IS-ENES3 Deliverable D10.3

Second release of the ENES CDI software stack

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

Authors: 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), C. Pagé (CERFACS), A. Stephens (UKRI), A. Spinuso (KNMI), M. Stockhause (DKRZ)

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

Release date: 21/12/2021

ABSTRACT

This deliverable illustrates the second release of the ENES 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. This document describes new 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	X
CO	Confidential, only for the partners of the IS-ENES3 project	

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	18/10/2021	Guillaume Levavasseur	Preliminary Structure
Collection and review of first round of contributions	15/11/2021	Guillaume Levavasseur	First inputs & general sections completion
Added icclim content and updated parts of C4I content	15/11/2021	Christian Pagé	icclim and C4I contents
Document formatting and last update	26/11/2021	Guillaume Levavasseur	Final structure
Version to review	17/12/2021	Guillaume Levavasseur	Last inputs & final formatting.
Submitted version	21/12/2021	Stephan Kindermann + Sylvie Joussaume + Guillaume Levavasseur	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. This report describes new developments and implementations for each ENES services and software, leading to the second release of the CDI.

As an update, the document follows a similar organisation as the D10.2 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 second Reporting Period (RP2). 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 RP2, to be pursued in RP3, 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 second technical overview of the progress achieved with the implementation of the requirements and architecture presented in D10.2 "First release of the ENES CDI software stack" [2]. The document addresses the deliverable D10.3 "Second 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 will be described in respect to the progress made, the issues encountered, how these have been solved and what is left to be addressed. It will be clear 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. In addition, any relevant upcoming developments will be highlighted together with plans for incorporation in a future release of the integrated ENES services (D10.5).

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 the release. Sections are broken down by component and where appropriate details of the development of new software and services are also described. As these are under development they may not at time of writing be available as public releases.

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).

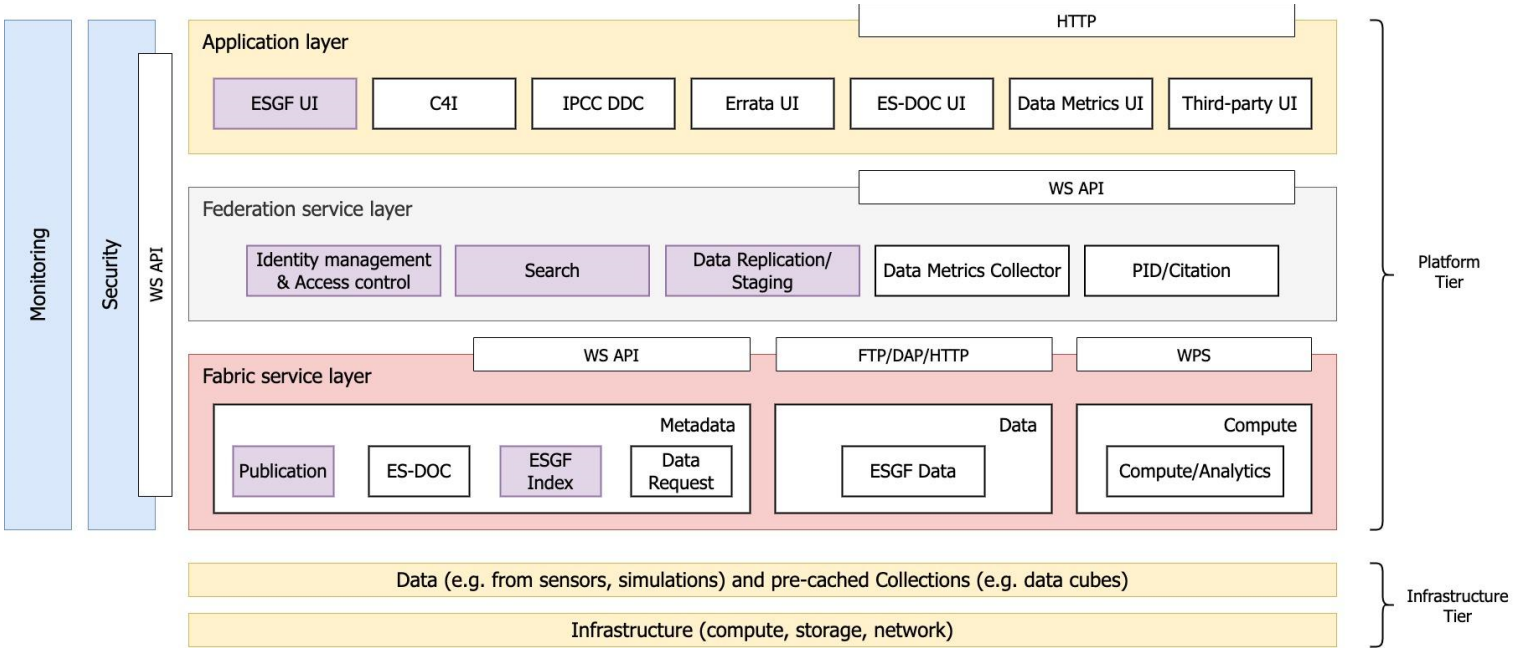

Figure 1. ENES CDI software stack architecture (purple boxes are ESGF services exploited in the ENES-CDI that are associated with collaborative development efforts carried out with partners outside Europe)

Figure 1 above depicts the ENES CDI layers. The document D10.2 [2] fully describes each layer and components with no major changes.

There have been no major updates to the architecture with respect to the first release of the ENES CDI software stack. Figure 2 below shows an updated version of the UML 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.

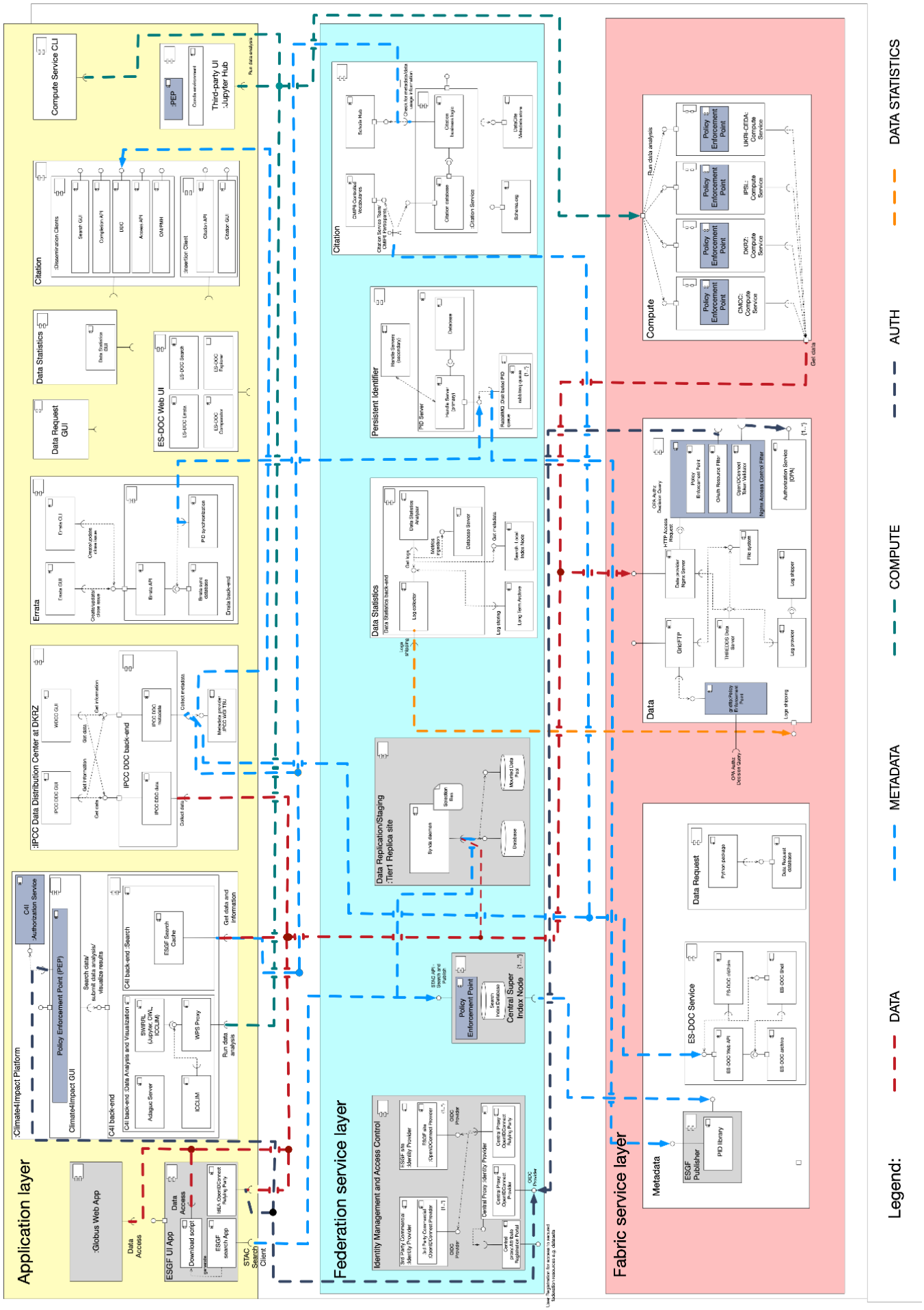

Figure 2. Updated component diagram of the ENES CDI architecture.

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	https://cera-www.dkrz.de/ords/f?p=127:LOGIN_DESKTOP::::::	CMIP6 Data Citation Service	Registration page and database backend and APIs to manage CMIP6 citation information	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/handlequeueconsumer	(restricted access)
IPCC Data Distribution Centre	At DKRZ: http://ipcc.wdc-climate.de At CEDA: http://www.ipcc-data.org/sim/	DDC	Long term curated archive supporting CMIP	-	-	-
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-	**master**

					errata-fe	
		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-ui/	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	0.5.4
		Roocs	ESGF-specific WPS	https://roocs.github.io/	https://github.com/roocs	**0.7**
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://climate4impact.eu/ DEV: https://dev.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 notebooksusing icclim	-	https://gitlab.com/is-enes-cdi-c4i/notebooks	master
		SWIRRL API	Software for Interactive Reproducible Research Labs	https://dev.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	**PROD: 4.2.20** DEV: 5.0b8

Table 1. Second 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 data related components in the ESGF, ie. 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 Second release details

These services are in heavy operational use and were just slightly 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).

[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

An ongoing project called "METAGRID"[7] aims to replace the CoG interface that is no more maintained by US partners. This new search interface is developed in collaboration with US ESGF partners (PCMDI). In the next release, focus must be made on the GridFTP/Globus[8] script for download. The future ESGF release to be deployed in early 2022 will introduce the implementation of Spatio-Temporal Assets Catalogs (STAC[9]). This technology will also offer some front-end solutions, such like STAC-server[10], to replace the old CoG interface.

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. IPCC WGI of the current IPCC cycle integrates data citations in the 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

Due to the deadline of the 6th Assessment Report from Intergovernmental Panel on Climate Change (IPCC) the Citation Service API was retired to ensure no changes of data citations until the publication of the Working Group I contribution. Therefore, modeling centers requested a possibility to maintain the citation information on experiment level, which was only possible via the GUI. Updates include adding article references and changing authors before DOI registration. 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

[7] https://github.com/aims-group/metagrid
[8] extension of the File Transfer Protocol (FTP) for grid computing
[9] https://stacspec.org/
[10] https://github.com/stac-utils/stac-server

experiment entries and thus enables them to focus on those citation entries with remaining tasks.

The automated processing monitoring 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 increase of the DOI registration made it possible to get back to hourly checks for new DOI registration tasks.

Finally, some bugs in GUI and the Scholix process were fixed.

3.2.3 Further tasks and next steps

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 will be closed on the PID side. The idea is to publish these links via Scholix. Options for usage of existing external Scholix Hubs have been investigated. A citation manager survey was carried out to gather feedback and set priorities for the next CMIP phase, CMIP7, Citation Service.

The adjustments to support input4MIPs datasets provided for the activity "CMIP6plus" will be made and priorities based on the survey results and analysis of further user requests will be defined.

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
- 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

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

3.3.1 Second release details

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 Next steps

In the future, besides continuous efforts to correct PIDs where flawed publication/unpublication 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 key findings 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 CMIP6 dataset list of CMIP6 data used in 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 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 archival workflow at DKRZ has been defined and metadata generation has started.

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 will provide the metadata in the agreed form, which requires 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 Guidelines is gathered to set priorities for the next IPCC Assessment Report AR7 and defines the transfer of knowledge and tools to the next assessment cycle.

3.4.3 Next steps

The long-term archival of the CMIP6 data subset underpinning the AR6 is the main task. Others are related to curation of data of older IPCC assessment reports (FAR to AR4) in order to export/map/import all data holdings of the DDC Partner DKRZ into the joint DDC data catalog and to the consolidation of the Quality Assessment software.

3.5 Errata

The ES-DOC Errata service has been the community's answer to an issue raised by the complexity of projects like CMIP5 and CMIP6. The aim is to provide a platform that enables users to record and track reasons that motivate dataset version change. This, of course, should result in a considerable improvement of data quality, given the proper implementation of errata information handling, i.e. timely issue reporting, accurate description and comprehensive updates. The process is well documented and the development team at IPSL offers guidance for first-time users to make the most of the platform. This chapter is a succinct summary of the design, implementation and the change log for the different components of the system.

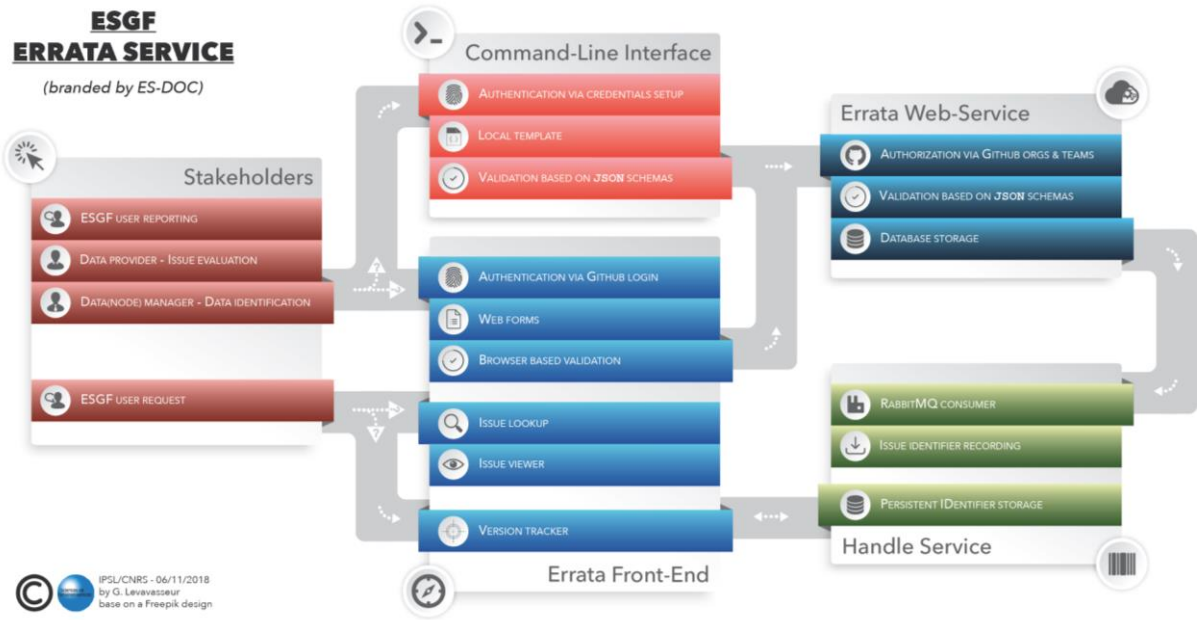

Figure 3. Errata service architecture

3.5.1 Brief reminder of system architecture

As a part of the ES-DOC ecosystem, the Errata Service offers a user-friendly front-end and a dedicated API to provide timely information about known issues affecting ESGF data (see Figure 3).

ESGF users can query for modifications and/or corrections applied to the data in different ways:

- through the centralized and filtered list of ESGF known issues;
- through the "PID lookup" interface to get the version history of a (set of) file/dataset(s).

Contributions to the Errata service are, for the time being, subject to access restrictions based on users' identity and affiliation. For the time being, the ES-DOC Errata system relies on Github Oauth service for authentication and authorization of write operations.

Entries to the service can be made through a lightweight CLI (command line interface) or a web-form.

The ES-DOC Errata system is integrated with the PID system. Each errata pushed to the platform is forwarded and persisted on the PID handles of the appropriate datasets. This enables reverse search and tree of version reconstruction. The system also exposes an API for external third parties wishing to query or archive errata information.

3.5.2 Opening issue registration to all users

We plan on releasing a new version of the backend that enables all users to help identify errata across the different ESGF supported projects, subject to moderation from the data providers. Up to now, an entry to the errata system (or an errata) is sensitive information about climate models and must conform to a specific set of guidelines and rules of formatting. These constraints, although understandable, add extra complexity to the usage of the platform, aim to ensure the highest quality of the information provided to the community. The ES-DOC Errata service is a community effort, and if the community chooses not to invest time and effort into feeding it, it cannot fully serve its purpose. These constraints are fully explained and detailed in the documentation: https://es-doc.github.io/esdoc-errata-client/.

3.5.3 Next steps

The next steps for the errata system are already underway. After a designing phase, the development of a new front-end and back-end to support every user to report issues through a data provider moderation should relax the stress on the latter to report and describe every issue encountered.

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 4).

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 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.

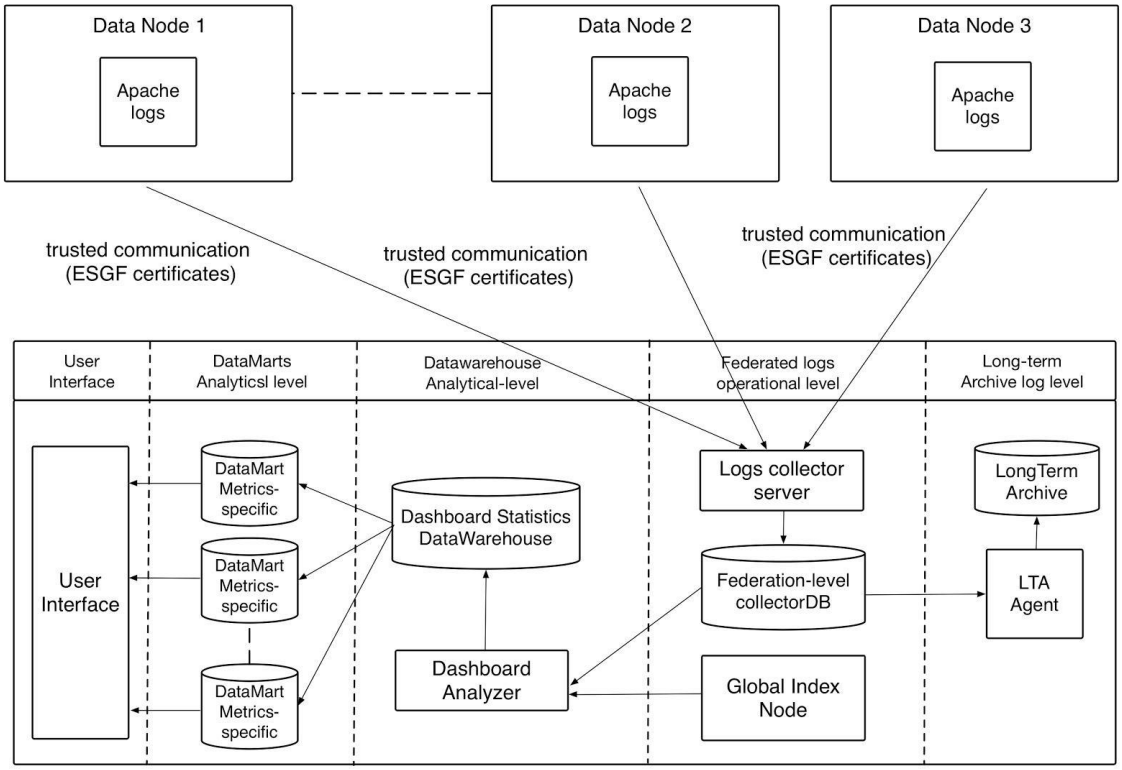

Figure 4. The ESGF Data Statistics Architecture

3.6.2 Second release details

The main developments of the second release of the software regarded:

- the integration of new data nodes (at the Linköping University (LIU) and at the National Center for Atmospheric Research (NCAR)) 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;

- the monthly backup activities of the statistics catalog, bug fixing, performance improving, periodic Logstash certificate update;
- the continuous optimisation of the operational chain and its extension to include new functionalities and the improving of the existing ones;
- the 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.

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.

3.6.3 Next steps

The next activities will be always connected with the enhancing of the user experience in order to provide a better and better understanding of the federation trends in terms of data usage and data publication. A continuous monitoring, as well as performance improving and bug fixing, will assure the proper working of the whole system. New European and not European data nodes will be integrated into the environment.

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") 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 Second release details

The replication software Synda is now distributed in the 3.4 version. The software is packaged through IPSL conda channel making it extremely user-friendly to set up in a dedicated conda environment. To this day, 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 5). We have opted to keep the old user-interface language for the time being to ease the transition for historical users.

At 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.

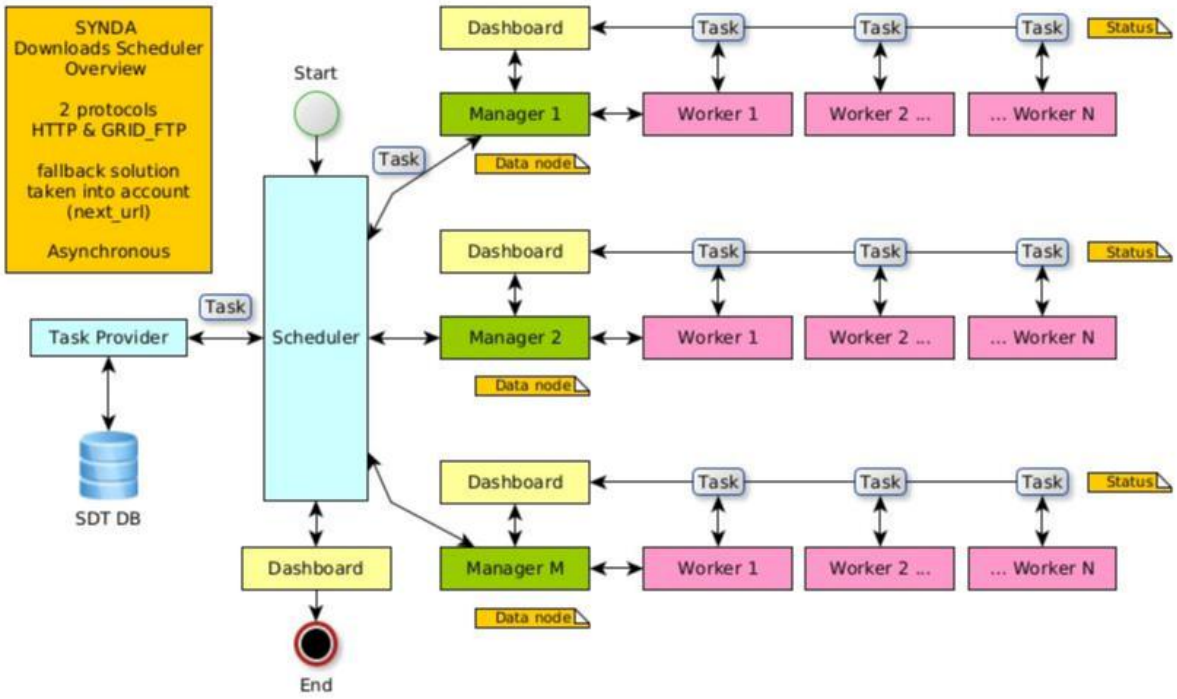

Figure 5. Synda scheduler overview

3.7.3 Next steps

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 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. These efforts will be further reinforced in the last release of the CDI, which will also include 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[12], 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, which are incremental improvements:

[12] http://cfconventions.org/

Software Description	RP2 changes
cfdm: a Python reference implementation of the CF data model. 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 checkerhttps://pypi.org/project/cfchecker/	Implements the new features introduced in CF-1.8 (bar netCDF hierarchical groups).
cf-python: a CF-compliant earth science data analysis library 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 still 1.0.33). 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[13]. A series of meetings was held, and will be reported on in D3.3 Standards Synthesis.

Plans for CMIP7 discussed at the Working Group on Coupled Models annual meeting (WGCM-24) focus on a community consultation to be run in 2022.

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

4.2.2 Next Steps

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 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 6 below. Simplifying features include:

- Removal of central nodes simplifies 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".

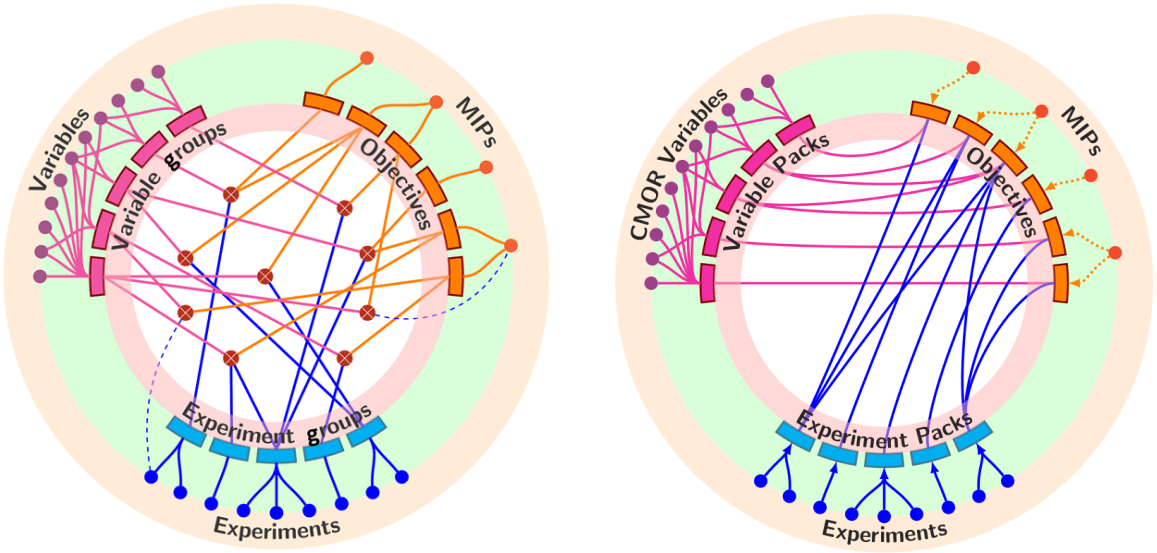

Figure 6: Connectivity diagram for the Data Request schema versions 1.0 and 2.0. (left) Data Request 1.0: nodes link to triples of Objectives, Variable Groups and Experiment Groups. (right) Data Request 2.0: Objectives link to Variable Packs and Experiment Packs

The next steps will be further meetings to review outcomes, followed by implementation and testing of the code. The objective is to have a stable implementation ready at the end of the project, well in advance of CMIP7.

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)

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.

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. They will adopt and further develop innovative technologies to better address the way researchers conduct their analyses, with built in mechanisms for FAIRness [3] and reproducibility [4].

5.1 Climate4Impact

Climate4Impact (C4I) is a portal that enhances the use of climate research data and analysis methods.

5.1.1 Brief reminder of the service

Currently, there are two versions of the service. The first one, which is available to the public[14], had been refactored during RP1 to migrate to a microservice based infrastructure. These developments have been reported in D10.1 and no further updates were applied, besides ensuring the regular operations of the portal. The second version, Climate4Impact v2[15], is a

[14] https://climate4impact.eu
[15] https://dev.climate4impact.eu

renewed release of the system, which also includes a new component for the management of the underlying analysis workspaces, the SWIRRL API[16].

5.1.2 Second release details

The C4I development team pursued this implementation in agreement with the consortium, in order to modernise the software technologies running the portal and to facilitate the support of the advanced components of the ENES-CDI. A first review of the C4I v2 has been already conducted by a team of selected external experts. Its outcome is reported and discussed in D7.2. The new portal is already available to users for testing and evaluation purposes.

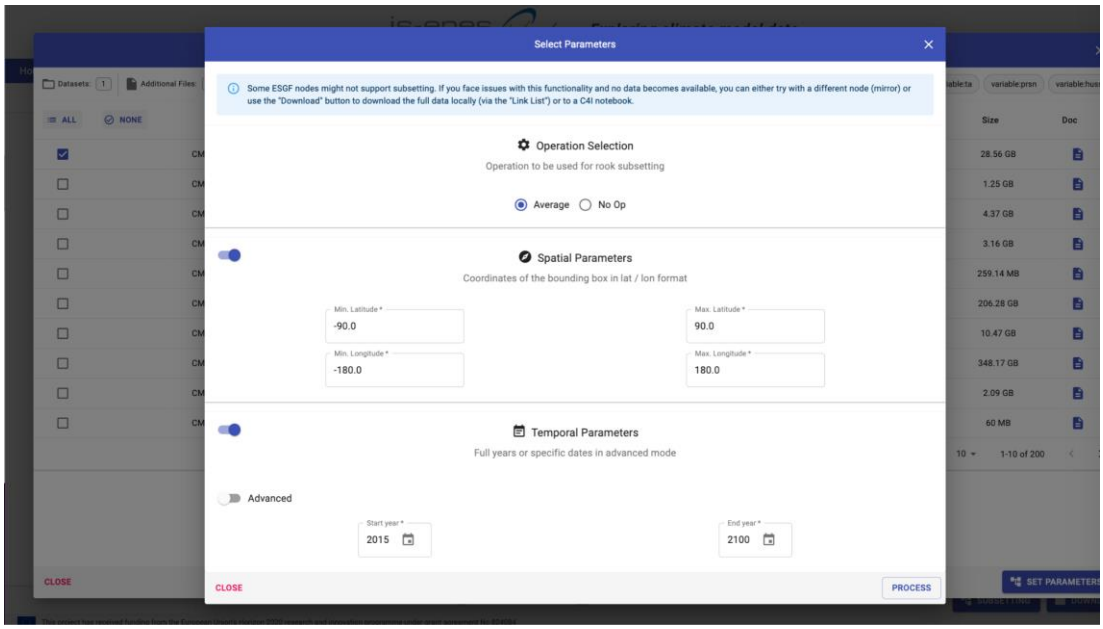

Figure 7. Subsetting workflow parameterization interface in C4I, as optional finalisation phase of the data selection user flow

Thereby, in RP2 our development efforts went into C4I v2, as follows: (1) integration of the new IdP and AAI, in line with the developments of ESGF Future Architecture; (2) implementation of sub-setting workflows within SWIRRL and development of interactive controls in the C4I GUI (see Figure 7). The workflows support OpenDAP and WPS and execute subsetting on selected data. This get stage to the user's workspace, managed by SWIRRL[17], which collect and re-purpose full provenance information; (3) Consolidation of the provision of notebooks in C4I via the SWIRRL API and integration with the ESGF Future Architecture IdP for authorisation checks; (4) improvements to the usability of the web-pages, especially concerning the portal's data and services discovery functionalities. This also includes the production of help-pages, as well as the activation of user support mechanisms, via online contact forms; (5) integration of model comparison pages, in cooperation with NLeSC and WP7.

16 https://gitlab.com/KNMI-OSS/swirrl/swirrl-api
17 https://zenodo.org/record/4264852#.YW8TCXmxWNZ

As anticipated in D10.1, users perform analysis in C4I v2 by means of JupyterLab notebook. This has been enriched with an extension that provides controls to store the code and the environment of the analysis onto Binder repositories in GitLab, as well as monitoring the execution of workflows and accessing provenance information for documentation or to trigger recovery actions. Users can access a collection of pre-set notebooks that can be uploaded and executed in C4I (see Figure 8). Finally, table 1 reports the software components in use and their source code repositories.

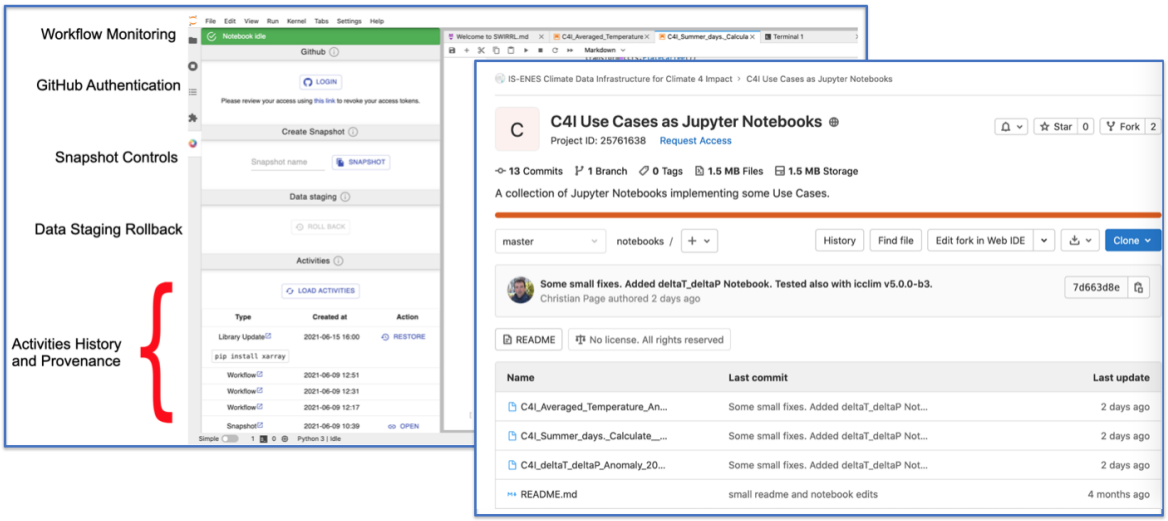

Figure 8. JupyterLab extension (on the right) and gitlab repository with notebook presets (left)

5.1.3 Next Steps

Further activities on C4I v2 will address the consolidation of the authorisation mechanisms which are needed for the dissemination of CORDEX data. This is conducted in cooperation with the data-nodes that will have to support the new ESGF IDentity Provider (IdP). Moreover, in the context of cross-work packages activities agreed by the consortium, RP3 will be dedicated to extend and refine the data-reduction workflows of C4I, aiming at the exploitation of the WPS services that will be deployed across multiple sites, besides the assessment and demonstration of the integration within C4I of the new ESGF data discovery service, based on STAC. Documentation from C4I v1 will be revisited and integrated in v2.

icclim 5.0 will soon be released with a much greater performance compared to version 4.x, especially for percentile-based climate indices. Code has been rewritten from scratch and is very robust. It is now based on xclim, but with a layer on top to support our specific features and to provide a backward compatible API with the previous releases 4.x. The provenance framework specified in D3.3 will be implemented. Documentation will also need to be rewritten. More C4I Jupyter Notebooks (using icclim) will be released, and can be used either with icclim 4.x or 5.x.

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 2021, progress has been a mixture of tasks that were planned for the year; tasks brought forward from 2022; and some unexpected work related to the commercial cloud hosting.

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 yet, but the needs of the obs4mips community have been discussed with a view to making a subset of the ES-DOC functionality available during 2022.

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 so that other members of the team (or anyone else) could learn how to deploy the ES-DOC services.

5.2.3 Next steps and progress on tasks

The documentation of the ES-DOC software stack itself has been brought forward and has been a considerable amount of work. The purpose of this documentation is to enable anyone to be able to maintain and develop the ES-DOC services in the future, if (or when) the current ES-DOC team is not able to support the project any longer. In addition, the knowledge transfer between existing staff will also help provide a better level of maintenance for the duration of the IS-ENES3 project.

During 2022, the highest priority activity is to complete the provision of all types of CMIP6 documentation. This means putting into production the automated "cdf2cim" process that automatically publishes the descriptions of every CMIP6 simulation (the cdf2cim service already running on the ESGF nodes has been providing raw information to the ES-DOC servers for some time, but no documents have been publicly published yet); and providing the ability of the CMIP6 groups to provide and publish conformance to numerical requirements documentation.

Support for obs4mips and, if appropriate, for the input4mips project will be provided.

5.3 Institutional compute service deployments at ENES CDI sites

The ENES CDI aims to foster the "data near processing capabilities" paradigm. In this second release, steps have been taken in this direction, in close collaboration with the activities performed in WP5/NA4, to provide an increasingly 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

As shown in Figure 9, the CMCC compute service, also named CMCC Analytics Hub, consists of multiple components: (i) an interface/GUI providing an Open (data) Science-ready environment for Data Science applications, interactive and exploratory data analysis, visualization, etc.; (ii) a workflow-enabled, secure, and interoperable front-end; (iii) an analytics framework back-end to perform data analysis at scale and support metadata management; (iv) a data collector and its local storage to gather the relevant datasets from ESGF and keep them synchronized with the remote repositories, as well as other auxiliary services for publication and sharing of results and code.

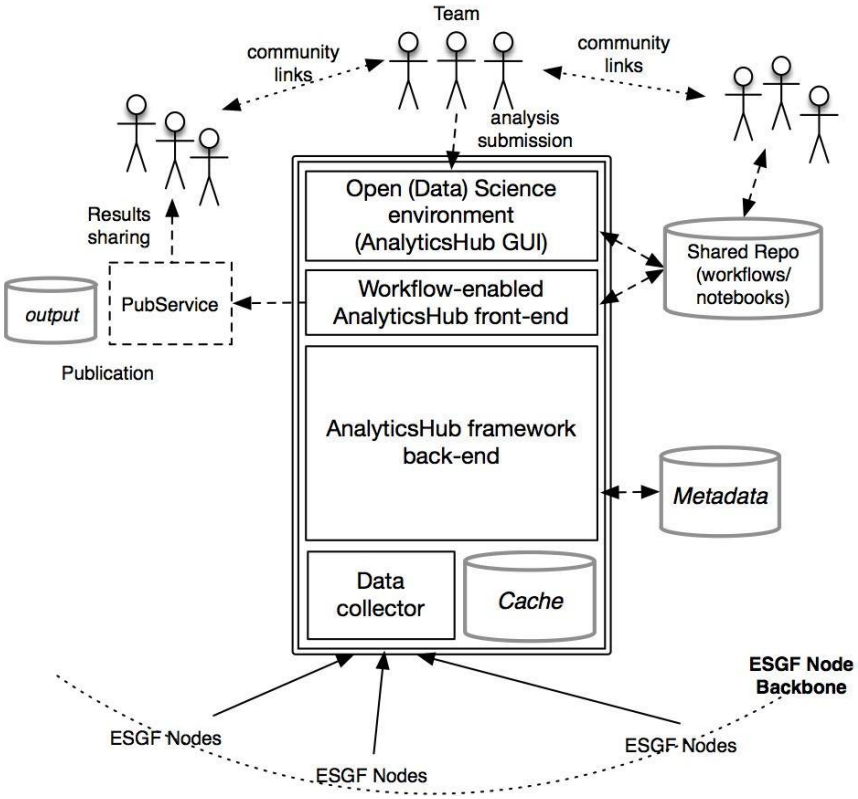

Figure 9. CMCC Analytics-Hub architecture

5.3.1.2 Second release details

According to the scheduling listed in the deliverable D10.2 [2], the following activities have been carried out towards the second release of the CMCC compute environment.

The CMIP6 data catalog has been extended with the inclusion of new files related to the temperature (*tas*) variable, with a *3hr* frequency and for the *historical*, *ssp245* and *ssp585* experiments.

The Jupyter-based data science environment has been updated and deployed on a more powerful server in order to better support user applications. In particular, the target programming language of the environment has been updated to Python version 3.9. Moreover, PyOphidia (the Python interface to the Ophidia framework) has been improved to better support results visualization on map and some bug fixing activities led to better operation of the Ophidia server. Finally, general improvements, at the level of users' management for the Ophidia system, have also been carried out.

On top of this environment, a scientific portal has been released, with the aim to provide users with timely information about the capabilities of the CMCC Analytics-Hub, as well as documentation to allow easy access to the Jupyter environment. It is available at the following link: https://ecaslab.cmcc.it/web/home.html, and consists of i) a general landing page (see Figure 10) introducing the environment and describing its main capabilities, from the JupyterHub service to the available libraries and Notebooks and the CMIP6 data catalogue; ii)

a second section represents a gallery with a list of Jupyter Notebooks already available into the environment; also, iii) two dedicated sections allow users to register and access the environment and iv) a final section with the contact details.

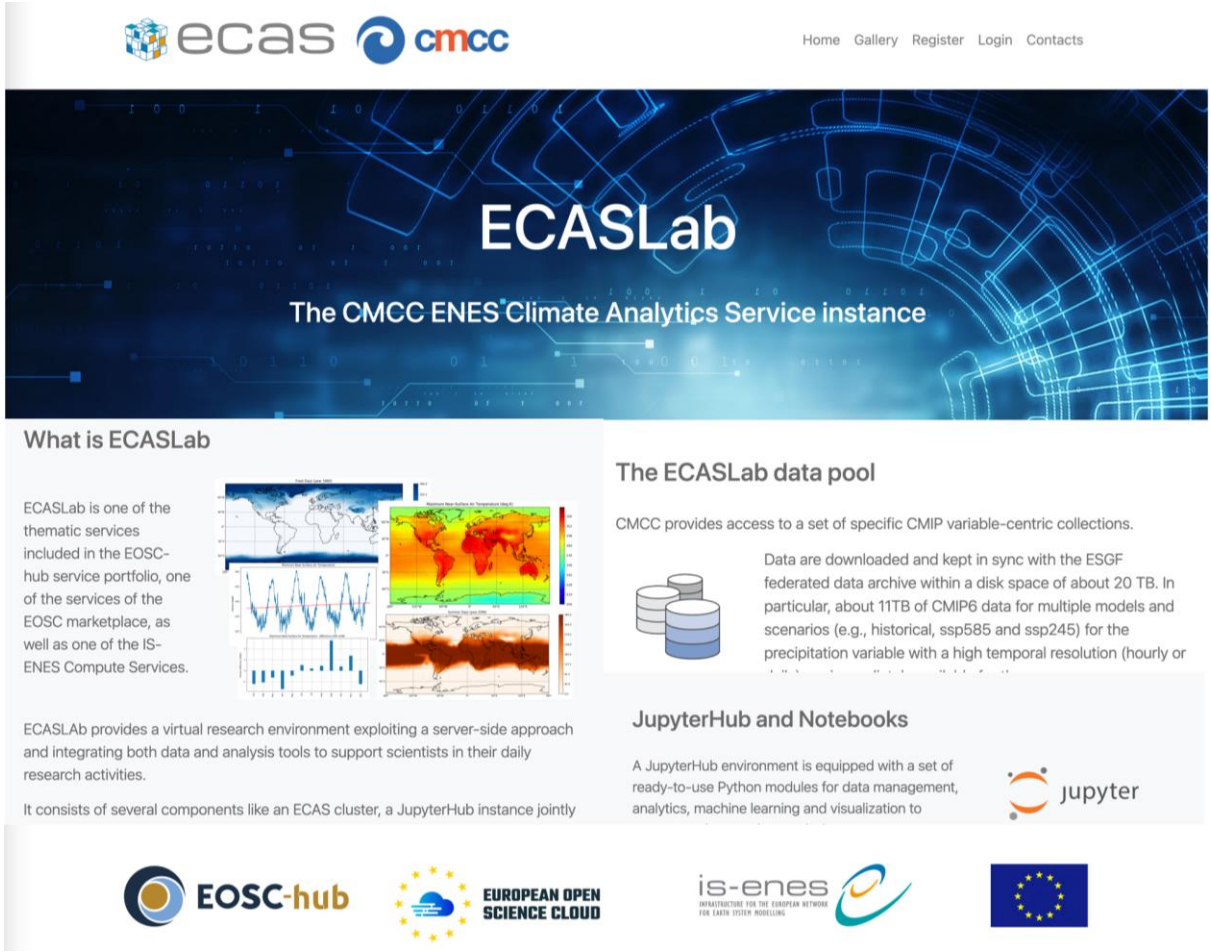

Figure 10. CMCC Analytics Hub portal landing page

5.3.1.3 Next steps

Future activities will be related to the improvement of the CMCC Analytics Hub in terms of i) inclusion of additional sections/functionalities of the scientific portal, ii) bug fixing and performance improvement of the overall environment, iii) extension of the data catalog with new variables based on user requirements and/or the most downloaded variables ranking available through the ESGF Data Statistics (see Section 3.6) user interface.

Moreover, in the next months, `Sproket`, the current download tool used to create and enrich the Analytics Hub data catalog, will be progressively replaced by `Synda`, which is well supported and documented and ensures increasing download and data catalog synchronization performance.

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 or values.

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, which has been developed in close collaboration with DRKZ, 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).

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;
- 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 Next steps

The next steps for the "roocs" stack will include:

- support for CORDEX data
- temporal averaging: monthly and annual

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).

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.

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

CEDA, 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/). 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 (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 11). The result of the subsetting operation is provided for download to the requesting client (Climate Data Store).

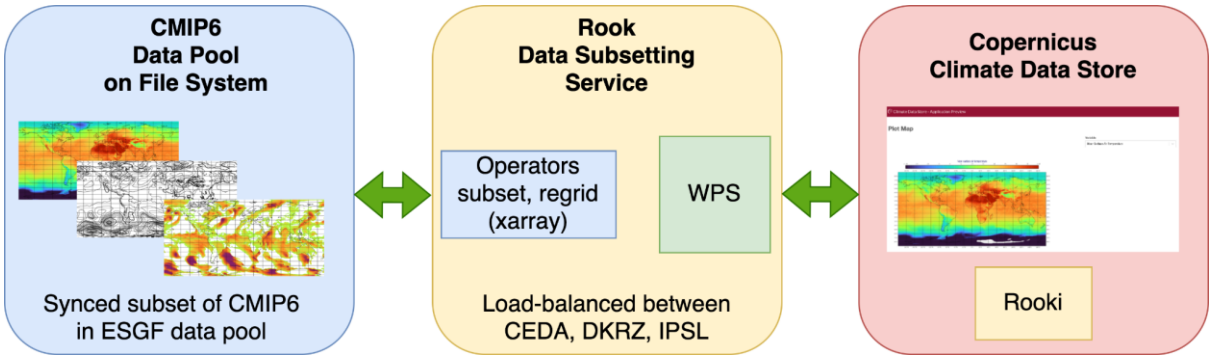

Figure 11. Rook service implementation

In addition to the subset operator we currently work on the average and regrid operator. The Rook service is using the OGC Web Processing Service Standard[18] 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

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

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 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 Second 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 VREs (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)[19]" developments. Functionalities now include temporal/spatial subsetting as well as regridding (testing phase).

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

5.3.3.4 Future work

Starting in 2022 a new HPC system will be deployed at DKRZ and thus preparations are ongoing to migrate the existing 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 will be 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 IPSL

The generic outline of the compute service at IPSL was provided as part of deliverable D10.2 [2] and has been reinforced in this release.

5.3.4.1 Brief reminder of the infrastructure

A third of the computing capacity has been renewed and increased by 25% bringing computing resources to 2500 CPU cores together with 8TB of RAM. The IPSL computing center deployed 4 GPU cores together with 256GB of RAM to answer IA computing facilities and climate services needs.

5.3.4.2 Second release details

Pre-configured Python virtual environments have been improved that activate mutualized and 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).
- "analyse" provides a base environment for data analysing including the well-known Xarray[20] and Dask[21] libraries that provide user-friendly I/O and parallel tasking.

Despite the compute service design at IPSL still mainly relies on generic remote access to dedicated login nodes (see [1], Figure 32), all above Python environments are finally accessible through a JupythHub recently deployed (https://data.ipsl.fr/jupyter) and open to any IPSL computing centre users.

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

[20] Xarray: http://xarray.pydata.org/en/stable/
[21] Dask: https://dask.org/

climate modelling production from IPSL and CNRM). The data access now benefits of "intake-esm"[22] catalogs that would ease transition towards STAC[23].

A new Web Processing Service has been deployed on a 8CPU machine and dedicated to Copernicus needs. This WPS is based on the "Rooc"[24] project led by CEDA and DKRZ partners.

A Kubernetes has also been deployed at IPSL but its production status is delayed due to missing staff allocated to the infrastructure itself. Priority was given to the IPSL JupytherLab deployment. For the time being, the Kubernetes is for educational use only.

5.3.4.3 Next steps

Dedicated Python environments are under development:

- "esmvaltool" aims to provide a pre-configured instance of the last version of the ESMValTool.
- "cmor" will provide a pre-configured instance of CMOR tool to standardized CMIP and CORDEX data produced by IPSL climate models.

The "Dask-jobqueue"[25] plugin will be used to interface Dask with the usual IPSL PBS (Portable Batch System) manager (ciclad-web.ipsl.jussieu.fr).

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

6 Identity Management and Access Entitlement

Under the ESGF future architecture initiative, a new implementation of the identity and access entitlement (IdEA) system has been under development since March. Since reporting for D10.2 all components have been completed to at least the level of prototypes. Integration testing has been conducted between partners trialing the new authentication technologies.

6.1 Current 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 OAuth 2.0 for delegation of authentication.

[22] intake-esm: https://github.com/intake/intake-esm
[23] STAC: https://stacspec.org/
[24] Rooc project: https://github.com/roocs/
[25] Dask-jobqueue: https://jobqueue.dask.org/en/latest/

- In some cases, where simple authentication is required, GitHub's OAuth 2.0 service is used.

6.2 Implementation status for Future Architecture Components

6.2.1 Authentication, single sign-on and user delegation

The new 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

The Identity Provider (IdP) Proxy 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 are based on customisations of the open source implementation Keycloak[26] from RedHat.

1. **IdP Proxy**: implementation complete; Docker image completed; Deployed for integration testing on JASMIN. Application is in the process of being deployed on AWS
2. **Federation site IdP**: completed. Deployment ready for CEDA. Docker image

6.2.3 Relying Party and Policy Enforcement Point Implementation

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.

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 12), PDP (aka. Authorisation Service, bottom right in Figure 12) and

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

Attribute Service. In the new system (see Figure 12), 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)[27]. 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 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. 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.

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

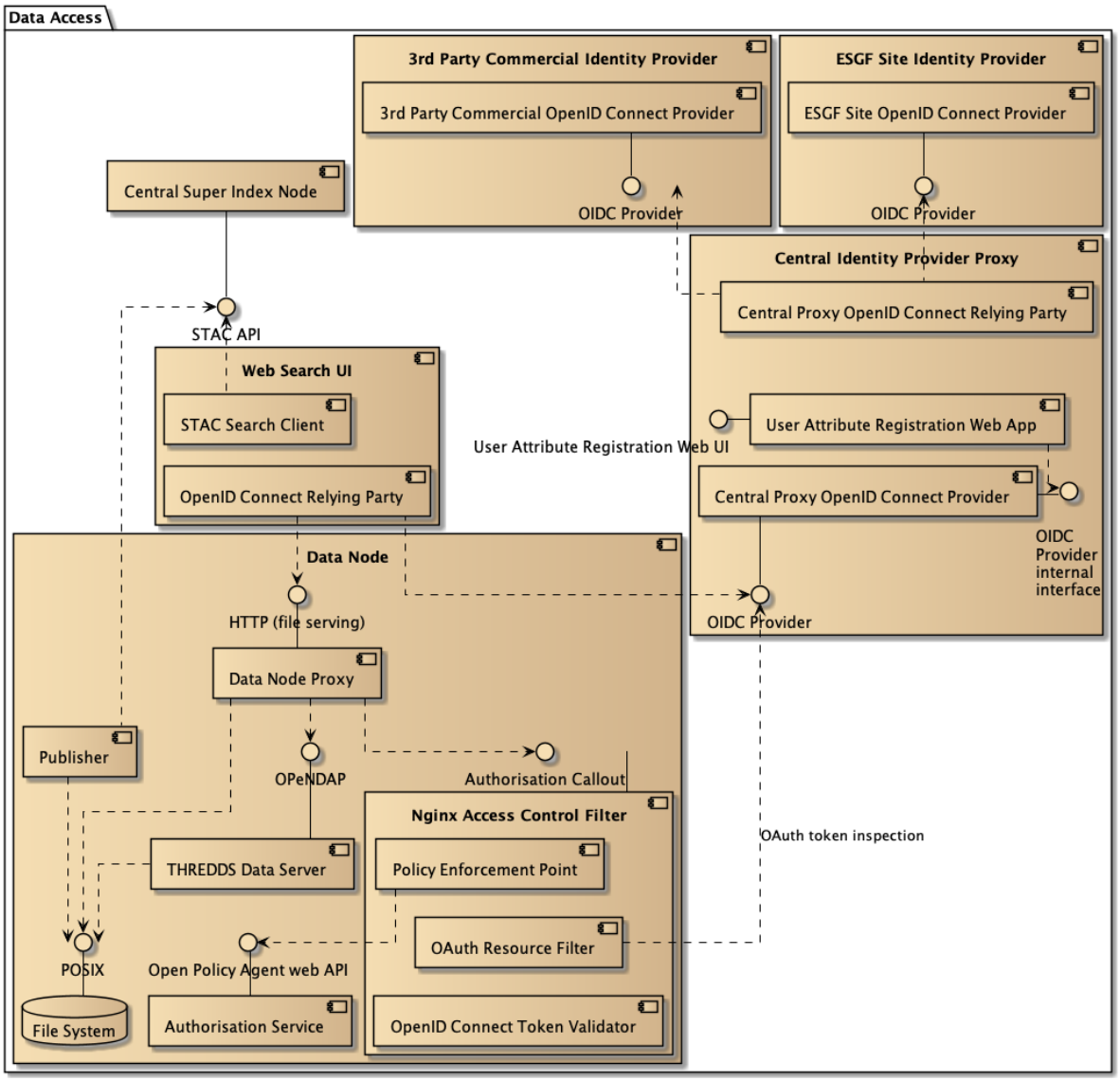

Figure 12. ESGF Future Architecture showing identity services and index and data nodes

7 Conclusions and main targets of the next release

This second 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 release of the ENES CDI [1] 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 (new nodes & widgets), and a major release of the data replication tool ("synda").
- 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). 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.

Important targets for the next and last release will:

- Address long-term archival of the CMIP6 data subset underpinning the AR6 and the curation of data for older assessments.
- Finalize the new errata interface opening issue registration to any users.
- Pursuing the development effort on "synda" replication tool and particularly its discovery module relying on the ESGF Search API.
- Deploy the Policy Enforcement Point (PEP) and federated Identity Providers on Tier 1 sites within the new ESGF architecture.

We will also consider further updates to the architecture : despite adoption of containerisation gaining attention, the acquisition of object storage capacity with some caching strategy will be a must towards data-ready analyses and hosting cloud-based services.

Finally we will take into account progress made with the integration of components produced in other work packages. For instance, we will pursue the connection of Climate4Impact to the

model evaluation system based on the ESMValTool[28] (M7.3), to provide access to relevant metrics in support of the selection of model data. This is in cooperation with WP9/JRA2. Progress will be reported in the third release (D10.5) of the ENES-CDI.

[28] https://cmip-esmvaltool.dkrz.de/

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.