IS-ENES3 Milestone M10.1
Technical requirements on the software stack

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

Authors:
S. Fiore (CMCC), P. Nassisi (CMCC), F. Antonio (CMCC), L. Barring (SMHI),
K. Berger (DKRZ), D. Hassell (UREAD-NCAS), M. Juckes (UKRI),
P. Kershaw (UKRI), S. Kindermann (DKRZ), G. Levavasseur (IPSL),
A. Nuzzo (CMCC), C. Pagé (CERFACS), A. Stephens (UKRI),
W. Som de Cerff (KNMI), M. Stockhause (DKRZ), T. Weigel (DKRZ)

Reviewers:
P. Kershaw (UKRI), M. Lautenschlager (DKRZ)

Release date:
March 26th, 2020

ABSTRACT

This report addresses the milestone M10.1 "*Technical requirements on the software stack*" of the IS-ENES3 project and provides a comprehensive list of technical requirements driven by the work done in WP5/NA4 "*Networking on data and model evaluation*" and WP3/NA2 "*Community engagement*" as well as by previous meetings and workshops in the community, like the ESGF F2F Conferences. It represents the first step towards the design of the ENES Climate Data Infrastructure (ENES CDI) software stack architecture. The document adopts the concepts of user stories and use cases to translate them into functional and non-functional requirements for the whole ENES CDI software stack.

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

TABLE OF CONTENTS

1. INTRODUCTION

This report lays the foundation for evolving and consolidating the ENES Climate Data Infrastructure (CDI) software stack that, according to the overall IS-ENES3 [1] project objective #3, aims to *support the exploitation of model data by both the earth system science community and the climate change impacts community*. The software components involved are developed and maintained as open source efforts by the IS-ENES partner institutes with contributions from the international Earth System Grid Federation (ESGF) [2] developers community.

The document addresses the milestone **M10.1 "Technical requirements on the software stack"** of the IS-ENES3 project, within the WP10/JRA3 "ENES Climate Data Infrastructure software stack developments". It provides a comprehensive list of technical requirements, driven by the work done in WP5/NA4 "Networking on data and model evaluation" and WP3/NA2 "Community engagement" as well as by previous meetings and workshops in the community (i.e. ESGF F2F Conferences). Moreover, it represents the first step towards the design of the ENES CDI software stack architecture that will be described in the Deliverable D10.1 "Architectural document of the ENES CDI software stack", to be delivered at month 18.

The document provides a general overview of the key software components currently developed for the ENES community and deployed as services at the European data centres, before delving into the details of the user stories, use cases analysis and list of technical requirements for the whole ENES Climate Data Infrastructure software stack.

The rest of the document is structured as follows: Section 2 proposes an introduction to the IS-ENES3 project, its main goals, stakeholders and correlation with ESGF and other European data ecosystems. A general overview of the current status of the ENES CDI is presented, introducing the list of software components such as the core data distribution and compute services, vocabulary management, documentation and impact study tools. Section 3 identifies the main functional and non-functional requirements for the overall software stack and goes into the details of each service involved. Section 4 presents a preliminary design of the architecture for the ENES CDI software stack while Section 5 drives the main document conclusions and makes a brief analysis of the outcomes. Finally, Appendix A reports a comprehensive summary of all the technical requirements in a tabular form while Appendix B presents the templates used in Section 3 to describe the scenarios in which the components are involved.

2. ENES Climate Data Infrastructure

2.1. Overview

Users from the climate modelling community, the climate impact community as well as interdisciplinary research domains rely on stable and consistent services to access and to process the high-volume climate model data of WCRP [3] reference simulations from CMIP [4] and CORDEX [5], hosted in distributed repositories across Europe. To this end, the ENES CDI is the primary source for climate model data in Europe (originating from both European and international modeling groups). It provides access to data from ESGF, from the WDCC [6] archival system and from the Climate4Impact portal [7], as well as processing services, documentation and standards. From the perspective of the software stack, the ENES CDI consists of a set of services related, for instance, to core data distribution/management, compute and analytics, vocabulary management, documentation and impact studies. The software developed in WP10 provides the building blocks for the setup of the ENES CDI service infrastructure managed in WP7; its long-term sustainability plan is part of WP2. Relevant, in terms of community requirements gathering and networking activities in the overall international context, are WP5 and WP3.

2.2. Main goal and high-level objectives in IS-ENES3

The IS-ENES3 project will evolve and consolidate the ENES Climate Data Infrastructure software stack according to the overall project objective #3 (*IS-ENES3 will support the exploitation of model data by both the earth system science community and the climate change impacts community*).

In IS-ENES3, the high-level objectives with respect to the ENES CDI software stack aim to:

- design, improve, extend and consolidate the ENES CDI software stack as a basis for a sustainable, streamlined and scalable climate model data distribution solution for users in the climate modeling and the climate impact research and modelling communities.
- Support interoperability of data files and archives for automated data processing through improved and extended standards and metadata.
- Provide an interoperable and flexible computing layer supporting scientific data analysis and processing within the infrastructure, by evolving existing solutions towards an integrated set of service offerings for end users (from climate researchers

to the climate impact research and modelling community including long-tail end users).

- Evolve the Climate4Impact platform towards a climate data analytics portal for impact scientists. This is done by providing advanced data processing services and data access services in the C4I portal. The functionality will be made available through user friendly interfaces (e.g., tailored search interfaces, guided wizards, Jupyter notebooks with use case examples).
- Maintain and develop the ES-DOC international documentation infrastructure to support CMIP6 and other MIPs as well as expand the scope of documentation to new areas for the climate modelling process, including model evaluation.

2.3. ENES CDI and ESGF

The Earth System Grid Federation is based on a software architecture that provides access to a federated data archive distributed across multiple sites (called ‘‘Nodes’’) that are geographically distributed around the world but can interoperate due to the adoption of a common set of services, protocols and APIs [8].

Through the ENES-CDI, IS-ENES3 provides the European contribution to ESGF in terms of software (WP10) and services (WP5). However, the ENES-CDI also provides room for (i) EU-level/national specific developments, which do not target ESGF directly, (ii) early software developments, which would need a pre-production phase for testing and validation before moving into a wider production environment.

Additionally, IS-ENES3 activities include (i) the contribution to the future ESGF architecture (WP5), which spans across a longer timeframe with respect to IS-ENES3 and (ii) the IS-ENES Sustainability plan (WP2), which of course includes the ENES CDI too.

The following table summarizes and clarifies the link between the aforementioned activities in the project.

Table 1. IS-ENES3 contribution to the ENES-CDI and ESGF

IS-ENES3 WP	Context	Target	Timeframe
WP2 “Governance, Sustainability and Innovation”	ENES CDI	Sustainability of the ENES CDI	Long term (decade)
WP5 “Networking on data and model evaluation”	ESGF	Software stack architecture	Long term (decade)
WP7 “Data standards, distribution and processing services”	ENES CDI	Operational Services	Short/Medium (IS-ENES3 timeframe)

WP10 “ENES Climate Data Infrastructure software stack developments”	ENES CDI	Software stack architecture and development	Short/Medium (IS-ENES3 timeframe)

From the table above, it can be easily inferred that ensuring the proper link between the ENES CDI software stack architecture and the future ESGF architecture is essential in terms of design perspective and towards a common long-term, sustainable objective. In this perspective, through the ENES CDI contribution, IS-ENES3 will contribute to short/medium term actions in the long-term ESGF roadmap with resources as in the past that are balanced with the US contribution.

2.4. ENES CDI and EOSC/Copernicus

The ENES CDI will be designed by also considering the ongoing efforts in the wider European *data ecosystem* and it will look forward to the EOSC [9] roadmap and evolution as well as to the Copernicus [10] landscape. In particular, the design of some specific components, like the compute service, will take into account the link with EOSC e-infrastructures (EGI [11], EUDAT [12]) and the Copernicus C3S tools and platform.

2.5. Stakeholders

As part of the ENES CDI architectural design phase, we provide in this section a list of stakeholders identified at this stage, starting from the bottom (IT platform) to the top (community governance and funding agencies).

Support

- Infrastructure-level (i.e. sys-admin IT facilities)
- Application-level (i.e. ESGF data node admin)
- User support (i.e. MIP technical support, ESM technical support)

Service providers

- Data service providers
- Metadata service providers
- Compute service providers

Data providers

- Modelling groups providing simulations output (i.e. MIP science team, ESM science team)

- Sensors providing observations (i.e. Obs4MIPs [13] experiment)
- Specialist data producers (e.g. downstream providers, datacube/cache providers, etc.)

Data publishers

- IT experts publishing data into ESGF

Scientific end users (Earth system scientists, climate change impacts scientists, climate data scientists)

- Analysis users/service
    - big data users (require access to large computational facilities, i.e. compute services via TNA)
    - medium data users (require access to medium-size computational facilities, i.e. compute services via VA)
    - small data users (do not require access to any computational facility, i.e. desktop users requiring access to a data node only for data download)
    - applications/services (machine-to-machine interaction)
- Article writer, reviewer, IPCC author

Community/Services Governance, project coordinators and funding authorities

- Funding agencies (i.e. EU commission)
- ENES board representatives, ENES Task Forces and IS-ENES PO
- ESGF Executive committee
- Governance bodies (CMIP panel, WIP, CF Convention)
- Project coordinators (i.e. MIP, ESM)

2.6. List of Services

The overall goal of the architectural design phase, which will be finalized by month 18 and reported in D10.1, is to define the ENES CDI software stack architecture. Before going into the details of the different levels of the stack (which is out of the scope of this report), we provide in this section the list of the ENES CDI list of services, with a brief description summarizing its current status and the main goal. Such material provides the proper background for Section 3, where the technical requirements will be defined and listed.

2.6.1. ESGF Data

The ESGF is a federation of several European and non-European institutes and provides a service for data publication and distribution. Therefore, the underlying ESGF software consists of multiple components, including a tool for data organization and preparation ('esgprep'), a tool for data publication ('esgpublish'), a federated search service and multiple download options such as HTTP, wget and Globus [14], GridFTP download.

To prepare the data for publication, 'esgprep' provides the functionality to organize the data on the filesystem, i.e. to build a homogeneous directory structure for better file management.

During the 'esgpublish' process, various metadata are extracted from the directory structure and also from the files themselves. This metadata follows a structure determined by the ESGF DRS (Data Reference Syntax), a set of controlled vocabularies to support the organisation of the data and search services. Different projects hosted on ESGF may adopt different DRS definitions according to the structure of the data being hosted and specific user requirements. The metadata extracted is published to a Solr [15] index in order to support faceted search.

In addition to the automated search via the ESGF SearchAPI, ESGF also provides a graphical search interface (CoG) for less experienced users. For data download, the ESGF supports the download of single files via HTTP as well as the download of many files using a script-based download and a Globus/GridFTP interface.

2.6.2. Citation

The Citation service is a standalone service able to create and maintain data references for CMIP6 data collections on model/MIP and experiment granularities. It consists of some core components working in close collaboration with other ESGF community services with the aim of making CMIP6 data citable in scholarly publications.

The citation service exploits the DataCite [16] service for Digital Object Identifier (DOI) registration and DataCite’s catalog metadata insertion and update.

The insertion of the citation entries is based on the project’s controlled vocabulary, developed and managed by communities of domain experts, that describes in detail the scientific properties of Earth system models and simulations. To easily insert citation entries, two kinds of services have been provided: a GUI based on Oracle Application Express (APEX) [17] and a REST API inserting DataCite-conformant metadata provided in JSON. The service contacts the ESGF Solr to check for newly published data and keep data references always up-to-date.

On the other hand, detailed citation information is provided through different kinds of services: (i) on dedicated DOI landing pages, (ii) as XML files for harvesting via OAI/PMH [18] and (iii) as JSON-LD [19] integrated in the DOI landing pages. An interface to the ESGF Data Node Manager and ES-DOC is also provided, to access status information on citation metadata and DOI registration.

CMIP6 data users are obliged to use these data citations in the reference lists of their publications. The citations will be visible in the ESGF CoG portals displaying core citation information and data license. Service usage will be monitored to offer DOI statistics and integration of articles citing the data as additional references.

Finally, user support is a key aspect of this service, put in place with a strong connection and collaboration with data providers and project partners in the ESGF community.

2.6.3. Persistent Identifier

Digital objects are often affected by changes in their location or ownership, making it hard to reference them through web URLs. In scientific publishing, it is therefore not recommended that any kind of material be referred to using URLs, as any future reader may have difficulties discovering relevant information in case the management of the respective web servers fails.

The Persistent Identifier (PID) service aims to address this issue, assigning a unique name that can be used to refer to a digital object and possibly retrieve it even through changes of object location or ownership. The concept of PIDs therefore relies on an act of explicit registration with a Naming Authority and making sure that the management processes of a repository such as ESGF covers such changes. As part of the registration, the identifier is associated with some minimal amount of information, including the object's current location. If the object location changes, the information in the registry can be updated so that the binding between identifier and current location does not break.

This concerns both the final stage of data when it receives a DataCite DOI and also preceding stages where data is shared via ESGF before formal data publication of the final datasets. The PID services described in this section manage PIDs used for the latter case.

The PID services consist of:

- an adapted Climate Model Output Rewriter (CMOR) [20] software library to generate PID names;
- an esgf-pid Python client library (connected to the ESGF publisher) to send off PID registration and update requests;
- a distributed messaging service, implemented as a RabbitMQ [21] federation with four entry nodes and one exit node and specifically configured to route esgf-pid registration and update requests;
- esgf-pid server-side components (Java servlet), deployed as add-ons to a Handle server and connected to the RabbitMQ exit node, to execute PID registration and update requests;
- modified ESGF search and view pages to display PID names and context information;
- a web-based data cart PID registration service to enable end-users to create custom data carts with stable PID names, and
- CMIP6 PID viewer information pages (dynamic web pages generated via a node.js/react.js server application) to display detailed PID context information on files, datasets and carts.

The fine-grain levels of files and datasets are generated and curated by the combination of CMOR, publisher, esgf-pid client and server components. The coarse levels are taken care of by the data cart PID web pages for registration and viewing. The PID viewer web pages are relevant for all levels.

2.6.4. IPCC Data Distribution Centre at DKRZ

The IPCC DDC at DKRZ has extended its services from long-term archival of the CMIP data snapshot in the DDC Reference Data Archive in the 5th Assessment cycle to additional services supporting the Assessment process.

These additional services consist of:

- contribution to developing a concept aimed at improving the transparency and traceability of IPCC key results;
- providing Virtual Workspaces for the IPCC authors as collaboration platforms and for easy access to the CMIP6 data (connected to the Data replication service);
- IPCC author support.

2.6.5. Errata

The Errata Service offers a user-friendly front-end and a dedicated API to provide timely information about known issues affecting ESGF data. ESGF users can query about 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).

All ESGF projects with a pyessv CV [22] are currently supported by the Errata Service. It allows identified and authorized actors of the corresponding modelling groups to create, update and close issues using either a lightweight CLI and/or a new and easy web form.

In June 2018, the IPSL moved the ESGF Errata Service to production phase; the services are available at [23].

As a part of the ES-DOC ecosystem, the Errata Service exploits the Persistent IDentifier (PID) attached to each dataset and file during the ESGF publication process. The documentation [24] has been fully revised to guide users through the errata procedure.

2.6.6. Data Statistics

The ESGF Data Statistics service is a software component responsible for federating, collecting, and reporting data usage and data publication statistics about the entire ESGF Federation, thus providing a better understanding of the amount of downloaded data, along with the most downloaded datasets, the data published in the federation, and so on. The latest release of the Data Statistics service relies on a pipeline based on industry-standard tools (Logstash [25] and FileBeat [26]). It provides a lightweight log collection system deployed on the ESGF data nodes responsible for sending data usage logs (filtering out sensitive information prior to dispatch) to the central collector node deployed at the CMCC SuperComputing Centre. Such log transfer is performed in a secure way, via trusted communication leveraging ESGF certificates.

Several nodes on the collector side host Logstash instances to gather log information; there is also a Data Statistics analyzer (big data) pipeline responsible for processing all log entries, as well as a Long-Term Archive module to address log preservation, and a User Interface [27] with a rich set of charts and reports that allows end users to visualize a wide range of statistics.

2.6.7. Data replication

ESGF relies on a two-tier architecture to enable efficient and reliable data delivery to end users: tier1 nodes replicate data from tier2 nodes and from other tier1 nodes, thus providing additional copies of often used data collections. The data replication service is used primarily at tier1 centres to coordinate and efficiently manage large-scale parallel data transfers from other ESGF data notes. The data replication service is based on the Synda software package deployed in operational environments at ESGF tier1 centres. Synda provides an easy interface to specify data collections to be replicated (in so called "selection files" based on a facet characterization of the datasets). Based on an internal database, it manages the transfer state of the matching data and the actual transfer is based on (parallel) HTTP and also GridFTP connections to remote data nodes. Complete data replication includes an additional ESGF data publication step to make replicated datasets also visible as replicas in the ESGF search index. Whereas Synda also provides some support for this, in the actual operational workflow of the IS-ENES tier1 ESGF sites (CEDA, DKRZ, IPSL), this is implemented as a separate step based on the (site-specific) existing ESGF publication workflow (often with manual intervention). Generally, there are additional aspects which need to be guided by policies to keep replica pools in sync with the changing ESGF data holdings and in sync between sites based on Synda.

As part of the ESGF publication of CMIP6, replicas are also marked as replicas in the PID system, thus allowing not only the derivation of the datasets version history but also the replica information from the persistent identifiers of datasets.

2.6.8. Compute & Analytics

Considering the large and continuing increase in the volume of climate data in recent years, the ENES Research Infrastructure (RI) can no longer solely focus on data storage and federated data access. The need to keep data processing as close as possible to data storage is now strongly required by providers and users. Within the IS-ENES3 project, Compute Services will be developed as part of WP10 to extend the service portfolio offered by the ENES CDI.

Several institutions participating in the IS-ENES3 project are involved in the compute service activities. Several approaches and solutions have been developed, which share some common aspects but also differ from each other due to the complexity of the target environment as well as the application use case they want to address in the compute and analytics area.

The list of institutions involved in the development of compute service solutions in the IS-ENES3 project includes: CMCC, DKRZ, UKRI, IPSL. Additionally, institutions developing applications/tools on top of compute services are CERFACS and KNMI. This activity will also intersect with the work on climate diagnostics and climate indices standards.

2.6.9. Climate4Impact

The IS-ENES Climate4Impact (C4I) platform has been in operation since 2011. It provides easier access to climate simulations for end users, especially the climate change impact modelling community.

The characteristics of the C4I Platform are as follows:

- Targeted at climate science researchers and the climate change impact community, to explore climate data, download required data, and perform on-demand analysis.
- Connected to ESGF web services.
- Visualization capability via ADAGUC Software [28].
- Provides Analysis Services using (Py)WPS [29] to perform calculations.
- Runs an operational service.
- Promotes Users' engagement: climate research community, climate impact community as well as interdisciplinary research community.

The C4I Platform provides a range of processing capabilities, from time and spatial subsetting (with a GUI using OGC WCS Standard [30]) to simple statistics such as time average to more complex calculations such as climate indices and indicators (provided by the ICCLIM software [31]). More complex tools such as statistical downscaling are also provided through an interface to the University of Cantabria Downscaling Portal [32].

2.6.10. ES-DOC

ES-DOC (Earth System Documentation) offers services for metadata search, comparison and creation, following the CIM standard (Common Information Model) [33]. It provides an environment to document the modelling workflow and WGCM [34] and the WIP [35] have tasked ES-DOC with documenting all aspects of CMIP6.

In December 2018, ES-DOC became operational for CMIP6, formally inviting every modelling institute to begin using the ES-DOC infrastructure to document the formulations of their general circulation models. This followed a development phase that, after modification from a select group of beta testers based in modelling centres, resulted in a procedure based on the pyesdoc library where (i) automatically generated spreadsheets, bespoke to each model, are disseminated via GitHub; (ii) modelling centres create documentation content by answering the questions in the spreadsheets and then request that these are published; (iii) as an automated service, ES-DOC converts the spreadsheets into CIM documents, ingests them into the ES-DOC archive, and makes them publicly available via the ES-DOC web-site.

This methodology has been designed to be generalisable to the other types of documentation that need to be collected notably the conformance to experimental requirements, the performance of simulations running on supercomputers, and the descriptions of differences between ensemble members. During 2019, 18 models from 10 institutes were wholly or partially documented, published, and made publically available via the ES-DOC website. The web application that delivers CMIP6 documentation [36] has been refactored to provide improved views of the content, and also be easily configured to display different types of document.

2.6.11. Climate Forecast (CF)

The conventions for CF (Climate and Forecast) metadata are designed to promote the processing and sharing of files created with the NetCDF file format. The CF conventions are increasingly gaining acceptance and have been adopted by a number of projects and groups as a primary standard. The conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities.

The CF Conventions are maintained by a global network of scientists. The ENES CDI supports this work through moderation of the CF Standard Name discussions board and by providing the CF Data Model and associated tools.

The CF Standard Names Tables is a list of agreed terms identifying physical quantities. This addresses a fundamental requirement for exchange of scientific data: namely, the ability to describe precisely the physical quantities being represented.

The CF Data Model sets out a formal semantic structure which identifies the fundamental elements of the CF conventions and shows how they relate to each other, independently of the NetCDF encoding.

2.6.12. Data Request

Data Request

The data request of the Coupled Model Intercomparison Project Phase 6 (CMIP6) defines all the quantities from CMIP6 simulations that should be archived. This includes both quantities of general interest needed from most of the CMIP6-endorsed model intercomparison projects (MIPs) and quantities that are more specialized and only of interest to a single endorsed MIP. The complexity of the data request has increased from the early days of model intercomparisons, as has the data volume. In contrast with CMIP5, CMIP6 requires distinct sets of highly tailored variables to be saved from each of the more than 200 experiments. This places new demands on the data request information base and leads to a new requirement for development of software that facilitates automated interrogation of the request and retrieval of its technical specifications.

The data request provides a database specifying scientific definitions of variables, metadata and formatting requirements, and variable priorities for individual experiments and objectives. It also provides a python library to facilitate exploration of the database and a web interface.

Data Request for Climate Indices

Climate indices play an important role in the practical use of climate and weather data. Their application spans a wide range of topics, from impact assessment in agriculture and urban planning, over indispensable advice in the energy sector, to important evaluation in the climate science community. Several widely used standard sets of indices exist through long-standing efforts of WMO [37] and WCRP Expert Teams (ETCCDI and ET-SCI), as well as European initiatives (ECA&D [38]) and more recently EU Horizon 2020 projects and Copernicus C3S activities. They, however, focus on the data themselves, leaving much of the metadata to the individual user. Moreover, these core sets of indices lack a coherent metadata framework that would allow for the consistent inclusion of new indices that continue to be considered every day.

The Data Request for Climate Indices aims to fill this gap by actively working with the larger community to establish a proposal for a climate index metadata standard that builds on the Climate and Forecasting Conventions. Currently, metadata for a majority of the widely accepted set of core indices are completed. The proposed metadata standard is available in a public repository, and python tools for transforming the spreadsheet information to a format suitable for inclusion in software are available. Tools are being developed for checking whether existing data files adhere to the proposed standard and signaling any deviations.

2.6.13. Identity Management and Access Entitlement

Identity management and access entitlement encompasses the ability to authenticate and authorise users in order to access secured resources such as datasets or computing services. In a federated system, these processes can involve distributed actors in order to implement what is required. This is inherently stateful with dependent services holding session information indicating an authenticated context and attribute information to determine authorisation.

In a federated system, dependent services (Relying Parties) delegate authentication to Identity Providers (IdPs) associated with a user's home institution. IdPs allow user authentication and secure delivery of user attributes associated with the user. In ESGF, the IdP is implemented with OpenID 2.0 for browser-based authentication and a MyProxy server for requesting limited-lifetime user certificates for non-browser based access. ESGF also supports OAuth 2.0 at some sites in the federation. This will be extended as part of a wider rollout together with OpenID Connect, a widely adopted standard for single sign-on which build upon OAuth 2.0 OpenID 2.0 will be deprecated. The IdP also includes services for User Registration and Attribute Services for distributed access control. The latter is a key feature of ESGF: resources deployed at sites in the federation can be secured with a Policy Enforcement Point (access control filter) which acts as a gatekeeper for access calling out to an Authorisation Service. The Authorisation Service makes access control decisions based on a local authorisation policy. This policy may have rules which require a user to be registered for attributes mediated by another site in the federation. For example, for CMIP5 data, ESGF partner PCMDI was the authority for overseeing data access. Users register with PCMDI for access rights and Authorisation Services deployed at sites hosting CMIP5 data query PCMDI's attribute service in order to check users' entitlement to CMIP5 data i.e. whether the given user has the required attributes indicating they are registered for access. This allows a virtual organisation model for ESGF whereby different projects can enforce an access policy spanning multiple nodes hosting secured resources.

2.7. Software components

This section highlights the link between the ENES CDI services and the list of software components associated with them, with a special focus on those developed in the context of the IS-ENES3 project. For each module, a reference to the GitHub repository is also provided.

Table 2. ENES-CDI software components

ENES-CDI Service	**Software components**
ESGF Data	*esg-publisher* doc: https://esgf.github.io/esg-publisher/index.html repo: https://github.com/ESGF/esg-publisher
	esgf prepare repo: https://github.com/ESGF/esgf-prepare
	esgf-pyclient repo: https://github.com/ESGF/esgf-pyclient
	CoG repo: https://github.com/EarthSystemCoG/COG
Citation	doc: http://cmip6cite.wdc-climate.de repo: internal gitlab software versioning at DKRZ
Persistent Identifier (PID)	*ESGF data publication pid library* repo: https://github.com/IS-ENES-Data/esgf-pid
	RabbitMQ federation doc: https://acme-climate.atlassian.net/wiki/spaces/ESGF/pages/107708573/PID+Services+Working+Team+esgf-pidwt (restricted access)
	PID consumer library doc and repo: https://gitlab.dkrz.de/esgf/handlequeueconsumer (restricted access)
IPCC Data Distribution Centre at DKRZ	*IPCC Data Distribution Centre at DKRZ:* http://ipcc.wdc-climate.de DDC web pages on server hosted at CEDA: http://www.ipcc-data.org/sim/
Errata	*ESGF Errata Service*: https://errata.es-doc.org/

	doc: https://es-doc.github.io/esdoc-errata-client/ repos: ● Web-Service: https://github.com/ES-DOC/esdoc-errata-ws ● Front-end: https://github.com/ES-DOC/esdoc-errata-fe ● CLI: https://github.com/ES-DOC/esdoc-errata-client
Data Statistics	*ESGF Dashboard UI:* http://esgf-ui.cmcc.it:8080/esgf-dashboard-ui-2020/ doc: ● https://acme-climate.atlassian.net/wiki/spaces/ESGF/pages/1043464194/Federated+data+usage+statistics+ESGF+Dashboard ● https://acme-climate.atlassian.net/wiki/spaces/ESGF/pages/1054113816/Proposed+ESGF+Usage+of+Filebeat+and+Logstash repo: https://github.com/ESGF/esgf-dashboard
Data Replication	*Synda replication software package* doc: http://prodiguer.github.io/synda/ repo: https://github.com/Prodiguer/synda
Compute	*ECAS* service: https://ecaslab.cmcc.it/jupyter/hub/login doc: https://ecaslab.cmcc.it/web/home.html repo: https://github.com/ECAS-Lab
	Ophidia doc: http://ophidia.cmcc.it/ repo: https://github.com/OphidiaBigData
	Birdhouse WPS framework doc: https://birdhouse.readthedocs.io/en/latest/ repo: https://github.com/bird-house security proxy: https://github.com/bird-house/twitcher
	ESGF-specific WPS framework under development for C3S prototype repos under development at: https://github.com/roocs underlying library: https://github.com/pydata/xarray
	Third party components: - *JupyterHub:* https://jupyter.org/hub - *xarray: http://xarray.pydata.org/en/stable/*
Climate4Impact	*C4I front-end, C4I backend, C4I storybook, C4I errorhandler, C4I front-end content, C4I search portal backend, C4I map preview, C4I frontend dataset preview*

	service: https://climate4impact.eu/impactportal/general/index.jsp repo: https://gitlab.com/is-enes-cdi-c4i
ES-DOC	*ES-DOC service and documentation*: http://es-doc.org CIM repo: https://github.com/ES-DOC/esdoc-cim-v2-schema pyesdoc repo: https://github.com/ES-DOC/esdoc-py-client CMIP6 content repos: https://github.com/ES-DOC-INSTITUTIONAL cdf2cim repo: https://github.com/ES-DOC/esdoc-cdf2cim
Climate Forecast (CF)	service: http://cfconventions.org/ doc: http://cfconventions.org/ repo: https://github.com/cf-convention/
Data Request	*Data Request* service: http://clipc-services.ceda.ac.uk/dreq/ doc: http://w3id.org/cmip6dr repo: https://pypi.org/project/dreqPy/
	Data Request (indices) Proposed metadata specification for climate indices, repository: https://bitbucket.org/cf-index-meta/cf-index-meta/src/master/
Identity Management and Access Entitlement	doc: https://github.com/ESGF/esgf.github.io/wiki/Security%7CInterfaceControlDocument service: Attribute and Authorisation Services repo: http://esgf.org/esgf-security/ service: OpenID Provider and Relying Party repo: http://esgf.org/esg-orp/ service: OAuth 2.0 and Short-lived Credential Service repo: https://github.com/ESGF/esgf-slcs-server

3. TECHNICAL REQUIREMENTS SPECIFICATION

3.1. Methodology

The main purpose of this section is to collect technical requirements for the ENES Climate Data Infrastructure, which will be consolidated to support the exploitation of model data by both the earth system science community and the climate change impact community. To this aim, to better understand the user needs and turn them into technical requirements for the software stack, the document is structured following some of the Agile principles [39], adopting the concepts of *user stories* and *use cases* [40] and translating them into functional and non-functional requirements.

The following definitions explain the two concepts and their features, pointing out the differences between them.

User story

A *user story* - some people call it a *scenario* - expresses a specific need of a user. It is a short description of something that users will do when they use the application/software, focused on the value or result they get from doing such a thing. They are written from the point of view of a person, using the service or application and in the language that the customer would use. A user story is usually written using the format:
"As an [actor] I want [action] so that [achievement]".

Use cases

A *use case* is a more detailed description of a set of interactions between the system and one or more actors, where an actor is either a user or another system. The use case is described as a table that includes:

- the *name* of the use case, assigned for further reference in the document;
- a *summary* or *brief description* of the use case. Unlike the user story, this description should be more system-oriented and explain the main features of the service in a few words;
- the system *actors*;
- the *basic course of events*, the normal flow that the user follows when using the system; in other words, what will usually happen, described as a series of steps;
- any *alternative paths* that users could meet, the exceptions in the system behaviour and what they could do to achieve their goal anyway;
- the *involved components*, that means, in the specific context of the ENES CDI, any other modules of the software stack involved in the system behaviour.

In other words, use cases are more about the behaviour that the technical team will have to build into the software. As mentioned in the bulleted list above, a use case will contain a lot of details, clearly describing everything that a developer needs to build to meet user needs.

Requirements
Starting from the use cases, the extraction of the requirements is a quite simple operation since it’s about translating the system behaviour into technical functional and non-functional requirements.

- *Functional requirement*: it defines a function of a system or its component, where a function is described as a specification of behavior between outputs and inputs [41]. Functional requirements may involve calculations, technical details, data manipulation and processing, and other specific functionality that define what a system is supposed to accomplish.
- *Non-functional requirement*: it imposes constraints on the design or implementation. (e.g. scalability, performance, security).

In section 3.3 of this document, a list of relevant user stories will be described. Each subsection will be then structured as follows:

1. a brief description of one or more user stories;
2. a detailed description of the related use case(s);
3. a comprehensive list of functional requirements (FR), with FR code and description. Requirements will include keywords like "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" according to the definitions provided in RFC 2119 [42];
4. a comprehensive list of non-functional requirements codes (with “(O)” in case of optional, (R) in case of RECOMMENDED). A complete definition/description is given in the following section (Section 3.2).

3.2. Non-functional requirements definition

This section provides a comprehensive list of non-functional requirements with code, short name and description. Their code will be used in the tables provided in Section 3.3.

[NFR#1] Transparency
A software is transparent if it hides the back-end complexity as well as low level technical details. Transparency is usually considered to be a good characteristic of a system because it shields the user from the system's complexity.

[NFR#2] Robustness
A software is robust if any exception raised during its execution, in any architecture and with any initial state, is caught by some exception handler. In other words, it means that the software is stable with regard to operation under stress or toleration of unpredictable or invalid input.

[NFR#3] Scalability

Scalability is an attribute that describes the ability of a software to manage a growing amount of work by adding resources to the system. A system that is described as scalable has an advantage because it is more adaptable to the changing needs or demands of its users or clients.

[NFR#4] Efficiency

The source code and software architecture attributes are the elements that ensure high performance once the application is in runtime mode. Efficiency is especially important for applications in high execution speed environments such as algorithmic or transactional processing where performance and scalability are paramount. An analysis of source code efficiency and scalability provides a clear picture of the latent business risks and the harm they can cause to customer satisfaction due to response-time degradation.

[NFR#5] Security

Security protects the system against malicious attacks and other hacker risks so that the software continues to function correctly under such potential circumstances. It is a necessary aspect to provide integrity, confidentiality and availability.

[NFR#6] Reusability

Software reusability refers to the design features of a software element (or collection of software elements) that enhance its suitability for reuse. The ability to reuse relies in an essential way on the ability to build larger things from smaller parts, and being able to identify commonalities among those parts.

[NFR#7] Extensibility

Extensibility is a software design principle defined as a system's ability to have new functionalities extended while the system's internal structure and data flow are minimally or not affected.

[NFR#8] Flexibility

A system is flexible if it is able to adapt to possible or future changes in its requirements.

[NFR#9] Usability

Usability is the degree to which a software can be used by specified consumers to achieve quantified objectives with effectiveness, efficiency, and satisfaction in a quantified context of use.

[NFR#10] Look and feel

The "look and feel" term can refer to a website, to aspects of a non-graphical UI (such as a command-line interface), as well as to aspects of an API. It describes its appearance and functionality. Particular attention to this aspect should be paid to address a very large adoption of all the provided functionality.

[NFR#11] Interoperability: Interoperability is the characteristic of a system, whose interfaces can be integrated easily to work with other systems through a wide adoption of open standard solutions.

[NFR#12] Access control: The application of access control should be policy-based deriving from the needs of a given project hosted within the federation. Access control should support all the required client interfaces (e.g. web-based, scripted, CLI, APIs). Access control should support delegation of user access rights to authorised third party applications and services in the federation.

[NFR#13] Deployability: In recent years, the time it takes for code to get into production after a commit by a developer has come under scrutiny. A movement known as DevOps has advocated a number of practices and technologies intended to reduce this time. If we call the time to get code into production after a commit the deployment time, we have defined a new quality attribute for complex systems—deployability [43].

3.3. User stories, use cases and technical requirements

The following subsections provide a set of user stories, use cases and technical requirements for the ENES CDI components listed in Section 2.6.

3.3.1. ESGF Data

Table 3. ESGF Data service specification

User story
US#1a: As a data provider, I want to distribute my data so that the files can be easily found from scientists all over the world.
US#1b: As a scientist, I want to search and download data using a “one stop shop” so that I do not have to access different sites

Use case specification	
Code: Name	**UC#1a**: Data publication

Associated User Story	**US#1a**
Summary/Description	Publish data to the ESGF
Actors	Data providers Data publishers
Basic Course of Events	1. The data provider submits the data to the selected data distribution centre. 2. The data centre makes sure the (meta-)data fulfills the ESGF publication requirements. 3. The data publisher publishes the data to the ESGF.
Alternative Paths	1. The data provider installs a ESGF node on its own, linked to one of the existing ESGF index nodes. 2. The data provider makes sure the data fulfills the ESGF publication requirements. 3. The data provider also acts as data publisher and publishes the data to the ESGF.
Involved components	For some projects: PID and Citation Service

Use case specification	
Code: Name	**UC#1b**: Data download
Associated User Story	**US#1b**
Summary/Description	Download data from ESGF
Actors	Data providers ● Specialist data producers (e.g. downstream providers, datacube/cache providers) Scientific end users
Basic Course of Events	1. The scientist searches for specific climate data in the ESGF (either via the SearchAPI or the GUI). 2. The scientist can (prior authentication and authorization steps

	which could optionally be required according to project constraints and data policies) either download single files or the whole data for complete selection, using one of the following download methods: a. HTTP download b. wget-script download c. Globus/GridFTP
Alternative Paths	None
Involved components	Depending on the search or download method: Synda, CoG, Globus

Functional Requirements [DATAFR]	
FR code	*FR description*
[DATAFR#1] Data publication	The data *must* be published to the ESGF. During the publication step, relevant metadata *must* be extracted and published to an index to support a facet based search according to a specific DRS.
[DATAFR#2] Data search	The data *must* be searchable. The search service *must* be user friendly and it *must* provide a faceted search, along with a web-based UI and an automated (machine-readable) search.
[DATAFR#3] Data download	The data *must* be downloadable and the system *must* provide different download methods to fit the needs of all types of users (i.e. big, medium and small data users).
Non-Functional Requirements	
NFR#1, NFR#2, NFR#3, NFR#4, NFR#5, NFR#6, NFR#7, NFR#8, NFR#9, NFR#10, NFR#11	

3.3.2. Citation

Table 4. Citation service specification

User story
US#2a: As a data provider, I want to receive a citation for the data published in ESGF so

that I can share this information in my ORCID [44] researcher profile and have an idea about the data impact.
US#2b: As an article writer, I want to cite some data in a publication so that article readers (and reviewers) could easily verify the consistency of the results presented.

Use cases specification	
Code: Name	**UC#2a**: Citation insertion
Associated User Story	**US#2a**
Summary/Description	New citation insertion and update of the associated metadata.
Actors	Data providers Support
Basic Course of Events	1. The user logs in into the citation service. 2. The user inserts the citation information through the GUI and/or the API. 3. The automated citation service checks for completed citations regarding available data and changes in provided metadata. 4. New or changed information is registered at DataCite: ○ a new DOI and the related metadata for new data, ○ a metadata update for changed information about existing data. 5. The user allows DataCite via the ‘search and link’ functionality to add data references to the ORCID profile.
Alternative Paths	None
Involved components	CMIP6 CV, GUI, API for modeling centres, DOI registration service, Metadata change service, DataCite services, ORCID services

Use cases specification	
Code: Name	**UC#2b**: Get citation information
Associated User Story	**US#2b**

Summary/Description	Get citation information about ESGF data
Actors	Article writer, reviewer
Basic Course of Events	1. A user gets some data for its publications through the ESGF CoG web interface. 2. The user follows the "show citation" link in the ESGF CoG. 3. The user gets the desired information from the citation landing page.
Alternative Paths	Step 2 has the following alternative paths (at the same level): ● use DataCite search: http://search.datacite.org ● use Google Dataset Search: https://datasetsearch.research.google.com ● access furtherInfoUrl link in NetCDF data header in 1.
Involved components	API for ESGF, ESGF CoG, DataCite. For alternative paths: ES-DOC, schema.org implementation on landing pages.

Functional Requirements [CITFR]	
FR code	*FR description*
[CITFR#1] CMIP6 compliance	The system *must* maintain citation information compliant with the requirements of the CMIP6 guide.
[CITFR#2] Insertion or update of citation information	The system *must* allow the data providers to insert new citation information for new data or to update citation metadata for existing data.
[CITFR#3] Access to citation information	The citation information *should* be easily accessible for the users who want to cite data.
Non-Functional Requirements	
NFR#1 (O), NFR#2, NFR#5, NFR#7 (O), NFR#9, NFR#10 (O), NFR#12	

3.3.3. Persistent Identifier (PID)

Table 5. Persistent Identifier specification

User story
US#3: As a scientific user working with larger sets of CMIP6 data, I want to get stable references to the data I downloaded and used, regardless of the preservation and quality status, so that I will be able to recall a summary list of data at a later stage and learn about possible new versions.

Use case specification	
Code: Name	**UC#3**: Data cart creation and query
Associated User Story	**US#3**
Summary/Description	For a custom CMIP6 data download cart, a user (data downloader) requests an individual reference to the cart ("collection") that remains stable even if data is redacted from ESGF or revised. The user will receive a persistent identifier (PID) for this collection as a stable reference. The cart may contain any combination of individual CMIP6 files and datasets. The user or any third party that the user might share the PID with, will be able to recall, at any later point in time, a summary of the data in the collection, including checksums of files, new versions, dataset relations, and, if still available, download links to originals or replicas. The PID and summary will remain available regardless of the status of the data.
Actors	Scientific end users
Basic Course of Events	1. The user gathers data of interest in a download cart. 2. The user requests an individual PID for the collection of data. The PID is given in its original Handle-based format and as a resolvable https URL. 3. At a later point, the user resolves the PID HTTPs URL through the browser and receives a summary page (dynamic web page). 4. The user uses the information on this page and on the linked follow-on pages, including links to individual elements in the collection and download links.
Alternative Paths	None
Involved components	ESGF PID services (data cart service, PID registration, PID viewer)

Functional Requirements [PIDFR]	
FR code	*FR description*
[PIDFR#1] PIDs assigned for CMIP6 files and datasets	CMIP6 files and datasets *must* bear Handle-type PIDs.
[PIDFR#2] Essential metadata registered for CMIP6 file and dataset PIDs	Handle-type PIDs for CMIP6 files and datasets *must* be registered with essential metadata for related items, originals/replicas locations, data state, technical/system metadata.
Non-Functional Requirements	
NFR#1 (O), NFR#2, NFR#3, NFR#5, NFR#6 (O), NFR#7 (O), NFR#8 (O), NFR#11	

3.3.4. IPCC Data Distribution Centre at DKRZ

Table 6. IPCC Data Distribution Centre services specification

User story
US#4: As a reader of the IPCC AR6, I want to access the underlying data so that I can use it for my research activity.

Use case specification	
Code: Name	**UC#4**: Data access
Associated User Story	**US#4**
Summary/Description	Some users of IPCC AR6 data have limited bandwidth for data download of the high-volume data of IPCC AR5 and the forthcoming IPCC AR6 data. An improved service to discover and select the desired datasets and an improved download service to reduce the volume of the transferred data is needed.

Actors	Support Scientific end users
Basic Course of Events	1. The DDC data user discovers data on the DDC web page 2. The DDC data user follows the DOI link to the data landing page hosted at DKRZ 3. The DDC user filters the data results down to the datasets of interest and downloads the selected datasets or dataset subsets.
Alternative Paths	● Data discovery via the WDCC portal instead of steps 1 and 2. ● The DDC user with a weak internet connection asks the DDC service to mail a subset of datasets for a selected domain to him/her.
Involved components	DDC web page, WDCC portal and download services

Functional Requirements [DDCFR]	
FR code	*FR description*
[DDCFR#1] Discovery	The system *must* be improved in terms of data discovery and selection possibilities in the portal.
[DDCFR#2] Download	The download service functionality *must* be improved, especially in terms of a flexible and effective reduction of the total data volume to be transferred.
Non-Functional Requirements	
NFR#1 (O), NFR#2, NFR#3, NFR#5, NFR#8 (O), NFR#9, NFR#10 (O), NFR#12	

3.3.5. Errata

Table 7. Errata service specification

User story
US#5a: As an end-user, I want to get annotations related to any known issues on the ESGF data I use so that I can get timely information about known issues.
US#5b: As a data manager/provider, I want to report/register an issue with some datasets

on behalf of my institute, so that I can provide end-users with timely information about known issues.

Use case specification	
Code: Name	**UC#5a**: Errata lookup
Associated User Story	**US#5a**
Summary/Description	Get the version history with the corresponding issues of one or several ESGF datasets [45].
Actors	Scientific end users ● Analysis users/service
Basic Course of Events	1. The user gets access to the Errata Service UI website. 2. The user clicks on the "Search" button on the top-right menu. 3. The user extracts the PID/Handle from the NetCDF files using basic NetCDF operators (NCO/CDO) [46]. 4. The user submits the PID/Handle through the web form (using the PID field or sending a list of IDs). 5. The user visualises the versioning history for each queried dataset with useful flags (e.g., initial version, latest version, queried version) 6. The user gets issue information for each version through the appropriate link. 7. The Errata Service redirects the user to the corresponding issue viewer that provides: a. The title of the issue b. The full description of the issue c. External links to related materials (graphics, additional landing pages) d. The status and severity of the issue. e. The list of affected ESGF datasets.
Alternative Paths	None
Involved components	PID Service

Use case specification

Code: Name	**UC#5b**: Issue life-cycle
Associated User Story	**US#5b**
Summary/Description	Manage an issue on behalf of a data provider affecting one or several ESGF datasets [47-49].
Actors	Support ● Infrastructure-level ● Application-level ● User support Data providers ● Modelling groups providing simulations output ● Specialist data producers Data publishers ● IT experts publishing data into ESGF
Basic Course of Events	1. The user gets access to the Errata Service UI website. 2. The user clicks on the “Log In” button on the top-right menu. 3. The user interacts with the GitHub OAuth box to grant access to the GitHub organisation. 4. The user asks the ES-DOC administrator to be part of the appropriate GitHub organisation in relation to the institute. 5. The user clicks on the “Create” button on the top-right menu. 6. The user fills the web form with the following information: a. Project (select in drop-down menu) b. Title of the issue (must be unique) c. Description of the issue d. Severity (select in drop-down menu) e. Optional external links for materials f. List of affected dataset ID in the appropriate format. 7. The user clicks on the “Save” button on the top-right menu. 8. The system validates the issue format against controlled vocabularies and syntaxes. 9. The Errata Service redirects the user to the new issue viewer: a. The system adds a unique identifier to the newly created issue. b. The system initializes the issue status to “new”. 10. The user updates the issue information at any time by clicking on the “Edit” button on the top-right menu within the Issue Viewer. Only the title and the project are unchanged.

	11. The user can close an issue by updating its status to “wontfix” or “resolved”.
Alternative Paths	None
Involved components	PID Service

Functional Requirements [ERRFR]	
FR code	*FR description*
[ERRFR#1] Data publication	Data with known issues *must* be published on the ESGF. PID registration *must* be enabled during the publication process.
[ERRFR#2] Issue viewer	The system *must* provide timely information about the issue: - unique identifier - unique title - description - status (new, on hold, wontfix, resolved) - severity (low, medium, high, critical) - optional materials urls - list of affected dataset ID - creation/update dates - author username - optional links with the ES-DOC vocabularies
[ERRFR#3] Issue list	The system *must* provide the full list of known issues. The list *must* be filtered by usual facets (e.g. project, model, variable, etc.)
[ERRFR#4] Issue management	The system *must* provide user-friendly ways to manage the issues: 1. A web form to be filled and automatically validated against controlled vocabularies and syntaxes depending on the project. 2. A Unix command-line interface that makes it easier to manage several issues. The system *must* provide the creation/registration, update and closing of an issue. The system *must* deal with some authorization layer to grant issue edition only to allowed users. The system *must* provide the retrieval/download of any issue for all users in JSON format with the list of affected datasets in TXT format.

Non-Functional Requirements
NFR#1, NFR#2, NFR#3 (O), NFR#4, NFR#5, NFR#6, NFR#7 (O), NFR#8, NFR#9, NFR#10 (O)

3.3.6. Data Statistics

Table 8. Data statistics service specification

User story
US#6: As an end user, I want to get a clear view of the data usage related to the climate datasets available via ESGF, in order to capture the data downloads and publication activity as well as the level of interest from the climate community in specific climate datasets, projects, variables, etc.

Use case specification	
Code: Name	**UC#6**: Usage and publication statistics
Associated User Story	**US#6**
Summary/Description	Visualise statistics about the data usage and the published datasets at the single site, European and global ESGF level. Specific interest is in EU data nodes and EU clients.
Actors	Community/Services Governance, project coordinators and funding authorities ● Funding agencies ● ENES board representatives, ENES Task Forces and IS-ENES PO ● ESGF Executive committee ● Governance bodies ● Project coordinators Scientific end users
Basic Course of Events	1. The user gets access to the Data statistics UI website. 2. The user visualises information about the total volume of published datasets as well as data usage statistics both at ESGF, European (including IS-ENES KPIs) and single-site level.

	More in particular, the user can display one or more of the following statistics views: a. Data publication statistics, split also by project (CMIP5, CMIP6, Input4MIPs, Obs4MIPs, CORDEX), giving an overall idea of how many datasets have been published for each project. b. The trend over time of data publication, as a general overview and for the CMIP6 and CORDEX projects (the same information for each data node of the Federation). c. For CMIP6, CMIP5 and CORDEX, specific details such as the volume and the number of published datasets per project-specific facets. d. Cross-project data usage statistics such as the number of downloads, the downloaded data volume and the number of replica downloads. e. Cross-project statistics over time, per host and per project, as a general overview of the federation or for a specifically selected data node. f. Project-specific data usage statistics such as the number of downloads, the downloaded data volume and the number of replica downloads, aggregated per project-specific facets or filtered by a selected data node. g. Project-specific data usage statistics about the most (top 10-20) downloaded datasets, experiments and variables. h. IS-ENES3 KPIs (i.e. successfully downloaded files and the data volume for European and non-European users to have detailed information about the usage statistics of data available through the IS-ENES3 data nodes). 3. The user exports the statistics in a CSV format.
Alternative Paths	None
Involved components	ESGF Data services (data nodes)

Functional Requirements [STATSFR]	
FR code	*FR description*

[STATSFR#1] Data publication statistics	Data publication statistics at project and ESGF level *must* be provided. The system *must* distinguish between distinct and replica datasets and provide the same information for the whole federation and for each data node. The system *must* also offer a view about the trend over time of data publications, as a general overview and for specific projects for at least CMIP6 and CORDEX. It *should* provide similar statistics for the other projects too. For the most important projects, like CMIP6, CMIP5 and CORDEX, the service *must* provide more specific details such as the volume and the number of published datasets per project-specific facets.
[STATSFR#2] Cross-project data download statistics	The service *must* provide **cross-project** data usage statistics such as the number of downloads, the downloaded data volume and the number of replica downloads. The statistics *must* be visualised over time, by host and by project, as a general overview of the federation or for a specifically selected data node.
[STATSFR#3] Project-specific data download statistics	The service must provide **project-specific** data usage statistics such as the number of downloads, the downloaded data volume and the number of replica downloads. The statistics must be reported as a general overview of the federation or for a specific selected data node; additionally, the statistics must be aggregated per project-specific facets. The service must also provide statistics about the most (top 10-20) downloaded data from different perspectives according to the most popular project-specific facets.
[STATSFR#4] Data statistics export	The user *must* be able to export the statistics at least in CSV format.
[STATSFR#5] IS-ENES3 KPIs	The Data Statistics service must provide detailed information about the data usage statistics related to the IS-ENES3 data nodes. For this reason, the system must visualise the number of successfully downloaded files by European and non-European users, and the related data volume.
Non-Functional Requirements	
NFR#1, NFR#2, NFR#3, NFR#4 (O), NFR#5, NFR#7, NFR#8, NFR#9, NFR#10 (O), NFR#13	

3.3.7. Data Replication

Table 9. Data replication service specification

User story
US#7: As a data manager, I want to replicate data collections from remote ESGF sites so that they are locally available and can be published as replicas to the ESGF federation.

Use case specification	
Code: Name	**UC#7**: replicate ESGF data collections
Associated User Story	**US#7**
Summary/Description	Make data replicas available at multiple ESGF sites to improve data access reliability and speed.
Actors	Support Data publishers
Basic Course of Events	1. Specify the replica data collections based on Synda selection files. 2. Start Synda using these selection files. 3. Monitor downloads e.g. for continuously unsuccessful transfer and update selection files. 4. Regularly retry selection files to update the latest versions of files.
Alternative Paths	For retracted files from remote centres also retract local copies.
Involved components	ESGF data node, Synda, GridFTP

Functional Requirements [REPLICFR]	
FR code	*FR description*
[REPLICFR#1] Persistent status	The replication service *must* continuously track the ongoing data transfers with respect to their status and persistently store this

tracking and recording	information to support restart etc.
[REPLICFR#2] Multiple transfer protocols and parallel transfers	The replication service *must* support multiple selectable transfer options, with the minimal requirement to support HTTP and Globus (GridFTP). The replication service *must* support configurable parallel data transfers.
[REPLICFR#3] Faceted data characterization	The replication service *must* support faceted specifications to characterize the data collection(s) which have to be replicated.
[REPLICFR#4] Automatic retry	The replication service *must* be able to run as a continuous service and must include support for automatic retry of failing data transfers.
Non-Functional Requirements	
NFR#2, NFR#3, NFR#9	

3.3.8. Compute & Analytics

Table 10. Compute service specification

User story
US#8: As a researcher interested in climate science I want to run data analysis on CMIP6 data from ESGF so that I can extract knowledge from the available datasets and derive new data products (i.e. indicators).

Use case specification	
Code: Name	**UC#8**: Compute services
Associated User Story	**US#8**
Summary/Descrip	Exploit a compute service to perform interactive or batch server-side

tion	analysis, from simple operations to more complex computation (i.e. climate indices)[1].
Actors	Scientific end users ● Analysis users/service ○ big data users ○ medium data users ○ applications/services
Basic Course of Events	End-user-to-application (interactive & batch mode): 1. The user establishes a secure session with the compute service. 2. The user defines a set of input data from what is available in the pool connected to the compute facility or gathers it from ESGF data nodes. 3. The user performs interactive/batch data analysis (i.e. from simple operations like subsetting, statistical, arithmetic, interpolation, etc. to more complex workflows/analyses). 4. The user downloads or stores the main outcomes for further analysis and sharing of the results. Application-to-application (batch mode): 1. An application establishes a secure session with the compute service. 2. It performs batch data analysis (i.e. from simple operations like subsetting, statistical, arithmetic, interpolation, etc. to more complex workflows analyses). 3. It transfers/stores the output for further analysis and sharing of the results.
Alternative Paths	0. The user authenticates to a data science interactive environment with a rich inventory of ad-hoc libraries for data analysis and visualization/reporting.
Involved components	ESGF Data services (data nodes) Identity management and access entitlement service

Functional Requirements [COMPFR]	
FR code	*FR description*

[1] Security aspects are only partially considered in the UC#8 as the main focus here is on the computing aspects. UC#13 provides more insights into identity management and access control (see section 3.3.13 on Identity Management and access entitlement).

[COMPFR#1] Server-side data analysis	Compute services *must* provide server-side data analysis (with compute and data co-location) since data download is no longer a viable option for users and the critical volume of the analysis cannot be properly handled with the available client-side data management tools.
[COMPFR#2] Subsetting	Compute services *must* support spatio-temporal data subsetting
[COMPFR#3] Simple operations	Compute services *must* be able to perform a range of basic operations like statistical, arithmetic, interpolation and transformation.
[COMPFR#4] Complex analysis	Compute services *should* support workflow-enabled analytics to manage more complex analyses (i.e. multi-model climate analysis and climate indicators).
[COMPFR#5] Access control	Compute services *must* adopt an appropriate solution for access control to ensure legitimate use of the computing resources by external users.
[COMPFR#6] API	Compute services *should* provide an API to build applications on top of them and grant programmatic access to the compute resources for the development of "data science" applications.
[COMPFR#7] Standard interface	A standard interface *should* be provided (such as the WPS specification) to ensure ease of access, wide adoption and interoperability with the most relevant access points of the community (e.g. Climate4Impact portal).
[COMPFR#8] High-level user environment for data analysis	Besides the Command Line Interface (CLI), the compute services *should* be accessible through a user interface equipped with tools for interactive data analysis/manipulation and visualization.
[COMPFR#9] Compute Service discovery	Compute Services (and algorithms) *may* be discovered using a search catalog.
Non-Functional Requirements	
NFR#2, NFR#3, NFR#4 (O), NFR#5, NFR#7, NFR#9, NFR#11, NFR#13	

3.3.9. Climate4Impact

Table 11. Climate4Impact specification

User story
US#9: As a climate data user, I want to be able to easily access climate data with proper guidance, so that I can choose and use properly the climate scenarios I need to perform my research and study.

Use case specification	
Code: Name	**UC#9**: Provide an intuitive UI with guidance to accessing climate data for climate change impact community users.
Associated User Story	**US#9**
Summary/Description	Web-based platform offering a UI and standard services to climate change impact community users, enabling access to climate data using ESGF data and services.
Actors	Scientific end users ● Analysis users/service ○ medium data users ○ small data users
Basic Course of Events	End-user-to-application (interactive & batch mode): 1. The user establishes a secure session using the C4I UI. 2. The user uses a wizard to search and select the climate scenarios to analyze. Guidance is provided by the UI. 3. The user saves this selection in the C4I User Space. 4. The user performs interactive/batch data analysis (i.e. from simple operations like subsetting, statistical, arithmetic, interpolation, etc. to more complex workflows analyses). Provenance and Lineage information will be generated. 5. C4I will output and repackage data according to selected climate scenarios and, optionally, to the results of the data analysis. 6. The user generates plots of climate scenario statistics. 7. The user downloads or stores the main outcomes for further analysis and sharing of the results.

Alternative Paths	● The user can choose to download just the data selection and not to perform any analysis/data processing using C4I. ● The user could need extra information on the climate scenarios and query information from ES-DOC.
Involved components	External Birdhouse WPS ESGF Data Nodes ES-DOC Service for Access Management (Keycloak [50]) ICCLIM open-source software for climate indices and indicators calculations

Functional Requirements [C4IFR]	
FR code	*FR description*
[C4IFR#1] Search Wizard Interface	Guided Search UI that uses the ESGF Search Service.
[C4IFR#2] Visualization Service	A Service and UI that enables users to generate plots from data files stored in their C4I user space.
[C4IFR#3] Documentation and Guidance	On-demand general documentation and guidance on climate scenario simulations and data made accessible by C4I.
[C4IFR#4] User Space Storage (Basket)	User Space Storage where users can store data analysis results as well as links to remote data files.
[C4IFR#5] Data Subsetting Service	On-demand data processing: subsetting (spatial, temporal), simple statistics, arithmetic, interpolations.
[C4IFR#6] Download Service	Provide download service for files stored in C4I users' space.
[C4IFR#7] Standard Climate Indices and Indicators	On-demand calculations of standard international community-defined climate indices and indicators performed on selected climate scenarios.

Calculation Service	
[C4IFR#8] Data repackaging Service	Automated service to repackage automatically users' selected climate scenarios given subsetting information (spatial and temporal). Outputs are stored in C4I users' space.
[C4IFR#9] Provenance/Lineage Service	Automated service that records provenance and lineage information when data processing and calculations are performed using C4I services.
Non-Functional Requirements	
NFR#1, NFR#2 (O), NFR#3, NFR#4 (O), NFR#5, NFR#6, NFR#9, NFR#10, NFR#13 (O)	

3.3.10. ES-DOC

Table 12. ES-DOC specification

User stories
US#10a: As an analysis user, I want to get documentation relating to all aspects of CMIP6 (or other projects), so that I can thoroughly understand the workflow, including experiment design, model formulation, simulations, model performance and hardware.
US#10b: As a data provider of documentation, I want to easily access and use the infrastructure so that I can enter and publish the documentation about my institute's activities.

Use case specification	
Code: Name	**UC#10a**: Documentation access via "further info" URL
Associated User Story	**US#10a**
Summary/Description	Get documentation on any aspect of the CMIP6 project.
Actors	Scientific end users ● Analysis users/service

Basic Course of Events	1. The user gets a CMIP6 dataset from ESGF. 2. The user navigates to the “further info” URL contained in the NetCDF file. 3. The user uses the links on the further info URL landing page to navigate to the documentation about the complete workflow that produced the data in the original file.
Alternative Paths	A lesser coverage of documentation could be findable, with much more effort on the user’s part, via peer reviewed journals and personal communication.
Involved components	ES-DOC archive and web services.

Use case specification	
Code: Name	**UC#10b**: Documentation access via API
Associated User Story	**US#10a**
Summary/Description	Get documentation via the pyesdoc or pyosl APIs
Actors	Data providers ● Modelling groups providing simulations output ● Sensors providing observations ● Specialist data producers Scientific end users ● Analysis users/service ○ small data users
Basic Course of Events	1. The user configures the Python pyesdoc or pyosl package to view the ES-DOC (or local) archive. 2. The user creates, reads, writes CIM documents using either library.
Alternative Paths	None
Involved components	The pyesdoc and pyosl packages, the CIM data model

Use case specification	
Code: Name	**UC#10c**: Provision of materials for collecting documentation
Associated User Story	**US#10b**
Summary/Description	CMIP6 groups require user-friendly means to provide their workflow documentation.
Actors	Data providers ● Modelling groups providing simulations output ● Sensors providing observations ● Specialist data producers
Basic Course of Events	1. ES-DOC admins provide pre-formatted spreadsheets via GitHub to the CMIP groups. These reflect their exact CMIP6 activities (i.e. which models they are using, which experiments they are running, etc.). 2. The user in a CMIP6 institute fills in the spreadsheets and returns them, via GitHub, to ES-DOC. 3. Automated ES-DOC services convert new content to archived CIM documents.
Alternative Paths	MOHC is using pyesdoc to create some (but not all) documentation, facilitated by their use of an internal database. This path is generally beyond the resources of most CMIP6 modelling groups.
Involved components	ES-DOC web server (ENES-CDI) and GitHub (external service)

Use case specification	
Code: Name	**UC#10d**: Provision of web based search and compare of documentation
Associated User Story	**US#10a**
Summary/Description	User-friendly and easy to maintain web-based access to workflow documentation
Actors	Data providers

	● Modelling groups providing simulations output ● Sensors providing observations ● Specialist data producers Scientific end users ● Analysis users
Basic Course of Events	1. The user visits the ES-DOC website[36]. 2. The user goes to the “explorer” link, which provides a viewer and download facility for archived documentation. 3. The user goes to the “compare” page to select attributes of multiple documents to be viewed side-by-side.
Alternative Paths	None
Involved components	ES-DOC web server

Functional Requirements [ESDOCFR]	
FR code	*FR description*
[ESDOCFR#1] Documentation publication	Documentation about the CMIP6 workflow *must* be published to the ES-DOC archive of CIM documents.
[ESDOCFR#2] Web-based documentation access	Documentation *must* be accessible to users via services available from [36]
[ESDOCFR#3] API-based documentation creation and access	Documentation *must* be creatable and accessible via a software API.
[ESDOCFR#4] Further Info URL	The further Info URL in every CMIP6 dataset *must* resolve to a landing page that collates documentation about the workflow that led to the creation of the dataset.
Non-Functional Requirements	
NFR#1, NFR#2, NFR#3, NFR#4, NFR#5, NFR#6, NFR#7, NFR#8, NFR#9, NFR#10,	

NFR#11 (O)

3.3.11. Climate Forecast (CF)

Table 13. CF service specification

User story
US#11a: As a data provider, I want to describe my data concisely and clearly, by using standard labels to identify variables, so that users who work with my data products can easily and accurately identify the physical quantities being represented.
US#11b: As a data provider, I want to encode the technical metadata associated with my data in a structured and logical way so that users are able to automate data analysis and reuse existing tools.

Use case specification	
Code: Name	**UC#11a**: Standard names
Associated User Story	**US#11a**
Summary/Description	A data provider wants to describe their data concisely and clearly, by using standard labels to identify variables, so that users who work with the data products can easily and accurately identify the physical quantities being represented. The CF Convention provides a mechanism for doing this through CF Standard Names.
Actors	Data Provider ● Specialist data producers (scientist bringing new data to publication, scientists supporting maintenance of community standards) Service providers ● Metadata service providers (moderator of the standard name discussions, maintenance of services supporting discussion and publication)
Basic Course of Events	1. A data provider proposes a new standard name for a variable. 2. The community discusses the new standard name. 3. The new term and its description are published in the Standard

	Name Table. 4. The new term is published in the NERC Vocabulary Service.
Alternative Paths	Long lists of closely related terms can be submitted as spreadsheets.
Involved components	GitHub repository issues page [external to ENES CDI]; CF standard name editor; CF Conventions website [external to ENES CDI] NERC Vocabulary Service [external to ENES CDI]

Use case specification	
Code: Name	**UC#11b:** CF Data Model
Associated User Story	**US#11b**
Summary/Description	As a data provider, I want to encode the technical metadata associated with my data in a structured and logical way so that users are able to automate data analysis and reuse existing tools.
Actors	Data Provider ● Specialist data producers (scientist bringing new data to publication) Service Provider ● Metadata service providers (developer of the CF-python library, informatics experts supporting the CF Convention).
Basic Course of Events	1. Data provider proposes a new semantic structure. 2. Review by CF Community 3. Implementation by development team
Alternative Paths	None
Involved components	CF Python code repository

Functional Requirements [CFFR]	
FR code	*FR description*
[CFFR#1]	Community discussion *must* be supported in order to reach agreements

Discussion board	on community standards.
[CFFR#2] Publication workflow	A mechanism to manage and publish complex technical documents *must* be supported.
Non-Functional Requirements	
NFR#1, NFR#2, NFR#10, NFR#11	

3.3.12. Data Request

Table 14. Data request service specification

User story
US#12a: As a specialist data producer or analyst user, I want to produce climate indices enforcing specific metadata conventions (i.e. Data Request for climate indices) so that they are semantically consistent with the proposed guidelines defined at the community level[2].
US#12b: As an analysis user, I want to check the metadata stored in already produced climate indices datasets against the Data Request for climate indices database, so that I can check whether they are consistent with the proposed guidelines defined at the community level.
US#12c: As a Model Intercomparison Project Coordinator, I wish to specify a list of climate variables that should be provided by modeling centres contributing to my project, and also determine the file format and metadata syntax to facilitate efficient analysis. The list of variables should be prioritised, and there may be different variables required for different experiments.

[2] For example, there are many alternative definitions of a climate index called "growing season length" and it is therefore important to have a metadata standard that is able to describe the differences. In essence, this metadata standard informs how to format the files to ensure the required interoperability in order to prevent the customer from inadvertently comparing or combining incommensurate quantities.

US#12d: As a modeling centre program manager, I need to plan our involvement in a model intercomparison project. Some of the objectives are not relevant to our research team or to our model, so I would like to spare our team unnecessary effort and only provide data related to the objectives we support.
US#12e: As a data delivery technician at a modeling centre, I need to know which variables should be produced from each model simulation, and how to format them. I'm dealing with thousands of variables and hundreds of experiments, so I need information in a form, which supports automation and can be integrated into workflows in our HPC system.
US#12f: As a potential user of CMIP data, I want to explore the scope of data requested and understand the physical meaning of the variables listed.

Use case specification	
Code: Name	**UC#12a**: Enforce semantic consistency of available climate indices
Associated User Story	**US#12a**
Summary/Description	Enforce the semantic consistency of new climate indices data products by exploiting the Data Request for climate indices service.
Actors	Data providers ● Specialist data producers Scientific end users ● Analysis users
Basic Course of Events	1. The user wants to produce a "standard" climate indicator. 2. To enforce the proper metadata, the user queries the Data Request for climate indices database. 3. The user enforces the retrieved metadata into the output file.
Alternative Paths	None
Involved components	Analysis/compute tools

Use case specification

Code: Name	**UC#12b**: Check consistency of available climate indices
Associated User Story	**US#12b**
Summary/Description	Check the semantic consistency of available climate indices data products by relying on the Data Request for climate indices database.
Actors	Scientific end users ● Analysis users
Basic Course of Events	1. The user wants to check the semantic consistency of a specific dataset. 2. the user runs a CF climate index checker which in turn queries the Data Request for climate indices database. 3. The CF climate index checker provides, either: a. a list of metadata which are not compliant with the relevant community standard or b. a success message saying everything is ok.
Alternative Paths	None
Involved components	CF index checker (client application)

Use case specification	
Code: Name	**UC#12c**: MIP support
Associated User Story	**US#12c**
Summary/Description	Provide support for Model Intercomparison Project Coordinators
Actors	Support ● User support Service providers ● Metadata service providers Community/Services Governance, project coordinators and funding authorities ● Funding agencies ● ENES board representatives, ENES Task Forces and IS-ENES

	PO ● ESGF Executive committee ● Governance bodies ● Project coordinators
Basic Course of Events	1. A MIP is designated as part of the CMIP process by the CMIP panel based on a scientific project plan. 2. The scientific project plan contains some technical information, there is confusion about the hand-over from the scientific scoping phase to technical implementation. 3. Lists of desired variables and experiments are generated by the MIP science teams, and a database specifying variable requirements for each experiment is compiled. 4. Lists of experiments are reviewed by teams working on ES-DOC and CVs. 5. Lists of variables are reviewed by the Data Request team and CF Conventions community -- the discussion includes clarifying specifications, identifying duplication with other MIPs, developing standards as needed, scoping data volumes and checking that this is realistic. 6. A database schema to support structured dissemination of the final request is developed and maintained. 7. Software tools which support access to the database are developed and maintained. 8. Versioned releases.
Alternative Paths	None
Involved components	Template for information provided by MIPs Guidance documentation for MIPs submitting information List ingestion software Issue tracker Python library for accessing the compiled database Web site providing navigable view of the database Database as XML file, with XSD schema Documentation about software and schema

Use case specification	
Code: Name	**UC#12d**: ESM project management support
Associated User	**US#12d**

Story	
Summary/Description	Provide support for ESM teams participating in Model Intercomparison Projects
Actors	Support: ● User support Service providers ● Metadata service providers Community/Services Governance, project coordinators and funding authorities ● Governance bodies ● Project coordinators
Basic Course of Events	1. Versions of the Data Request are developed and released (US#12c[3]). 2. Python library and schema enable estimation of data volumes to support project planning. 3. Feedback from modeling centre to MIP teams on priorities and feasibility. 4. Revisions implemented through US#12c.
Alternative Paths	None
Involved components	Python library for accessing the compiled database; Web site providing navigable view of the database; Issue tracker; Database as XML file, with XSD schema; Documentation about software and schema;

Use case specification	
Code: Name	**UC#12e**: Modeling centre technician
Associated User Story	**US#12e**
Summary/Description	Provide support for ESM technician to implement the delivery of requested data
Actors	Support:

[3] Note that some ESM teams participate in delivery of US#12c through active participation in the CF community and the Data Request team.

	● User support Service providers ● Metadata service providers
Basic Course of Events	1. Versions of the Data Request are developed and released (US#12c). 2. Feedback on technical implementation and scientific issues is received. 3. Scientific feedback (e.g. variable specification does not make sense) is fed back to MIP teams, technical feedback feeds into software and schema enhancements. 4. Schema is frozen to enable workflow integration. 5. Content is partially frozen to enable implementation to start.
Alternative Paths	None
Involved components	Python library for accessing the compiled database; Web site providing navigable view of the database; Issue tracker; Database as XML file, with XSD schema; Documentation about software and schema;

Use case specification	
Code: Name	**UC#12f**: CMIP archive user
Associated User Story	**US#12f**
Summary/Description	Provide support for potential users who want to plan use of data from the archive.
Actors	Analysis users/service: ● big data users (need to develop tools and workflows); ● medium data users ● small data users
Basic Course of Events	1. Versions of the Data Request are developed and released (US#12c). 2. Web interface allows users to discover variables through a flexible search page. 3. Variables are linked to related variables and underlying CF standard names.

	4. Python library supports programmatic exploration for advanced users.
Alternative Paths	None
Involved components	Python library for accessing the compiled database; Web site providing navigable view of the database; Issue tracker; Database as XML file, with XSD schema; Documentation about software and schema;

Functional Requirements [DREQFR]	
FR code	*FR description*
[DREQFR#1] Database (DR climate indices)	The service *must* offer a Data Request for climate indices database.
[DREQFR#2] Programmatic access	Programmatic access to the Data Request for climate indices database *must* be provided to support query both from end-users and applications.
[DREQFR#3] Versions	The Data Request for climate indices database *should* manage versions to ensure that consistency is associated with a specific version of the database, over time.
[DREQFR#4] Request ingestion	The CMIP Data Request service *must* provide a transparent mechanism for submission of variable specifications and output requirements.
[DREQFR#5] Database (DR CMIP)	The CMIP Data Request *must* contain all the information in a structured form supporting programmatic access. There *must* be extensive checking of compliance with syntactical specifications.
[DREQFR#6] Web Interface	A navigable web interface *must* be provided to support exploration of the database.
[DREQFR#7] Software	A software library *must* support programmatic access to the CMIP Data Request.
[DREQFR#8] Component versions	All components *must* be version controlled, with a transparent relationship between versions; the web site must support viewing of all long-term support versions (a subset of published versions).

[DREQFR#9] Schema	The CMIP Data Request requires a schema which supports full specification of the variables, their priorities and other metadata needed to support user requirements. Tools are needed to support maintenance of the schema.
Non-Functional Requirements	
NFR#1[4], NFR#2, NFR#7, NFR#9(O), NFR#11(O)	

3.3.13. Identity Management and Access Entitlement

Table 15. Identity management and Access Entitlement service specification

User story
US#13: As a researcher interested in climate science I want to run data analysis on secured datasets from ESGF so that they can extract knowledge from the available datasets and derive new data products (i.e. indicators).

Use case specification	
Code: Name	**UC#13**: Identity Management and Access Entitlement with Compute services
Associated User Story	**US#13**
Summary/Description	Exploit a compute service to perform interactive or batch server-side analysis, from simple operations to more complex computation (i.e. climate indices)[5].
Actors	Scientific end users ● Analysis users/service ○ big data users ○ medium data users ○ applications/services

[4] The CMIP data request develops in parallel with other CMIP services and relies on developments in the CF community

[5] This use case extends the one provided in the UC#8 (see section 3.3.8 on Compute & Analytics), which focused on computing aspects only. It provides more insights into identity management and access control.

Basic Course of Events	End-user-to-application (interactive & batch mode): 1. The user establishes a secure session with the compute service. This could be an out of band action where the user authenticates with an IdP and obtains a token, or the compute service client prompts the user to authenticate and obtains the token for them. 2. The user defines a set of input data from what is available in the pool connected to the compute facility or gathers it from ESGF data nodes. 3. The user has selected data which is secured. The compute service must obtain delegated rights from the user to access the secured data on the user's behalf using the user's security credentials. The compute service must indicate to the user if the user's access permissions are not sufficient to access the required data. 4. The user performs interactive/batch data analysis (i.e. from simple operations like subsetting, statistical, arithmetic, interpolation, etc. to more complex workflows analyses). 5. The user downloads or stores the main outcomes for further analysis and sharing of the results. Application-to-application (batch mode): 1. The application establishes a secure session with the compute service using a token previously obtained by the user. 2. The user has selected data which is secured. The compute service must obtain delegated rights from the user to access the secured data on the user's behalf using the user's security credentials. The compute service must indicate to the user if the user's access permissions are not sufficient to access the required data. 3. The compute service performs batch data analysis (i.e. from simple operations like subsetting, statistical, arithmetic, interpolation, etc. to more complex workflows analyses). 4. It transfers/stores the output for further analysis and sharing of the results.
Alternative Paths	0. The user authenticates to a data science interactive environment with a rich inventory of ad-hoc libraries for data analysis and visualization/reporting.
Involved components	ESGF Data services (data nodes) Identity Provider, Authorisation service, Attribute Service, Compute Service Policy Enforcement Point, Data Node Policy Enforcement

	Point

Functional Requirements [IDFR]	
FR code	*FR description*
[IDFR#1] Delegation	The system *shall* support delegation of access rights.
[IDFR#2] Delegation with service chaining	The system *shall* support the ability for multiple services to be chained together whilst at the same time supporting the delegation of the users access rights throughout the chain.
[IDFR#3] Role based access control	The system *shall* support role or attribute-based access control to secure access to data *and* processing services.
[IDFR#4] Tokens	The system *shall* support the use of tokens to represent an authenticated session and / or validated access rights for a given user.
[IDFR#5] IdP trust	It *shall* be possible for the user to authenticate with any of a range of Identity Providers trusted within the federation.
Non-Functional Requirements	
NFR#11[6](R), NFR#12[7](R), NFR#13	

[6] The system *should* seek to use widely adopted standards and off-the-shelf/open source where possible so as to avoid the need to build and maintain its own solutions and in order to maximise the possibility of interoperability with other systems.

[7] The application of access control *should* be policy-based deriving from the needs of a given project hosted within the federation. For example, CMIP6 does not require access control but some resources (e.g. computational) may require registration and restrictions applied to access.

4. PRELIMINARY ARCHITECTURAL INSIGHTS

Even though the ENES CDI software stack architecture will be defined, documented and presented in the deliverable D10.1 at month 18, a preliminary architectural sketch is proposed below as a result of this initial phase on technical requirements elicitation and architectural discussions.

The ENES CDI stack is part of a multi-tier architecture organized according to the following Tiers (*yellow*):

- **Exploitation**: it refers to the stakeholders that include the scientific community and data providers as well as community governance, public/private sectors (either as users or as external applications) and funding agencies.
- **Platform**: it includes the ENES CDI stack, a suite of software components that can be selectively deployed according to specific needs and goals. It consists of the following layers *(purple)*:
    - **Fabric**: it provides basic data, metadata and compute services;
    - **Federation**: it federates multiple services at the Fabric layer, thus providing federation-level capabilities (unified view) as well as a single entry point to the user;
    - **Application**: it provides end-users applications to (i) perform data analysis, (i) get access to documentation, (iii) run/visualize climate indicators, (iv) report data usage/publication metrics, etc.

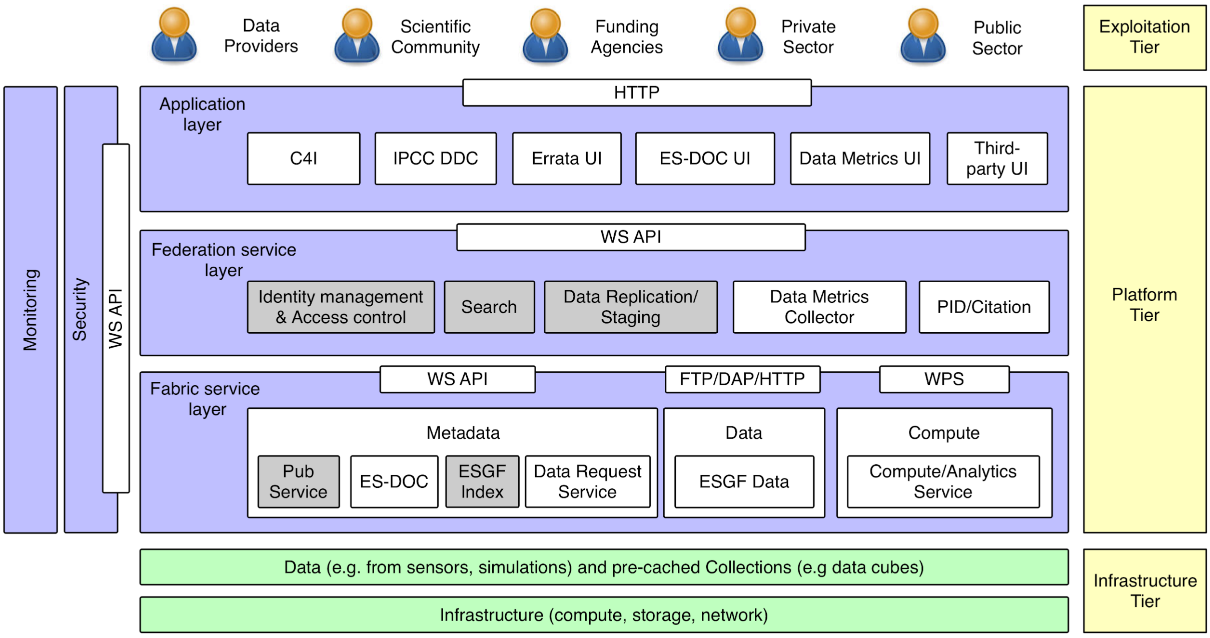

Fig. 1. ENES CDI software stack architecture

    - **Security**: it is an orthogonal layer of the stack that goes across the Fabric, Federation and Application layers. It includes, among the others, firewall settings, OS/applications/services security updates, etc.
    - **Monitoring**: as in the case of Security, it represents a cross-architectural layer that addresses monitoring aspects at different levels (e.g. from infrastructure to services up to applications, with different sets of metrics).

    The picture also highlights the main protocols by which each layer exposes its services; they are very diverse on the *Fabric* layer and more convergent towards Web Service API and HTTP respectively on the *Federation* and *Application* layers.
    The Platform Tier relates to any service of the ENES CDI that *could* be deployed according to a PaaS or SaaS approach in a public or private virtualized environment.
- **Infrastructure** (*green*): it consists of compute, storage and network physical resources and, on top of it, the *Data*, which relate to sensors and to the output of numerical simulations.
  The Infrastructure Tier *could* be virtualized, thus providing access to the resources according to a IaaS approach either in a public or private cloud environment.

Finally, in grey, a set of ESGF services exploited in the ENES-CDI that are associated with collaborative development efforts carried out with partners outside Europe.
Such an architectural stack will be further developed and discussed over the next months, as a starting point towards the final ENES CDI architecture that will be presented in D10.1.

5. CONCLUSIONS

This report addresses the milestone M10.1 “Technical requirements on the software stack” of the IS-ENES3 project and provides a wide list of technical requirements driven by the work done in WP5/NA4 “Networking on data and model evaluation” and WP3/NA2 “Community engagement” as well as by previous meetings and workshops in the community (i.e. ESGF F2F Conferences). It represents the first step towards the design of the ENES CDI software stack architecture that will be delivered by month 18 in the deliverable D10.1. Such a list of requirements will drive the implementation of the ENES CDI during the IS-ENES3 project and will be periodically revised to (i) adapt to changes with respect to existing needs as well as to (ii) include emerging needs which were not captured before and (iii) assess the degree of fulfilment in the software validation process.

The report describes the applied methodology for formal requirements gathering and documentation collection, and it also provides a comprehensive set of user stories, use cases and a list of technical requirements for the whole ENES Climate Data Infrastructure.

As a result of this phase, preliminary architectural insights have been also proposed in this milestone document, though the final ENES CDI software stack architecture will be defined, documented and presented in the deliverable D10.1.

The report provides background information for the technical scoping in WP2 Task3 “IS-ENES Sustainability”. The ENES-CDI architectural diagram will be coordinated with the ENES-RI[8] (Research Infrastructure) architecture at the management level and consolidated in D10.1.

[8] The ENES-RI roughly consists of the ENES-CDI, IS-ENES HPC and Tools and ESiWACE services.

REFERENCES

[1] IS-ENES project https://is.enes.org/
[2] Earth System Grid Federation https://esgf.llnl.gov/index.html
[3] World Climate Research Programme https://www.wcrp-climate.org/
[4] CMIP https://www.wcrp-climate.org/wgcm-cmip
[5] CORDEX https://cordex.org/
[6] WDCC https://www.dkrz.de/up/systems/wdcc
[7] Climate4Impact (C4I) portal https://climate4impact.eu/
[8] L. Cinquini, D. Crichton, C. Mattmann, J. Harney, G. Shipman, F. Wang, R. Ananthakrishnan, N. Miller, S. Denvil, M. Morgan, Z. Pobre, G. M. Bell, C. Doutriaux, R. Drach, D. Williams, P. Kershaw, S. Pascoe, E. Gonzalez, S. Fiore, R. Schweitzer, "The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data", Future Generation Computer Systems, Volume 36, 2014, Pages 400-417, ISSN 0167-739X, http://dx.doi.org/10.1016/j.future.2013.07.002.
[9] EOSC portal https://www.eosc-portal.eu/
[10] Copernicus https://climate.copernicus.eu/
[11] EGI https://www.egi.eu/tag/eosc/
[12] EUDAT https://www.eudat.eu/
[13] Obs4MIPs https://esgf-node.llnl.gov/projects/obs4mips/
[14] Globus https://www.globus.org/
[15] Apache Solr https://lucene.apache.org/solr/
[16] DataCite - https://datacite.org/index.html
[17] Oracle APEX https://apex.oracle.com/it/
[18] Open Archives Initiative Protocol for Metadata Harvesting https://www.openarchives.org/pmh/
[19] JSON for Linking Data https://json-ld.org/
[20] CMOR https://cmor.llnl.gov/
[21] RabbitMQ https://www.rabbitmq.com/
[22] Pyessy-archive - Archive of standard vocabularies written in pyessv notation https://github.com/ES-DOC/pyessv-archive
[23] Errata Service - https://errata.es-doc.org/
[24] ES-DOC Errata Documentation - https://es-doc.github.io/esdoc-errata-client/
[25] LogStash https://www.elastic.co/logstash
[26] Filebeat https://www.elastic.co/beats/filebeat
[27] S. Fiore, P. Nassisi, A. Nuzzo, M. Mirto, L. Cinquini, D. Williams, G. Aloisio, "A Climate Change Community Gateway for Data Usage & Data Archive Metrics across the Earth System Grid Federation", 11th International Workshop on Science Gateways (IWSG 2019), 12-14 June 2019, Ljubljana, Slovenia
[28] ADAGUC http://adaguc.knmi.nl/
[29] PyWPS https://pywps.org/

[30] OGC WCS http://www.ogc.org/standards/wcs
[31] ICCLIM Documentation: https://icclim.readthedocs.io/ Source Code: https://github.com/cerfacs-globc/icclim
[32]UC Downscaling Portal https://www.meteo.unican.es/en/portal/downscaling
[33] CIM standard https://www.dmtf.org/standards/cim
[34] Working Group on Coupled Modelling https://www.wcrp-climate.org/wgcm-overview
[35] WGCM Infrastructure Panel https://www.wcrp-climate.org/wgcm-cmip/wip
[36] ES-DOC Explorer https://explore.es-doc.org
[37] World Meteorological Organization https://public.wmo.int/en
[38] ECA&D https://www.ecad.eu/
[39] Agile principles https://agilemanifesto.org/principles.html
[40] Requirements 101: User Stories vs. Use Cases https://www.stellman-greene.com/2009/05/03/requirements-101-user-stories-vs-use-cases/
[41] Fulton R, Vandermolen R (2017). "Chapter 4: Requirements - Writing Requirements". Airborne Electronic Hardware Design Assurance: A Practitioner's Guide to RTCA/DO-254. CRC Press. pp. 89–93. ISBN 9781351831420.
[42] Key words for use in RFCs to Indicate Requirement Levels https://www.ietf.org/rfc/rfc2119.txt
[43] Len Bass, NICTA, Sydney, Australia https://www.oreilly.com/library/view/software-quality-assurance/9780128025413/xhtml/Deployability.xhtml
[44] ORCID https://orcid.org
[45] Errata Look-up web page - https://es-doc.github.io/esdoc-errata-client/lookup.html
[46] NetCDF operators https://www.unidata.ucar.edu/software/netcdf/workshops/most-recent/third_party/index.html
[47] Errata create web page - https://es-doc.github.io/esdoc-errata-client/create.html
[48] Errata update web page - https://es-doc.github.io/esdoc-errata-client/update.html
[49] Errata close web page - https://es-doc.github.io/esdoc-errata-client/close.htm
[50] Keycloak https://www.keycloak.org/

GLOSSARY

Table 16. Glossary definitions

Acronym	**Explanation**
ADAGUC	Atmospheric Data Access for the Geospatial User Community
APEX	Oracle Application Express
API	Application Programming Interface
C4I	Climate4Impact
CDO	Climate Data Operators
CF	Climate Forecast
CIM	Common Information Model
CLI	Command Line Interface
CMIP	Coupled Model Intercomparison Project
CMOR	Climate Model Output Rewriter
Copernicus C3S	Copernicus Climate Change Service
CORDEX	Coordinated Regional climate Downscaling Experiment
CSV	Comma-Separated Values
CV	Controlled Vocabulary
DDC	Data Distribution Centre
DOI	Digital Object Identifier
ECAS	ENES Climate Analytics Service
ECA&D	European Climate Assessment & Dataset
EGI	European Grid Infrastructure
ENES	European Network for Earth System Modelling
ENES CDI	ENES Climate Data Infrastructure
EOSC	European Open Science Cloud
ES-DOC	Earth System Documentation

ESGF	Earth System Grid Federation
EU	European
EUDAT	European Association of Databases for Education and Training
F2F	Face to Face
FR	Functional Requirement
GridFTP	File Transfer Protocol (FTP) for Grid Computing
GUI	Graphic User Interface
HTTP	Hypertext Transfer Protocol
IaaS	Infrastructure-as-a-Service
ICCLIM	Indice Calculation CLIMate
IdP	Identity Provider
Input4MIPs	Input Datasets for Model Intercomparison Projects
IPCC	Intergovernmental Panel on Climate Change
IPCC AR6	IPCC Assessment Report 6
IPCC TG-Data	Task Group on Data Support for Climate Change Assessments
IS-ENES	InfraStructure for the ENES modelling
IT	Information Technologies
JRA	Joint Research Activity
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
KPI	Key Performance Indicator
MIPs	Model Intercomparison Projects
NA	Networking Activity
NCO	NetCDF Operators
NetCDF	Network Common Data Form
NFR	Non-Functional Requirement

OAI/PMH	Open Archives Initiative Protocol for Metadata Harvesting
Obs4MIPs	Observations for Model Intercomparisons Project
OGC	Open Geospatial Consortium
PaaS	Platform-as-a-Service
PCMDI	Program for Climate Model Diagnosis & Intercomparison
PID	Persistent Identifier
PO	Project Officer
REST	Representational State Transfer
RI	Research Infrastructure
SaaS	Software-as-a-Service
Sys-admin	System administrator
TNA	Trans-National Access
TXT	Text
UI	User Interface
URL	Uniform Resource Locator
VA	Virtual Access
WCRP	World Climate Research Programme
WCS	Web Coverage Service
WDCC	World Data Center for Climate
WGCM	Working Group on Coupled Modeling
WIP	WGCM Infrastructure Panel
WMO	World Meteorological Organization
WP	Work Package
WPS	Web Processing Service
XML	eXtensible Markup Language

APPENDIX A: SUMMARY OF REQUIREMENTS

List of functional requirements

Table 17. List of functional requirement codes and names for each ENES CDI service

Requirement code	**Requirement name**	**Service**
[DATAFR#1]	Data publication	ESGF Data
[DATAFR#2]	Data search	ESGF Data
[DATAFR#3]	Data download	ESGF Data
[CITFR#1]	CMIP6 compliance	Citation
[CITFR#2]	Insertion or update of citation information	Citation
[CITFR#3]	Access to citation information	Citation
[PIDFR#1]	PIDs assigned for CMIP6 files and datasets	Persistent Identifier (PID)
[PIDFR#2]	Essential metadata registered for CMIP6 file and dataset PIDs	Persistent Identifier (PID)
[DDCFR#1]	Discovery	IPCC DDC at DKRZ
[DDCFR#2]	Download	IPCC DDC at DKRZ
[ERRFR#1]	Data publication	Errata
[ERRFR#2]	Issue viewer	Errata
[ERRFR#3]	Issue list	Errata
[ERRFR#4]	Issue management	Errata
[STATSFR#1]	Data publication statistics	Data Statistics
[STATSFR#2]	Cross-project data download statistics	Data Statistics
[STATSFR#3]	Project-specific data download statistics	Data Statistics
[STATSFR#4]	Data statistics export	Data Statistics

[STATSFR#5]	IS-ENES3 KPIs	Data Statistics
[REPLICFR#1]	Persistent status tracking and recording	Data Replication
[REPLICFR#2]	Multiple transfer protocols and parallel transfers	Data Replication
[REPLICFR#3]	Faceted data characterization	Data Replication
[REPLICFR#4]	Automatic retry	Data Replication
[COMPFR#1]	Server-side data analysis	Compute & Analytics
[COMPFR#2]	Subsetting	Compute & Analytics
[COMPFR#3]	Simple operations	Compute & Analytics
[COMPFR#4]	Complex analysis	Compute & Analytics
[COMPFR#5]	Access control	Compute & Analytics
[COMPFR#6]	API	Compute & Analytics
[COMPFR#7]	Standard interface	Compute & Analytics
[COMPFR#8]	High-level user environment for data analysis	Compute & Analytics
[COMPFR#9]	Compute Service discovery	Compute & Analytics
[C4IFR#1]	Search Wizard Interface	Climate4Impact
[C4IFR#2]	Visualization Service	Climate4Impact
[C4IFR#3]	Documentation and Guidance	Climate4Impact
[C4IFR#4]	User Space Storage (Basket)	Climate4Impact
[C4IFR#5]	Data Subsetting Service	Climate4Impact
[C4IFR#6]	Download Service	Climate4Impact
[C4IFR#7]	Standard Climate Indices and Indicators Calculation Service	Climate4Impact
[C4IFR#8]	Data repackaging Service	Climate4Impact
[C4IFR#9]	Provenance/Lineage Service	Climate4Impact

[ESDOCFR#1]	Documentation publication	ES-DOC
[ESDOCFR#2]	Web-based documentation access	ES-DOC
[ESDOCFR#3]	API-based documentation creation and access	ES-DOC
[ESDOCFR#4]	Further Info URL	ES-DOC
[CFFR#1]	Discussion board	Climate Forecast (CF)
[CFFR#2]	Publication workflow	Climate Forecast (CF)
[DREQFR#1]	Database (DR climate indices)	Data Request
[DREQFR#2]	Programmatic access	Data Request
[DREQFR#3]	Versions	Data Request
[DREQFR#4]	Request ingestion	Data Request
[DREQFR#5]	Database (DR CMIP)	Data Request
[DREQFR#6]	Web Interface	Data Request
[DREQFR#7]	Software	Data Request
[DREQFR#8]	Component versions	Data Request
[DREQFR#9]	Schema	Data Request
[IDFR#1]	Delegation	Identity Management and access entitlement
[IDFR#2]	Delegation with service chaining	Identity Management and access entitlement
[IDFR#3]	Role based access control	Identity Management and access entitlement
[IDFR#4]	Tokens	Identity Management and access entitlement
[IDFR#5]	IdP trust	Identity Management and access entitlement

Non-functional requirements matrix

Table 18. Mapping between services and set of non-functional requirements

	NFR# 1	NFR# 2	NFR# 3	NFR# 4	NFR# 5	NFR# 6	NFR# 7	NFR# 8	NFR# 9	NFR# 10	NFR# 11	NFR# 12	NFR# 13
ESGF DATA	M	M	M	M	M	M	M	M	M	M	M	-	-
CITATION	O	M	-	-	M	-	O	-	M	O	-	M	-
PERSISTENT IDENTIFIER	O	M	M	-	M	O	O	O	-	-	M	-	-
IPCC DDC AT DKRZ	O	M	M	-	M	-	-	O	M	O	-	M	-
ERRATA	M	M	O	M	M	M	O	M	M	O	-	-	-
DATA STATISTICS	M	M	M	O	M	-	M	M	M	O	-	-	M
DATA REPLICATION	-	M	M	-	-	-	-	-	M	-	-	-	-
COMPUTE & ANALYTICS	-	M	M	O	M	-	M	-	M	-	M	-	M
CLIMATE4 IMPACT	M	O	M	O	M	M	-	-	M	M	-	-	O
ES-DOC	M	M	M	M	M	M	M	M	M	M	O	-	-
CLIMATE FORECAST	M	M	-	-	-	-	-	-	-	M	M	-	-
DATA REQUEST	M	M	-	-	-	-	M	-	O	-	O	-	-
IDENTITY MNG & ACCESS ENTITLEMENT	-	-	-	-	-	-	-	-	-	-	R	R	M

Legend:
M = MUST O = OPTIONAL R= RECOMMENDED
NFR#1: Transparency NFR#2: Robustness NFR#3: Scalability
NFR#4: Efficiency NFR#5: Security NFR#6: Reusability
NFR#7: Extensibility NFR#8: Flexibility NFR#9: Usability
NFR#10: Look & feel NFR#11: Interoperability NFR#12: Access control
NFR#13: Deployability

APPENDIX B: TEMPLATES

Table 19. Table template for user story, use case and requirements specification

User story *[if possible I would suggest one user story encompassing user's needs associated to this software component]*
US#n: As an [actor] I want [action] so that [achievement].

Use case(s) specification *[please add one table for each use case]*	
Code: Name	**UC#n**: use case name
Associated User Story	**US#n**
Summary/Description	
Actors	
Basic Course of Events	1. step one 2. step two 3.
Alternative Paths	
Involved components	

Functional Requirements [<code_assigned>**FR]**	
FR code	*FR description*
[<code_assigned>**FR#1]** <brief_name>	
[<code_assigned>**FR#2]** <brief_name>	
Non-Functional Requirements	
List of non-functional requirements codes (codes must be taken from Section3.2). Please add (O) for those which are OPTIONAL and R for those which are RECOMMENDED. Default MUST.	