H2020-EINFRA-2017

EINFRA-21-2017 - Platform-driven e-infrastructure innovation
DARE [777413] "Delivering Agile Research Excellence on European e-Infrastructures"

D7.1 Requirements and Test Cases I

Project Reference No	777413 — DARE — H2020-EINFRA-2017 / EINFRA-21-2017
Deliverable	D7.1 Requirements and Test Cases I
Work package	WP7: IS-ENES/Climate4Impact Use Case
Tasks involved	T7.1 Requirements Elicitation and Prioritisation
Type	R: Document, report
Dissemination Level	PU = Public
Due Date	30/06/2018
Submission Date	20/07/2018 The extension was in agreement with the PO
Status	Draft
Editor(s)	Christian Pagé (CERFACS)
Contributor(s)	Christian Pagé (CERFACS), Alessandro Spinuso (KNMI)
Reviewer(s)	Andreas Rietbrock (KIT)
Document description	This deliverable will report on the identification, assessment and prioritisation requirements based on specific Test Cases and the targeted users. It will take into account the needs based on the current and envisioned RIs and e-infra European Landscape offering. Those

	requirements will be adjusted in later phases by taking into account the related DARE Tools and Services.

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v1.0	25/05/2018	First version	CERFACS
v2.0	08/06/2018	Second version	CERFACS
v3.0	04/07/2018	Version for internal review	CERFACS / KNMI
v4.0	13/07/2018	internal review	Andreas Rietbrock
v5.0 final	16/07/2018	Ready for submission	CERFACS

Executive Summary

The deliverables objectives are to determine and assess the requirements of the Climate Generic Use Case with respect to the DARE Platform Architecture. The Climate Use Case is targeting several potential user categories. Requirements need to be determined according to different needs and different user categories as tailored interfaces can be built on top of what the use case will provide.

From the generic workflows, requirements have been extracted and detailed. Needed components and interfaces in an architectural context are being discussed.
Available infrastructures, e-infrastructures, processing tools, interfaces and standards were assessed and identified as possible components for implementation. Also, a list of essential information to be passed through the interfaces was compiled. Finally, three levels of implementation have been identified, for increasing complexity and more complete integration of the Use Case. The provenance aspects will be very important as they are involved in almost all the workflow steps.

Many decisions for the overall architecture will need to be taken: the component definitions, the existing tools, and the interfaces implementation will highly depend on which architecture choices will be taken. For example, the use of the generic EUDAT GEF Service could be seen as a way to generically encapsulate the DARE platform. Alternatively, it also could be seen as a way for the DARE Platform to deploy calculations on-demand. The successful implementation and adoption of the generic climate Use Case will depend on those choices.

Table of Contents

List of Terms and Abbreviations

Abbreviation	Definition
ESGF	Earth System Grid Federation
C4I	climate4impact
EGI	European Grid Infrastructure
ENES	European Network for Earth System modelling
ENES CDI	European Network for Earth System modelling Climate Data Infrastructure
EUDAT CDI	EUDAT Common Data Infrastructure

1 Introduction

1.1 Purpose and Scope

The deliverable objectives are to determine and assess the requirements of the Climate Generic Use Case with respect to the DARE Platform Architecture. The Climate Use Case is targeting several potential user categories, because tailored interfaces can be built on top of what the use case will provide, through an Application Programming Interface (API).

The purpose of the deliverable is to build a requirements list for this use case, that needs to cover a large number of variations in the workflow along the generic aspects, according to different needs and different user categories. This list will be used to design a proper architecture for the DARE Platform, and ensure that all needs of this generic use case are covered.

1.2 Approach and relationship with other Work Packages and Deliverables

The approach used here will be to detail the use case itself, related to different aspects such as user categories and user needs, generic aspects, and research infrastructure landscape. This deliverable is closely linked to WP2 that will design the architecture, and D2.1 (DARE Architecture and Technical Positioning).

1.3 Methodology and Structure of the Deliverable

The structure of this deliverable is as follows. First, the generic Climate Use Case will be summarized, followed by a description of the user needs using a detailed description of several workflows. Requirements will be extracted from those workflows and detailed. Components and interfaces needed in an architectural point of view will be discussed.

2 Climate Use Case Summary

This section will summarize the Climate Use Case at a high-level, along with some information about its motivation and generic aspects.

2.1 Motivations

The scientific climate modeling community has done a lot of research on climate change within the past several decades, and climate models have improved significantly. Large international intercomparison of model output experiments are regularly carried out, to evaluate the climate models' performances, and called CMIP. Those simulations are used in periodic IPCC reports to guide stakeholders on climate change impacts in the future climate.

Access to climate simulations are important for the climate change impact community, who performs research and assesses the impacts of climate change on society in general. There are several types of users that are part of this heterogeneous community, ranging from impact modelers, PhD and Post-Doc students, research engineers, climate researchers, to practitioners, etc. Those users have different needs, and different technical and scientific knowledge.

The normal workflow for climate researchers has always been to download the data files, then post-process and analyze the data locally on their own systems. For other types of users, it is rather through the use of platforms and interfaces to access the data, that can reformat it into easier more user friendly data formats.

The current climate simulations archive is stored on an international federation of data nodes (Earth System Grid Federation, ESGF). For the last intercomparison exercise, CMIP5, the total size of the archive is on the order of 2 Pb. For the upcoming CMIP6 the archive size will be on the order of 30 to 50 Pb. Even currently, with a 2 Pb archive size, users have difficulties to use all data they need, because of time and efforts it takes to download and process all data. Many users also do not have the proper technical knowledge, or local technical infrastructure.

The IS-ENES consortium has developed a platform for tailored and easier access to climate simulations, called climate4impact (C4I https://climate4impact.eu/). This platform provides several services as well as on-demand data processing. It is part of the ENES Climate Data Infrastructure (CDI). Currently, the operational processing backend for C4I is the C4I server itself. Some prototypes exists to delegate calculations to: a) EGI FedCloud using the EUDAT GEF Services; b) ESGF Computing Nodes using the CWT API.

There is a strong need for all categories of users to have on-demand processing capabilities closer to the data storage. There is also need that the whole workflow is more streamlined, hiding technical details such as: a) the fragmentation of data into separate data files; b) different versions of the same data; c) replicated version stored into several ESGF data nodes. Having proper lineage (provenance).

2.2 Current scientific workflows for different user categories

A lot of work has been done within several European Projects, in climate science, on analyzing and identifying users needs with respect to scientific workflows for data analytics. We can cite, notably, IS-ENES and IS-ENES2 as well as CLIPC and CIRCLE2; however, several others and some related projects have also contributed to the analysis. The following user categories can be identified:

1. Climate Scientific Researchers and Engineers (seniors, post-docs, PhD students, research engineers …)
   - They are in the same scientific domain as the data.
2. Climate Change Impact Modellers and Researchers in other scientific domains
   - They are mainly working on the impacts of climate change on several aspects, such as crops, agriculture, land use, water management, tourism, hydrology, hydro-geology, etc.
3. Practitioners/Boundary Workers
   - They are working as interfaces between end users and climate researchers.
4. Stakeholders
   - They are developing/implementing policies based on climate change information, given by boundary workers. Those boundary workers are experts that are providing written reports to stakeholders based on their requests.

Climate simulation data are stored in an international network of peer-to-peer federated data servers, with replication of core datasets among them. The current archive has a size on the order of 2 Pb, and contains all simulations from the last Climate Model Intercomparison Program (CMIP5) used as a basis for IPCC reports.

Climate Scientific Researchers and Engineers

The usual workflow is still to firstly download needed data files locally, and then analyze the data on local servers. The difference within the past few years is that the download is more centralized in institutions, so that downloaded data is better organized to be used easily by researchers and engineers. Often, data is post-processed to be uniformized, as in many case the datasets are still heterogeneous. Additionally, post-processing servers are often centralized with the data, so users are no longer analyzing data on their own computer station.

A few attempts to provide centralized post-processing servers for climate data along with the whole data archive exist, but not all researchers can get access to those storage and computing resources. Those centralized solutions are often national (e.g. DRIAS Climat in France, KNMI Climate Explorer in the Netherlands, klimatscenarier in Sweden) or institutional (Jasmin by BADC in the UK, CICLAD by IPSL in France, …), and several front-ends exist with different interfaces, such as:

- Python Jupyter notebooks
- OGC Web Services
- OpenDAP and or THREDDS NetCDF
- Standard http and/or ftp
- Shell access
- Tailored Web Interfaces

Climate Change Impact Modellers and Researchers in other scientific domains

Users of this domain often rely on Climate Research Engineers to provide them tailored and reformatted datasets. They typically do not have the technical knowledge to handle the format and complexity of the original data format and often work with ASCII or GIS files, with Excel and/or GIS software on Windows computer systems. They also have very limited technical resources and support, such as storage, bandwidth and computing power and need a lot of support from climate researchers in interpreting and using properly climate simulations.

Specific tailored web platforms have been designed to fulfill the needs of this user community, but mostly for Climate Change Impact Modellers. There are still some technical aspects involved in using those platforms, such as C4I https://climate4impact.eu/ and CLIPC http://www.clipc.eu/ . On th other hand, national specific platforms such as DRIAS http://www.drias-climat.fr/ are often less technical and offer bias-corrected data on a subset of available data. Those platforms provide several services such as visualization, search, quicklooks, downloads, processing, guidance, to ease the access to climate datasets with more intuitive interfaces than ESGF data nodes.

The current workflow of those users highly depends if they use or do not use these dedicated web platforms. The most common workflow at this time is to use web platforms to get access to the data they need, and . So if we compare the workflow of those users compared to the workflow of climate researchers, the difference is only at the beginning and the end of the workflow, because of the difference in interaction with the web platforms, and needed guidance.

Practitioners/Boundary Workers

These users have a lot of experience in interpreting and analyzing climate simulations, but they still need many interactions with climate researchers, especially on guidance. The act as a bridge between climate researchers and stakeholders, also sometimes industry. They provide complete and summarized reports based on stakeholders' needs.

Their workflow is similar to those of climate change impact modellers, but in addition they need to producing synthetic reports for their clients. However they often need less guidance (scientific and technical) as they have more knowledge because of their experience.

Stakeholders

They rely on summarized reports written by Boundary Workers (see category above), but sometimes also by scientific researchers themselves (such as IPCC reports). Their needs are very specific and concise, as they have little time to evaluate those reports in order to define actions.

2.3 Current limitations and problems given needs

With the rapid increase in the total data volume to be processed, users need urgently robust and scalable solutions. The latest climate simulation archive was on the order of 2 Pb (2014), the next one, being generated right now, will have a size of about 20 to 50 Pb (2018). This means that:

- Data will be stored at several locations
  - Core simulation datafiles will be replicated in several locations
  - Experiments are segmented into several datafiles, but grouped into same locations
- Data reduction is mandatory, as near as possible to the data based on network bandwidth
- Data processing by data providers and/or between data providers and end users will be necessary
- No scheduler exists to choose from which data location one user needs to get data from, as it will depend on the **combination** of
  - Server proximity wrt network bandwidth capacity and current availability
  - Available processing power for data reduction and processing
  - Availability of specifics data node
- Standardization will be a key to automate every aspects of the workflow
- No provenance standard and model is currently in use, but a prototype exists
- No workflow language is used to perform data access and execute data processing

- Datafiles are not abstracted to users. This complexify data access, because needed data can span several data nodes, several datafiles, etc.

3 Components, interfaces, tools: identifying missing parts

3.1 Current status: available infrastructures, e-infrastructures, interfaces, components

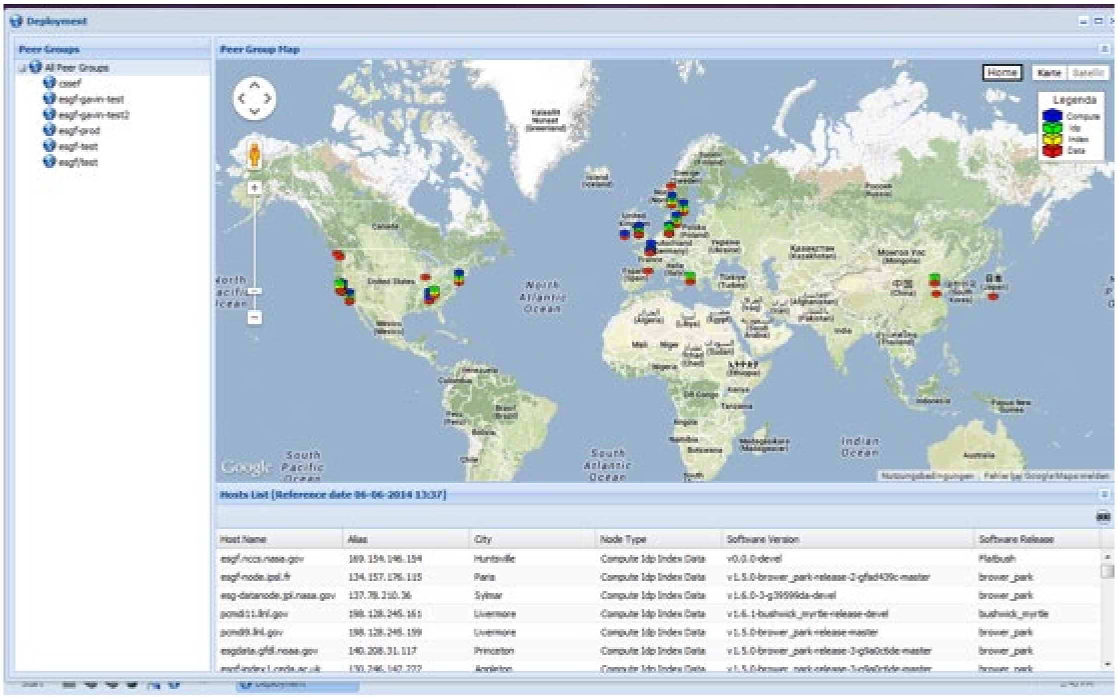
ESGF

The Earth System Grid Federation (ESGF) Peer-to-Peer (P2P) enterprise system is an open-source collaboration that develops, deploys and maintains an operational software infrastructure for the management, dissemination, analysis of model output, and observational data.

ESGF P2P is an architecture specifically designed to handle large-scale data management for worldwide distribution. It is a federation of data nodes (and soon also computing nodes) with dynamic replication of core datasets. Model simulations, satellite observations, and reanalysis products are all being served from the ESGF P2P distributed data archive.

The ESGF security architecture supports OpenID and PKI based authentication for services, and OAuth2 is currently being deployed. Applications are secured with an authorisation filter, middleware to intercept incoming requests and refer authorisation decisions to an Authorisation Service over a SAML/SOAP interface.

Each ESGF data node exposes its data catalog using a THREDDS Server (TDS). The THREDDS Data Server (TDS) is a web server that provides metadata and data access for scientific datasets, using OPeNDAP, OGC WMS and WCS, HTTP, and other remote data access protocols. TDS is developed and supported by Unidata.



Hosts List [Reference date 06-06-2014 13:37]

Host Name	Alias	City	Node Type	Software Version	Software Release
esgf.nccs.nasa.gov	169.154.146.154	Huntsville	Compute Idp Index Data	v0.0.0-devel	Flatbush
esgf-node.ipsl.fr	134.157.176.115	Paris	Compute Idp Index Data	v1.5.0-brower_park-release-2-gfad439c-master	brower_park
esg-datanode.jpl.nasa.gov	137.78.210.36	Sylmar	Compute Idp Index Data	v1.6.0-3-g39599da-devel	brower_park
pcmdi11.llnl.gov	198.128.245.161	Livermore	Compute Idp Index Data	v1.6.1-bushwick_myrtle-release-devel	bushwick_myrtle
pcmdi9.llnl.gov	198.128.245.159	Livermore	Compute Idp Index Data	v1.5.0-brower_park-release-master	brower_park
esgdata.gfdl.noaa.gov	140.208.31.117	Princeton	Compute Idp Index Data	v1.5.0-brower_park-release-3-g9a0c6de-master	brower_park
esgf-index1.ceda.ac.uk	130.246.142.222	Appleton	Compute Idp Index Data	v1.5.0-brower_park-release-3-g9a0c6de-master	brower_park

Figure 1: ESGF Network

EUDAT CDI

The EUDAT Collaborative Data Infrastructure (CDI) is essentially a European e-infrastructure of integrated data services and resources to support research. It consists of a network of nodes providing services for upload and retrieval, identification and description, movement, replication and data integrity. It also provides additional services that are needed to operate the infrastructure.

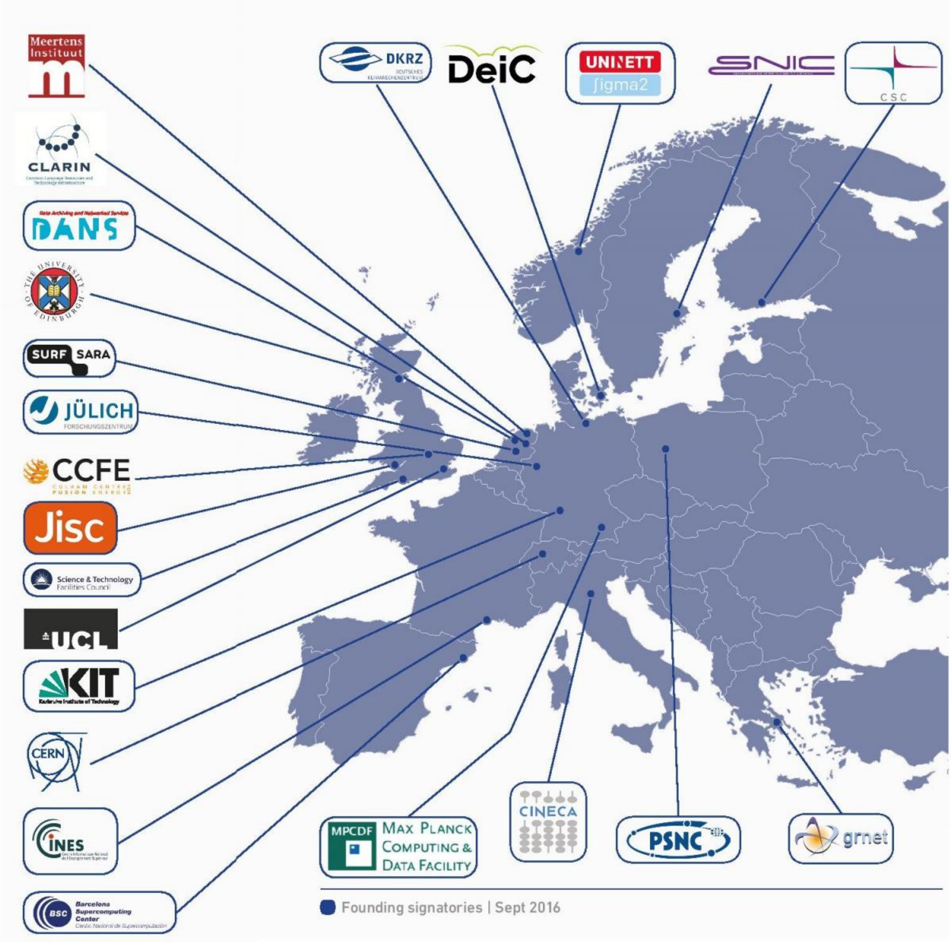

Meertens Instituut
CLARIN
DANS
The University of Edinburgh
SURF SARA
JÜLICH
CCFE
Jisc
Science & Technology Facilities Council
UCL
KIT
CERN
CINES
Barcelona Supercomputing Center
DKRZ
DeiC
UNINETT Sigma2
SNIC
CSC
MPCDF Max Planck Computing & Data Facility
CINECA
PSNC
grnet
Founding signatories | Sept 2016

Figure 2: EUDAT CDI

The EUDAT CDI provides a collection of services (the B2 Services):

- B2NOTE: Service that allows you create, search and manage annotations on research data.
- B2SAFE: Data replication and long term preservation service to implement data management policies on your research data across multiple administrative domains in a trustworthy manner.
- B2HANDLE: Tools for managing persistent identifiers within the EUDAT Collaborative Data Infrastructure enabling you to register data, and thus making possible to reference or cite them in the future.
- B2DROP: Secure and trusted cloud storage to store and exchange data, stipulating how, with whom and for how long, while accessing up to 20Gb of storage. It also allows automatic desktop synchronization of large files
- B2SHARE: Repository for shareable digital objects to improve your data sharing and publishing and guarantee long-term persistence of your locally-stored data.
- B2FIND: B2FIND is a discovery service based on metadata steadily harvested from research data collections from EUDAT data centres and other community repositories.
- B2STAGE: Reliable, efficient, light-weight and easy-to-use service to transfer research data sets between EUDAT data resources and (High-Performance) Computing workspaces.

- B2ACCESS: EUDAT federated cross-infrastructure authorisation and authentication framework for user identification and community-defined access control enforcement.

The EUDAT CDI is also developing future services. The EUDAT Generic Execution Framework (GEF) enables execution of containerized software tools on data stored in the EUDAT CDI.

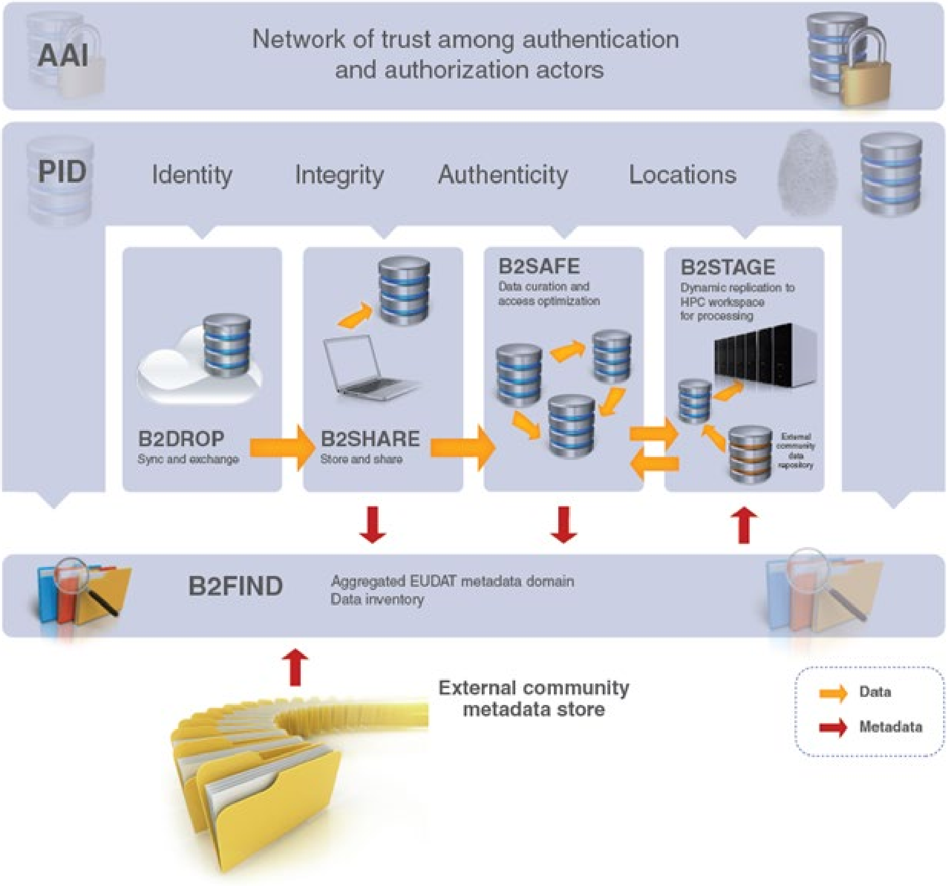
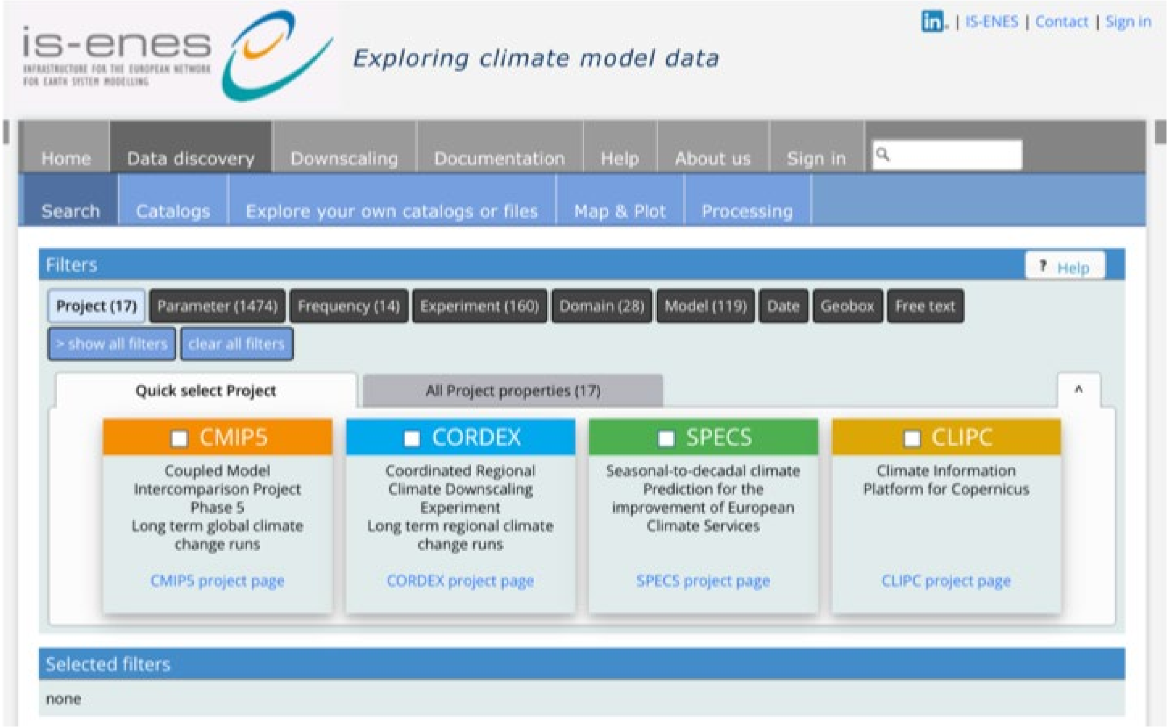

AAI
Network of trust among authentication and authorization actors
PID
Identity
Integrity
Authenticity
Locations
B2SAFE
Data curation and access optimization
B2STAGE
Dynamic replication to HPC workspace for processing
External community data repository
B2DROP
Sync and exchange
B2SHARE
Store and share
B2FIND
Aggregated EUDAT metadata domain
Data inventory
External community metadata store
Data
Metadata

Figure 3: EUDAT Services

ENES CDI C4I

The ENES Climate Data Infrastructure provides data access for climate research model data starting with providing data from CMIP5/CMIP6 and continuing with CORDEX. It includes ESGF European data nodes and portals. It also includes the climate4impact (C4I) Operational Platform which ease access to model data for the climate impact research community.

The C4I provides several services and standards, accessible through APIs. It also exposes on-demand services using ISO/OGC standards such as WPS and WMS. The main services of the C4I platform are visualization, search, download, processing (on-demand, also providing subsetting), statistical downscaling, authentication/authorization delegation to ESGF data nodes (and in the near future to computing nodes). All those services can be accessed programmatically or alternatively with an easy to use web interface. C4I can also access any OpenDAP and THREDDS remote catalog, dataset and file(s). It provides users a ‘user space’ to store intermediate results along with a basket feature to store data selections.



Coupled Model Intercomparison Project Phase 5
Long term global climate change runs

Coordinated Regional Climate Downscaling Experiment
Long term regional climate change runs

Seasonal-to-decadal climate Prediction for the improvement of European Climate Services

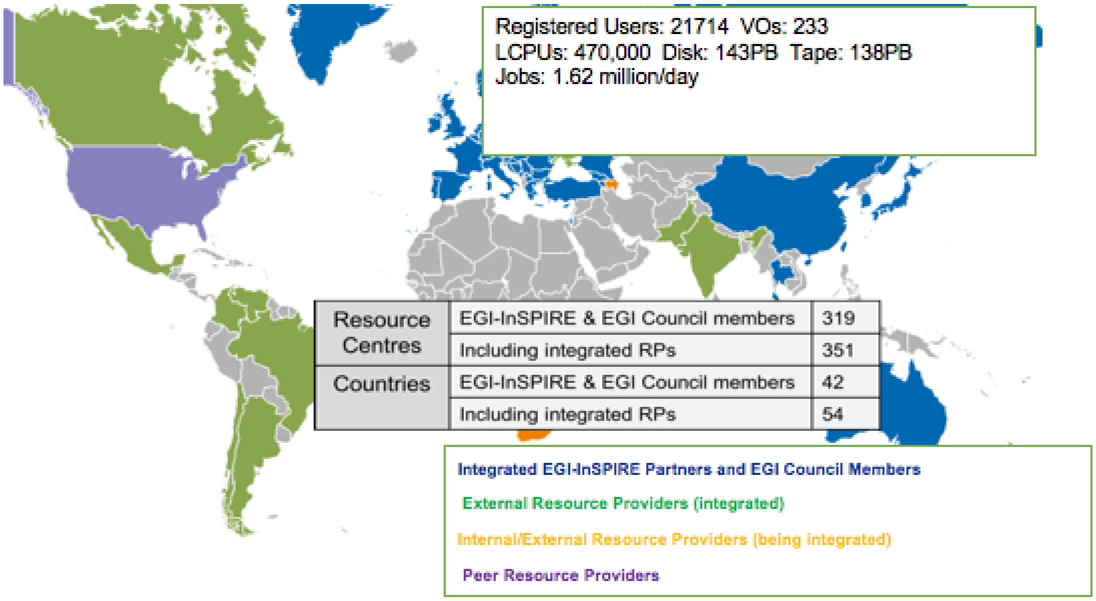
Climate Information Platform for Copernicus

Figure 4: ENES CDI climate4impact.eu

EGI FedCloud

The EGI Federated Cloud is part of EGI infrastructure. It's a seamless network of private clouds and virtualised resources, built around open standards and focusing on the requirements of the scientific community. The FedCloud is targeted at researchers and research communities that need to access resources. The EGI FedCloud currently federates OpenStack, OpenNebula and Synnefo technology based clouds.

To use the EGI FedCloud one has to request specific resources at specific providers and launch VMs. The REST technology is used to get the list of certified cloud providers from the EGI Application Database, and for each provider a list of available Virtual Appliances and resource templates can be retrieved. To develop applications that can deploy calculations on the FedCloud, one has to use the jOCCI-api Java library. It Implements transport functions for rendered OCCI (Open Cloud Computing Interface) queries. It is built on top of jOCCI-core and currently provides HTTP transport functionality with set of authentication methods and a basic request interface to easily communicate with OCCI servers.



Registered Users: 21714 VOs: 233
LCPUs: 470,000 Disk: 143PB Tape: 138PB
Jobs: 1.62 million/day

Resource Centres	EGI-InSPIRE & EGI Council members	319
	Including integrated RPs	351
Countries	EGI-InSPIRE & EGI Council members	42
	Including integrated RPs	54

Integrated EGI-InSPIRE Partners and EGI Council Members
External Resource Providers (integrated)
Internal/External Resource Providers (being integrated)
Peer Resource Providers

Figure 5: EGI

3.2 Authentication/Authorization Systems

EUDAT B2ACCESS

B2ACCESS is a secure Authentication and Authorization platform developed by EUDAT. B2ACCESS can be integrated with any service. When B2ACCESS is integrated with a given service, the user may log in by using different methods of authentication:

- Home organisation identity provider
- Google account
- EUDAT ID

It supports several methods of authentication via the users' primary identity providers (OpenID, SAML, x.509), as well as EduGain. It permits group, community and service managers to specify authorisation decisions. It is based on the Unity IDM technology.

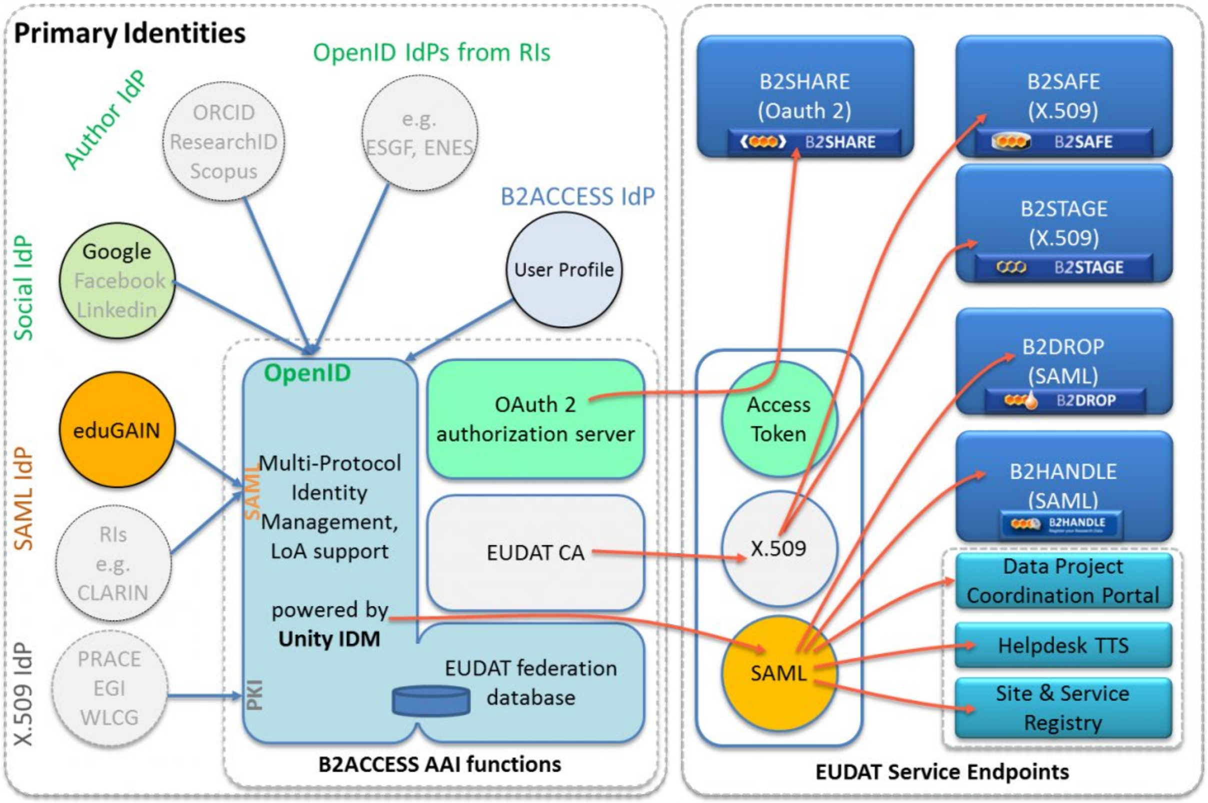


Figure 6: EUDAT B2ACCESS

ESGF OpenID

Up to now, OpenID 2.0 was used as the authentication technology for ESGF, but as it is being deprecated, ESGF is moving toward OpenID Connect (OIDC) is an authentication layer on top of OAuth 2.0, an authorization framework. It should be deployed in 2018.

EGI Certificates

EGI Infrastructure as a Service (IaaS) Cloud Resources can be accessed through Virtual Organizations (VOs). A VO is a grouping of IaaS cloud provider from the EGI federation, who allocate capacity for a specific user group. There are generic VOs too. A VO needs to be joined before accessing EGI IaaS cloud resources, while higher level services (PaaS, SaaS) do not always require VO membership. VO membership is controlled in EGI by X.509 certificates.

3.3 Metadata Standards

EUDAT CDI

The EUDAT CDI provides a metadata service (B2FIND) to search through metadata records.

NetCDF CF Convention

The conventions for CF (Climate and Forecast) metadata are designed to promote the processing and sharing of files created using the NetCDF API. 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 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.

ESGF CMOR

All data published on ESGF Data Nodes need to comply with strict metadata rules, with proper facets. The ESGF publisher can be configured with metadata rules for each experiment. For CMIP5 data, all published data must comply with the CMOR standards https://cmor.llnl.gov/ and is also using a specific DRS for file naming and directory structure.

ENES CDI C4I

C4I is assuming that data files follow the NetCDF CF Convention http://cfconventions.org/

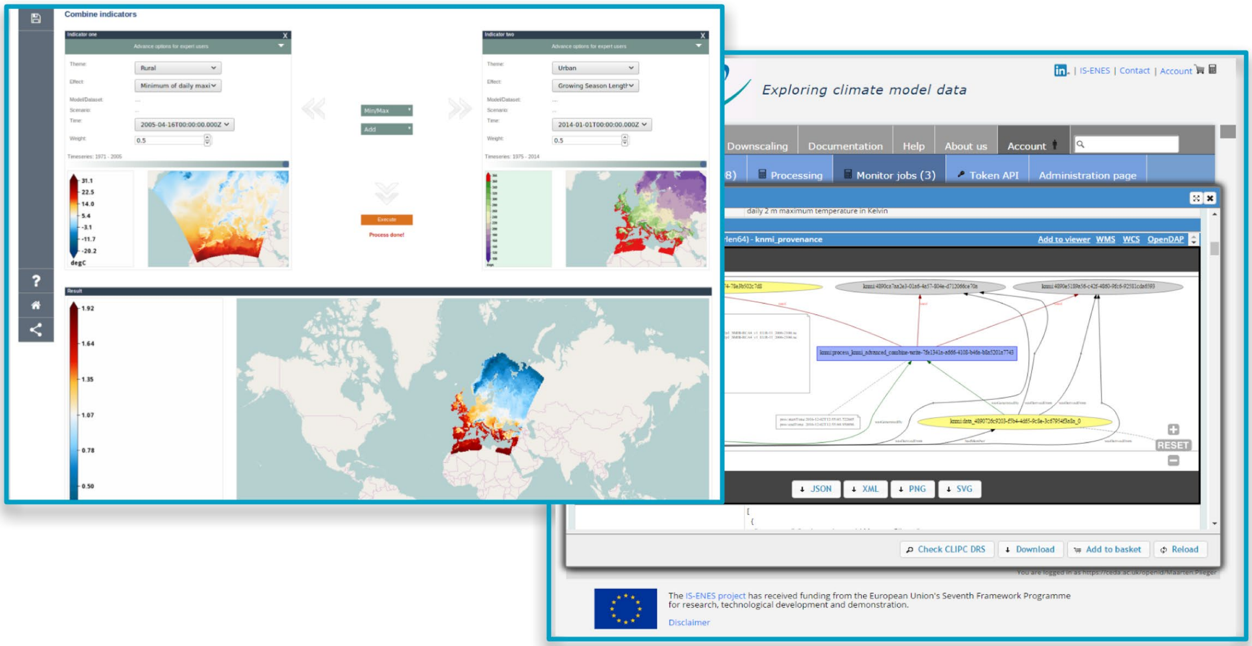
3.4 Lineage/Provenance

Users of the DARE platform should be able to explore and validate the origin and means of automated products as well as the results obtained by their own computations by means of lineage information represented in a comprehensive provenance data model. Here the relationships between the input data used by the functions adopted by the user-defined methods (or workflows) and the produced output, including the applied parameterization, should be described with substantial metadata. To guarantee coverage, consistency and interoperability of the collected provenance data, we will investigate how to integrate climate concepts and metadata within a standard model provenance such as PROV-DM https://www.w3.org/TR/prov-dm/.

To achieve a satisfactory coverage of the provenance data, the computational model that characterises the DARE workflows should merge with domain specific information, possibly combining the automated extraction of metadata according to the adopted data formats and schemas (as shown in the previous section) with user specific additional properties. We will initially rely on existing extensions of common model for processes and workflow provenance, such as ProvONE (https://purl.dataone.org/provone-v1-dev) and its extension S-PROV, the latter developed in the context of DARE. This baseline will be further elaborated to reach a satisfactory level of granularity of the lineage information and richness of context, with respect to application and system requirements. We will investigate what is available within existing and emerging controlled vocabularies and ontologies (http://iridl.ldeo.columbia.edu/ontologies/).

We will start from evaluating the experience and the results obtained within the CLIPC project (www.clipc.eu). The project aim was to offer a climate impact toolkit to evaluate, rank and combine climate impact indicators. These indicators synthesise effects of future climate change with relevance to a specific sector and business. Applications include research or decision- support in policy and practice. Thus, CLIPC allows users to combine indicators into a new indicator. For instance, they can add up climate indicators or create a difference map. The service required an instrumented and systematic provenance system that could capture the interactive execution of a combine function driven by the users. It has to be able to extract the relevant metadata and link all data products and intermediate results that are used for the production of a new indicator.

The approach is to develop a workflow with the dispel4py technology, that captures provenance assertions and data dependencies consistently with the adopted conventions and metadata, as agreed within the project. Hence, the workflow, that can be exposed as a WPS, delivers interoperable provenance reports that are accessible from a database or embedded in the NetCDF output files, which are the new impact indicators. Figure 7 and Figure 8 show two visualisations and access modes of the produced lineage. The former adopts standard PROV visual tools and conventions integrated in the climate4impact.eu portal, while the latter shows the same trace accessed interactively through the S-ProvFlow system. The S-ProvFlow will be the one used for lineage visualisation and management technology supported and further developed within DARE.



Theme: Rural
Effect: Minimum of daily maxi
Time: 2005-04-16T00:00:00.000Z
Weight: 0.5
Timeseries: 1971 - 2005

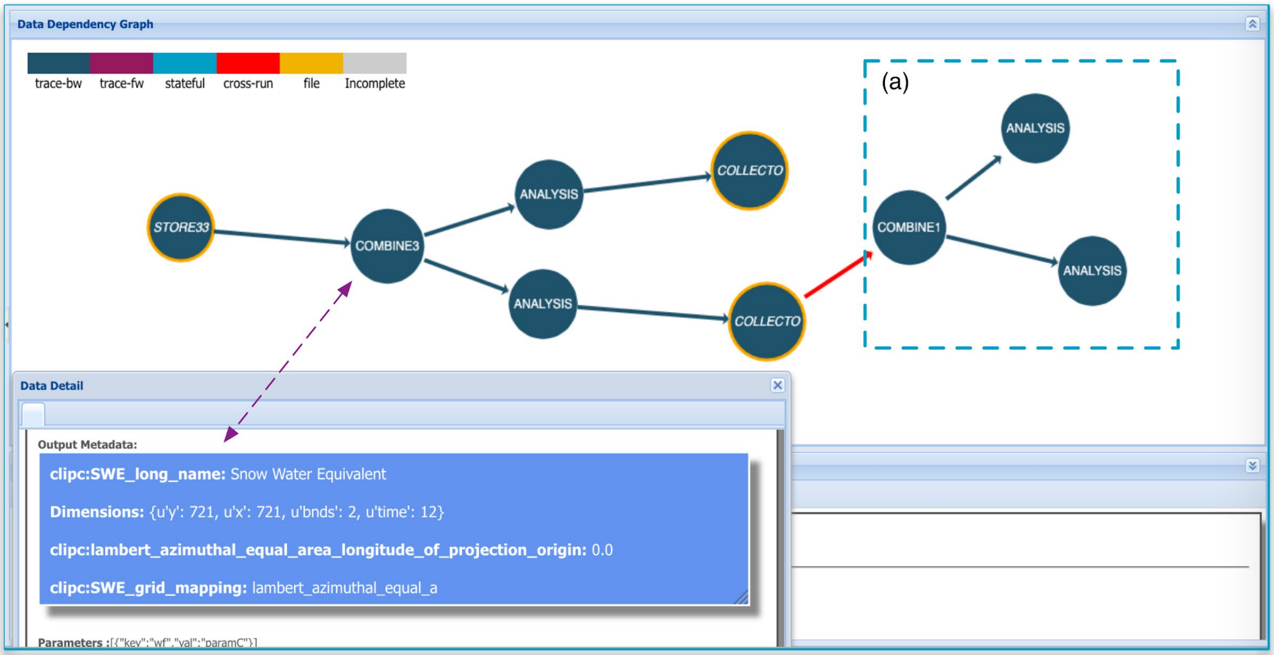
Theme: Urban
Effect: Growing Season Length
Time: 2014-01-01T00:00:00.000Z
Weight: 0.5
Timeseries: 1975 - 2014

daily 2 m maximum temperature in Kelvin

len64) - knmi_provenance

The IS-ENES project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration.
Disclaimer

Figure 7: CLIPC: GUI for the combine function and a view of the provenance trace for the resulting climate indicator within the climate4impact.eu portal.



Output Metadata:

clipc:SWE_long_name: Snow Water Equivalent

Dimensions: {u'y': 721, u'x': 721, u'bnds': 2, u'time': 12}

clipc:lambert_azimuthal_equal_area_longitude_of_projection_origin: 0.0

clipc:SWE_grid_mapping: lambert_azimuthal_equal_a

Parameters :[{"key":"wf","val":"paramC"}]

Figure 8: CLIPC: S-ProvFlow Visualisation of a portion of the lineage (backwards navigation of the data dependencies) of

a climate indicator obtained by the workflow implementing a combine function. The use of the automated metadata extraction from the NetCDF files guarantees the consistent representation of the application context, as seen in the frame in the bottom left. The square (a) instead shows that an indicator produced in a previous analysis has been reused in another computation (cross-runs). This is obtained by the automated generation and attribution of uuids to the results, and their consistent re-use in the lineage traces. This is achieved thanks to the data-format awareness of the dispel4py workflow technology.

ESGF Computing Nodes

In order to deal with larger data volumes and the need of data reduction near data storage (saving network bandwidth of both sides user and server), the ESGF partners have created a Compute Working Team (CWT) on defining the API of the future ESGF computing nodes: https://github.com/ESGF/esgf-compute-api . Those will have direct access to the ESGF data nodes storage system. Each ESGF data node manager will have the possibility to provide a compute service through a computing node. Each ESGF computing node will be certified by validating the API implementation as well as the processing functions.

A test computing node has been installed, but no operational server exists up to now. and will be deployed lated in 2018. The API Authentication is not operational: for now an API short-lived key can be generated to access the computing node API.

ENES CDI C4I

The climate4impact (C4I https://climate4impact.eu/) platform provides an on-demand processing service using the OGC WPS (ISO) standard. Several processing functions have been implemented, some are prototypes, some are operational. Calculations are performed on the C4I server itself, but delegation of calculations is needed as the C4I server is not powerful enough to support a large number of parallel processing requests.

The software used is either the icclim open-source software (http://icclim.readthedocs.io/en/latest/), or directly python code since the WPS implementation C4I is using is pyWPS 1.x.

A processing delegation prototype has been developed using the EUDAT GEF deployed on the EGI FedCloud. Another prototype has been developed using the ESGF Computing Node API, on the CWT test server.

EGI FedCloud

The EGI FedCloud provides computing resources. A prototype of using the EUDAT GEF service to deploy a docker container on an EGI FedCloud VM to perform some calculations have been tested successfully.

4 Summary of Requirements

4.1 Open Questions

EUDAT Services

Which EUDAT Services are candidates to be used in the *DARE* platform? How to integrate with dispel4py?

It is expected that the future **EOSC** will provide services through the **EUDAT B2 Services**. Potential services are listed below, along with some explanations. It is also to be considered that most of those services are integrated, which means that there are interfaces between them so that they can be used together.

- **EUDAT B2ACCESS:** provides an authentication and authorization service. It can be used as an integrated authentication system within *DARE*.

- **EUDAT GEF**: provides a framework and an API as a generic processing/execution service. It uses containerization technology (docker) to deploy the backend of the service on external resources. It can be used to manage the workflow within the *DARE* platform.
- **EUDAT B2DROP:** storage system for small datasets (max 20 Gb), like Dropbox (it is using the owncloud technology). Within *DARE*, it can be used to store processing results before those are retrieved on external platforms, such as C4I.
- **EUDAT B2SHARE:** storage system for large datasets. It supports metadata records, and has a stable API. It supports access policies, as well as PIDs. Data integrity is ensured by using checksums. The *DARE* platform could be interfaced with the **B2SHARE API** to access and store datafiles that are processed.
- **EUDAT B2NOTE:** annotation system for files stored into **B2SHARE**. It could be used within *DARE* to annotate processed datafiles that are stored into **B2SHARE**.
- **EUDAT B2HANDLE**: PID Service. It could be used to assign PIDs to the output files resulting from processing occurring within the *DARE* platform.

EGI FedCloud

The EGI FedCloud provides computing resources. We should consider testing the deployment of *DARE* components onto the EGI FedCloud, maybe using the EUDAT GEF?

Data Life Cycle and Workflow Description and System

The DARE architecture should describe and support the whole Data Life Cycle (DLC) of the Use Case. The main component in the DLC is the workflow that supports the whole event chain. The Use Case will require passing of information from the interactive user platform (C4I) to the DARE platform, such as:

1. input file(s)/dataset(s),
2. processing function(s),
3. function(s) parameters,
4. output file(s) storage,
5. provenance and lineage information creation,
6. PID assignments,
7. authentication,
8. processing status and progression.

The initiating platform C4I will need to specify the information listed in 1, 2 and 3, 7, and maybe even 4 (to be evaluated). This specification should be done using a workflow language, either using a known workflow engine, or using a parameter file like JSON.

4.2 Missing parts

Here we identify missing parts to support the Use Case using the DARE platform, notably to interface the *DARE* platforms with external infrastructures and components. The technical solution still need to be designed. The different levels shown here are from increasing complexity and functionality, so Level 1 are mandatory, while Level 2 are extra functionalities, and Level 3 are optional but desired features.

Level 1

- DARE (dispel4py) <-> ESGF Data Nodes. Using C4I ESGF credentials delegation? *Workflow; Authentication.*
- DARE (dispel4py) <-> ENES CDI C4I (using WPS). *Workflow; Provenance; Authentication.*
- C4I <-> DARE (dispel4py): send/receive instructions through WPS. *Workflow; Authentication.*
- dispel4py -> mapping of calculation/processing functions (e.g. icclim python package). *Workflow; Provenance.*

Level 2

- DARE (dispel4py) <-> B2DROP. *Workflow; Authentication.*
- C4I: GUI. *Workflow; Authentication.*

Level 3

- DARE (dispel4py) <-> ESGF Computing Nodes. Using C4I ESGF credentials delegation? *Workflow; Provenance; Authentication.*

5 General Conclusions

Starting from users' needs and from a generic Use Case in the climate scientific domain, specific requirements were identified. Available infrastructures, e-infrastructures, processing tools, interfaces and standards were assessed and identified as possible components for implementation. Also, a list of essential information to be passed through the interfaces has also been identified. Finally, three levels of implementation have been identified, for increasing complexity and more complete integration of the Use Case. The provenance aspects will be very important as they are involved in almost all the workflow steps.

Many decisions for the overall architecture will need to be taken: the component definitions, the existing tools, and the interfaces implementation will highly depend on which architecture choices will be taken. For example, the use of the generic EUDAT GEF Service could be seen as a way to generically encapsulate the DARE platform, or it could also be seen as a way for the DARE Platform to deploy calculations on-demand. The successful implementation and adoption of the generic climate Use Case will depend on those choices.