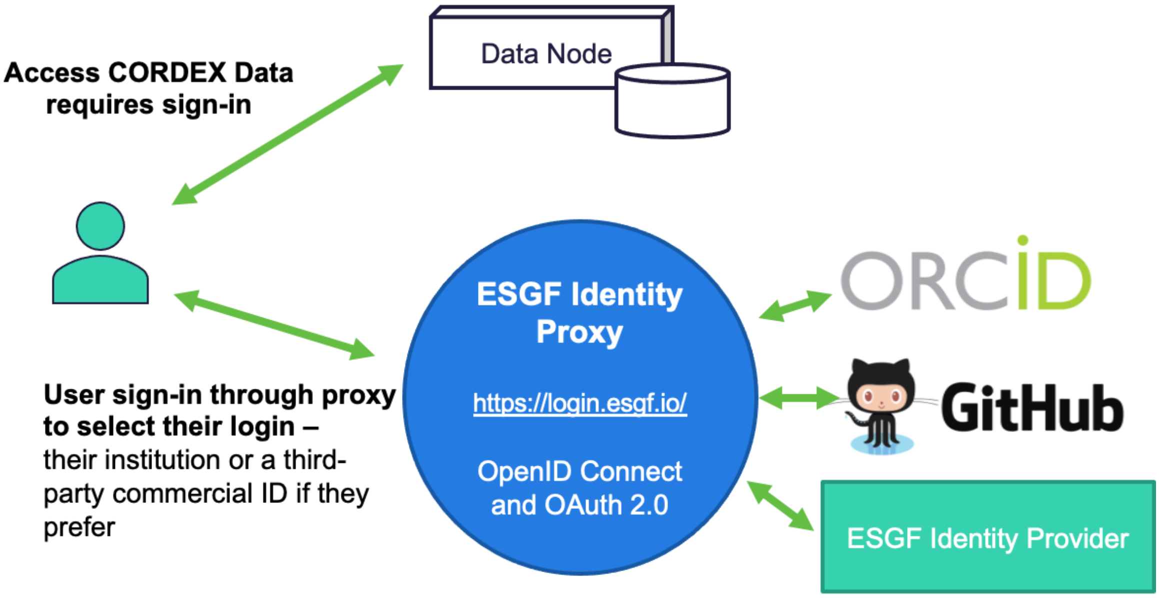
6.2.1 Authentication, single sign-on and user delegation

The new ESGF system adopts the OAuth 2.0 framework for user delegation and OpenID Connect for single sign-on use cases. These also enable authentication using tokens passed in

HTTP request headers and provide a simpler alternative to X.509 client certificate-based authentication used in the original ESGF system for command line use cases.

6.2.2 IdP Proxy and Federation Site IdP Implementations

In the existing model for ESGF, organisations participating in the federation typically deployed an IdP (Identity Provider) enabling their users to login and access secured data across the various data nodes deployed around the world. This led to a complex and confusing arrangement for users. Furthermore, the technology used OpenID 2.0 and had inadequate arrangements for managing the trust relationships between identity providers and ESGF data nodes that relied on them to mediate user login. For the new ESGF system implemented as part of the IS-ENES3 project, consulting current best practice for research federations a new model has been adopted based on work by the AARC project (https://aarc-project.eu/architecture/). This is illustrated in the diagram below:


Access CORDEX Data requires sign-in
Data Node
ESGF Identity Proxy
https://login.esgf.io/
OpenID Connect and OAuth 2.0
ORCID
GitHub
ESGF Identity Provider
User sign-in through proxy to select their login – their institution or a third-party commercial ID if they prefer

Rather than many organisations each deploying their own Identity Provider, there is a single new service, an Identity Proxy which mediates login requests for users authenticating with ESGF services. This simplifies the integration of ESGF services because rather than each service needing to interface with lots of different identity providers, they only need to link with a single identity proxy service.

The Identity Provider (IdP) Proxy then, is a special arrangement of the traditional model of Identity Provider ⇔ Rely Party pattern for single sign-on. The Proxy provides an intermediary between Relying Parties (in this case, ENES CDI services requiring authentication and authorisation) and IdPs. Supported IdPs include those sites in the federation wishing to host such services and also a number of external commercial IdPs such as Google and GitHub. Both IdP Proxy and federation site IdP use the industry-standards OpenID Connect and OAuth 2.0.

The implementation used for ESGF is based on customisations of the open source Keycloak[35] software from RedHat.

1. **IdP Proxy**: implementation complete; Docker image completed; Deployed for integration testing on JASMIN.
2. **Federation site IdP**: completed. Deployment ready for CEDA. Docker image

6.2.3 Relying Party and Policy Enforcement Point Implementation

The Relying Party (RP), Policy Enforcement Point (PEP) and Policy Decision Point (PDP) are key and fundamental components that enable a service offering secured resources to integrate with the rest of ESGF's system for authentication and authorisation. Without these, it is not possible for example for a data node to provide the required secured access to CORDEX data. To give another example, if a compute service deployed in the ENES RI needs to be secured it will need to have these components installed as an integral part of the service.

- A Relying Party (RP) is a component that implements the interactions necessary with an IdP to secure a given service enforcing authentication with single sign-on.
- A Policy Enforcement Point (PEP) is a component that *enforces* authorisation access control decisions for a service. The PEP refers to a Policy Decision Point (PDP) or authorisation service in order to make the access control *decisions* themselves.
- PDPs make decisions based on user attributes, access policies related to resources and in some cases, other factors related to the environment, for example access restricted to certain temporal constraints.

Both PEP and RP lend themselves well to a filter architectural pattern in which they front access requests to the application to be secured and enforce the access constraints. This is illustrated in Figure 11 below.

6.2.4 Federated Authorisation

As stated above (section 6.1), the existing legacy ESGF system uses SAML interfaces for authorisation interactions. The main actors are the PEP (See component inside Nginx Access Control Filter in Figure 11), PDP (aka. Authorisation Service, bottom right in Figure 11) and Attribute Service. In the new system (see Figure 11), the per-application PEPs are replaced by a generic PEP deployed as part of the Kubernetes Ingress Controller (Nginx) or if not using Kubernetes, via a standalone Nginx deployment. Since reporting in D10.2, the authorisation system has been redesigned to use OPA (Open Policy Agent)[36]. OPA uses a declarative policy language called Rego. This will replace the existing bespoke XML-based policies used for ESGF. OPA also provides a RESTful API for the interface between PEP and PDP. Another significant change is the communication of user attributes for authorisation decisions. In the existing ESGF system, the authorisation service *pulls* user attributes from a central Attribute

[35] https://www.keycloak.org/

[36] https://www.openpolicyagent.org

Service. In the new system as was proposed in D10.2 [2] we move to a push model for the communication of user attributes. Consequently, the Attribute Service is deprecated. Instead, when a user signs in with the central proxy, the proxy adds in all the user’s attribute entitlements and returns this to the Relying Party in the single sign on authentication flow. Consequently, the PEP has visibility of these attributes and can *push* these across the interface to the PDP such that the PDP then enact access control decisions based on these user attributes and information about the secured resource being requested.

1. PEP (implemented in Nginx auth plugin and standalone Python Django application - see preceding section). Django helper application to Nginx implements an OPA web service client callout to PDP to get authorisation decisions
2. Authorisation Service (PDP). Open source implementation of OPA is in the Go programming language. The OPA service provides a web service API to the PEP. The OPA deployment parses and enforces a policy file written in Rego.
3. Attribute Registration interface. Users register for access to secured resources through this web interface (https://login.esgf.io/registration/). User attributes are registered with the central IdP. When a user signs in, these attributes can be communicated to Relying Parties across the interface between IdP and SP. A standalone Django application has been implemented for this which integrates the Keycloak API.

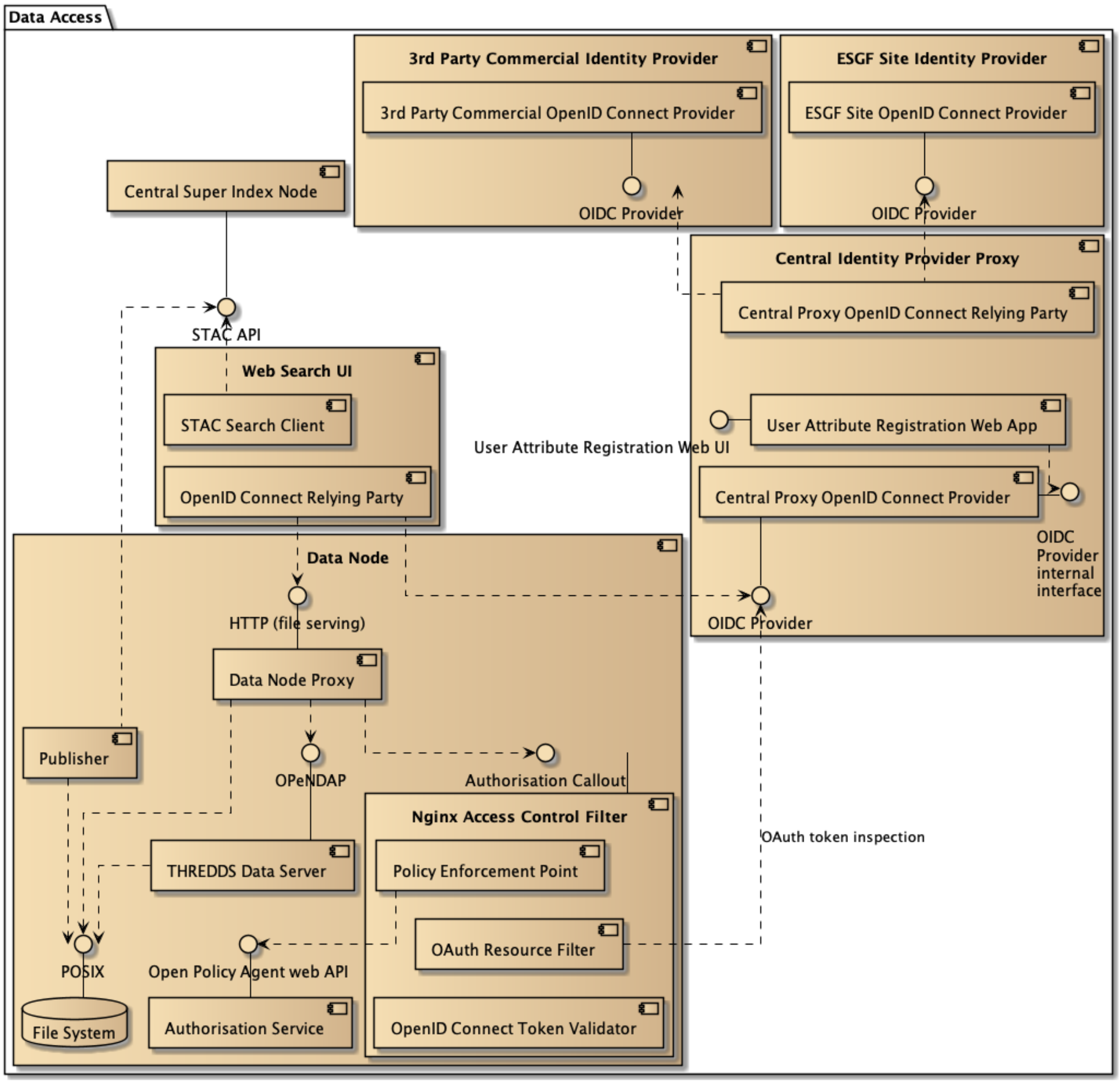

Data Access
3rd Party Commercial Identity Provider
3rd Party Commercial OpenID Connect Provider
OIDC Provider
ESGF Site Identity Provider
ESGF Site OpenID Connect Provider
OIDC Provider
Central Super Index Node
STAC API
Central Identity Provider Proxy
Central Proxy OpenID Connect Relying Party
Web Search UI
STAC Search Client
OpenID Connect Relying Party
User Attribute Registration Web UI
User Attribute Registration Web App
Central Proxy OpenID Connect Provider
OIDC Provider internal interface
Data Node
HTTP (file serving)
OIDC Provider
Data Node Proxy
Publisher
OPeNDAP
Authorisation Callout
Nginx Access Control Filter
OAuth token inspection
THREDDS Data Server
Policy Enforcement Point
OAuth Resource Filter
POSIX
Open Policy Agent web API
File System
Authorisation Service
OpenID Connect Token Validator

Figure 11. ESGF Future Architecture showing identity services and index and data nodes

6.3 Operational Status: next steps required

Overall, the status for the IdEA components can be summarised:

- A completely new architecture for access control services has been defined as part of the broader ESGF Future Architecture work
- A complete working implementation of that architecture has been implemented during the IS-ENES3 project. This software is being used for the ENES-RI
- A complete integration test and demonstration of the new system has been completed. This is an end to end working system for access to secured CORDEX data from CEDA using the Climate4Impact portal.

The system needs to be made fully operational in the ENES RI and in ESGF. The following steps are required to make this transition successful:

1) All nodes across the federation need to upgrade to the new future architecture deployment of ESGF. To date, GFDL (deployed on Amazon Web Services) and CEDA have been upgraded. ORNL, LLNL and DKRZ have systems ready to deploy or are in a pre-production status (i.e. services are part deployed, being tested but not fully ready for operation).
2) In ongoing collaboration with US counterparts and in discussion with the ENES-RI consortium it has become clear that a single identity proxy is not an adequate solution to serve the needs of the ESGF federation. This is a policy decision made by the European organisations making up the ENES-RI. Rather than a single identity proxy, there will be two: US ESGF partners will use an identity proxy based on a service from Globus; European partners will use identity proxy based on the EGI CheckIn service. As of writing, initial discussions between US and European teams indicate that it will be possible to develop a technical solution to integrate and make interoperable Identity Proxy services in the US (Globus) and Europe (EGI CheckIn). Such a solution will need to address issues including how to manage consistent user identifiers and user attribute information between the two services. This is necessary in order to ensure seamless access to federated resources across nodes around the world.

7 Conclusions and main targets after IS-ENES3

This final release of the ENES CDI marks a new step in the improvement of quality of all of the pre-existing software components, a step towards a fully interconnected infrastructure and the refined view of the next steps to be pursued.

Data services have been mainly consolidated in their production version to cope with some issues and improve the overall stability of the system. As identified in the first and second release of the ENES CDI [1,6] important targets have been addressed:

- Interconnections of data services have been mainly consolidated in their production version to accommodate larger data-streams and improve the overall stability of the system. These include consolidation of ESGF publication tools to cope with some identified issues, the improved curation and consistency of the ESGF PID collection, the continuous integration and optimization of the ESGF Data statistics collection, in terms of inclusion of new data nodes, activities related to the incoming CMIP7 project and addressing of new requirements emerged during the last ESGF Hybrid F2F meeting in Toulouse. The data replication tool “synda” has moved into a new “esgpull” library relying on Python 3 and asynchronous paradigm to download data from the ESGF efficiently.
- Due to the increasing impact of data deluge in recent years, a key aspect is the ability of the analytics layer to scale up by accommodating larger data streams and computations, thus pursuing the near-data processing paradigm. Moreover, machine learning and artificial intelligence algorithms will help with the analysis of rapidly increasing volumes of Earth system data.
- Interactive services such as Climate4Impact and the Analytics-Hub had major improvements in terms of their user-facing interfaces, search and computational

capabilities, especially addressing the integration of more flexible development environments based on notebooks (JupyterLab). In C4I these have been delivered as part of advanced reproducible workspaces (SWIRRL), allowing execution of workflows, with automated provenance recordings and versioning of the users' methods and computational contexts [3].
- Taking into account the official specification and new features of the Climate and Forecast standards in its version 1.8 and 1.9.
- Redesigning a federated identity and access entitlement (IdEA) by prototyping the required stack components for a new implementation that adopts the OAuth 2.0 framework for user delegation and OpenID Connect for single sign-on use cases. The system is used to authenticate and register users in C4I, providing access to computational workspaces. Moreover a test is ongoing between KNMI and CEDA to enable authorised access data to CORDEX by C4I workflows.
- Addressed long-term archival of the CMIP6 data subset underpinning the AR6 and the curation of data for older assessments.
- The new errata interface opening issue registration to any users has been developed.
- Pursued the development effort on “synda” replication tool and particularly its discovery module relying on the ESGF Search API.
- Deployed the Policy Enforcement Point (PEP) and federated Identity Providers on Tier 1 sites within the new ESGF architecture.

The current ENES CDI will be supported past the end of the IS-ENES3 project, as agreed between the partners. Maintenance will continue, and future developments that are presented here in this deliverable will continue, partly funded by other projects and by in-kind contributions. This will ensure users that those services will continue to be available within the next few years.

8 References

[1] S. Fiore, et al. *D10.1 - Architectural document of the ENES CDI software stack,* https://zenodo.org/record/4309892#.X9CSbS2ZMn1

[2] A. Spinuso, et al. *D10.2 - First release of the CDI software stack*, https://zenodo.org/record/4450012#.YaCjq73MJqs

[3] Goble, Carole, et al. "FAIR computational workflows." Data Intelligence 2.1-2 (2020): 108-121. https://doi.org/10.1162/dint_a_00033

[4] Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3.1 (2016): 1-9.

[5] Alessandro Spinuso, Mats Veldhuizen, Daniele Bailo, Valerio Vinciarelli, Tor Langeland; SWIRRL. Managing Provenance-aware and Reproducible Workspaces. *Data Intelligence* 2022; 4 (2): 243–258. doi: https://doi.org/10.1162/dint_a_00129

[6] G. Levavasseur, et al. *D10.3 - Second release of the ENES CDI software stack*, https://doi.org/10.5281/zenodo.7728921