H2020-EINFRA-2017

EINFRA-21-2017 - Platform-driven e-infrastructure innovation

DARE [777413] "Delivering Agile Research Excellence on European e-Infrastructures"

D7.3 Pilot Tools and Services, Execution and Evaluation Report I

Project Reference No	777413 — DARE — H2020-EINFRA-2017 / EINFRA-21-2017
Deliverable	D7.3 Pilot Tools and Services, Execution and Evaluation Report I
Work package	WP7: IS-ENES/Climate4Impact Use Case
Tasks involved	T7.3 IS-ENES Pilot Development; T7.4 Evaluation
Type	R: Document, report
Dissemination Level	PU = Public
Due Date	30/06/2019 Deadline extended to 31/07/2019 in agreement with PO
Submission Date	31/07/2019
Status	Final Draft
Editor(s)	Christian Pagé (CERFACS)
Contributor(s)	Christian Pagé (CERFACS)
Reviewer(s)	Iraklis Klampanos (NCSRD)
Document description	A report detailing the set of DARE components activated for the use case, their configuration parameters and possible customization steps (if needed). It will also report on pilot execution benchmarks to assess the usability and performance of the pilot. At early stage the pilot prototype will be evaluated, and results will feed in the next stages.

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v1.0	03/06/2019	First version	Christian Pagé
v2.0	01/07/2019	Draft ready for internal review	Christian Pagé
v3.0	24/07/2019	Internal review completed	Iraklis Klampanos
v4.0	29/07/2019	Final Draft submitted	Christian Pagé

Executive Summary

The objectives of this deliverable are to evaluate the Climate Science Domain pilot prototype. The evaluation is done by targeted users of the DARE platform, and the results feed into the next stages of development. The targeted users are software developers of the Climate Research Infrastructure, and the evaluation session has been organized in Utrecht, Netherlands, on June 17th, 2019. From 16 participants, 8 has taken time to fill the evaluation form. The results show a great interest in what the DARE Platform can bring and provide, and the evaluation has been quite positive. Regarding a future adoption of the DARE Platform by the community, results and discussions show that, overall, the adoption should take place. However, more work needs to be done, both on some additional features as well as proper dissemination and explanation. A future evaluation, of a closer to operational version of the pilot, will also need to demonstrate the benchmarks and added capabilities of the pilot implementation.

Table of Contents

List of Terms and Abbreviations

Abbreviation	Definition
C4I	climate4impact
C4I WPS	Climate4impact Web Processing Service
EGI	European Grid Infrastructure
ENES	European Network for Earth System modelling
ENES CDI	European Network for Earth System modelling Climate Data Infrastructure
EOSC	European Open Science Cloud
ESGF	Earth System Grid Federation
EUDAT CDI	EUDAT Common Data Infrastructure

1. Introduction

1.1 Purpose and Scope

The deliverable objectives are to analyze and report about a first evaluation of the DARE developments related to the Climate Domain Pilot.

This deliverable will consist of a report detailing the architecture and the updated schema of the use case, as designed for this first prototype. There will be an emphasis on set of DARE components used. It will also report on the training and the feedback of the targeted users that attended the training. Since it is the first version of the report, as it is an early stage, it will not report yet on pilot execution benchmarks and performance aspects of the pilot. This will be done in the next versions of the deliverable. The evaluation and results will feed in the next stages, as this first evaluation is done on a first prototype.

1.2 Approach and relation to other Work Packages and Deliverables

Evaluation has also been done in WP6 for the Seismology Domain and reported in D6.3. Both Work Packages have adopted the same methodology for the training and similar structure in the deliverables. The targeted users are not identical for those two domains - WP7 evaluation focuses more on developers - but nonetheless the approach is similar.

1.3 Methodology and Structure of the Deliverable

The structure of this deliverable is as follows. First, a summary of the Climate Domain Use Case will be presented, then the technical implementation of the first prototype of this Use Case will be detailed. Evaluation results will be shown and discussed, followed by lessons learned that are feeding into the plan for the next development phases.

2. Climate Use Case First Prototype Implementation

The Use Case has been presented in detail in the DARE Description of Work. The whole detailed description will not be reproduced here, but a summary will be presented instead, to help better understand the evaluation.

2.1 Test Case Description

The objectives of this Use Case are to:

- Enable the delegation of C4I Platform Data Analytics and Processing to the DARE infrastructure. Typically, data reduction on the order of 90% can be achieved, depending on the users' analyses.
- Streamline and ease the whole data lifecycle.
- Definitely move away from a download-then-analyze type of workflow.

The Use Case also considers the following:

- Interoperability with EUDAT by using its standards and services
- Coordination with C3S and the future EOSC on interoperability
- Coordination with the IS-ENES Data Task Force

Generic Use-case Characteristics

- Objective: Generate a multi-model multi-scenario time series of the surface temperature for CMIP5 data

Expected Outcomes

Demonstrate an end-to-end solution based on the DARE platform for the heterogeneous base of the climate-change impact community end-users, dealing properly with the large amount of data needed to perform their research and applications.

This Use Case is generic enough in the sense that all interfaces and components that will have to be developed for this use case will also be useful for most of data analysis workflows currently needed in the climate science domain.

The first design sketch of the Use Case was as follows:

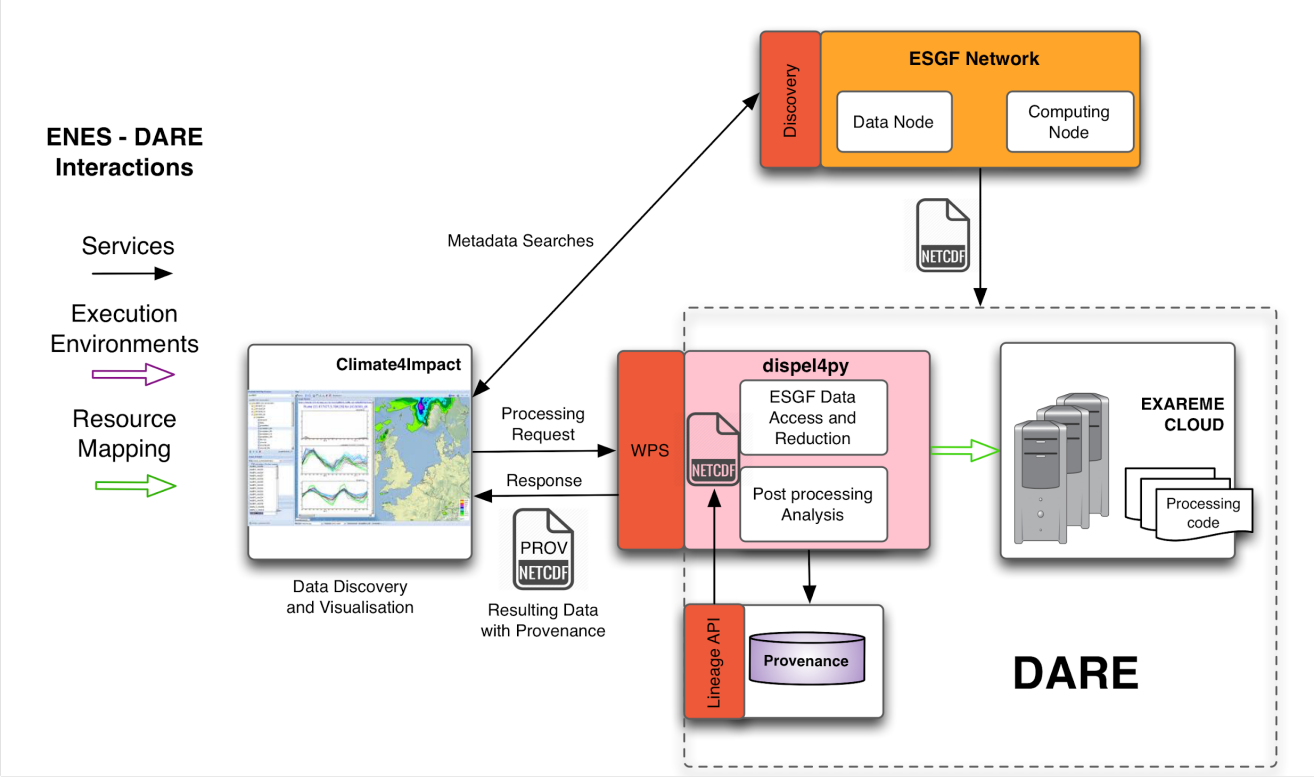


Figure 1: *Initial version of the DARE Climate Use Sketch*

The idea of this Use Case is to delegate calculations related to data analysis triggered by end users on the IS-ENES Climate4Impact (C4I) portal. Currently, the calculations take place on the portal front-end server. This approach is not scalable and has a negative impact on the performance of the front-end.
In this use case, the DARE components take care of the processing as well as input and output, seamlessly and transparently to the end users, also adding provenance and lineage information. The C4I platform queries the Climate Research Infrastructure ESGF and retrieve URLs of data files using metadata queries according to facets. Those URLs are then forwarded with the processing request to the DARE Platform. The deployment of those DARE components must be easy. It is intended to hide the complexity of underlying e-infrastructures and technologies to the software developer of platforms like C4I or data processing workflows.
The updated Use Case schema is now the following:

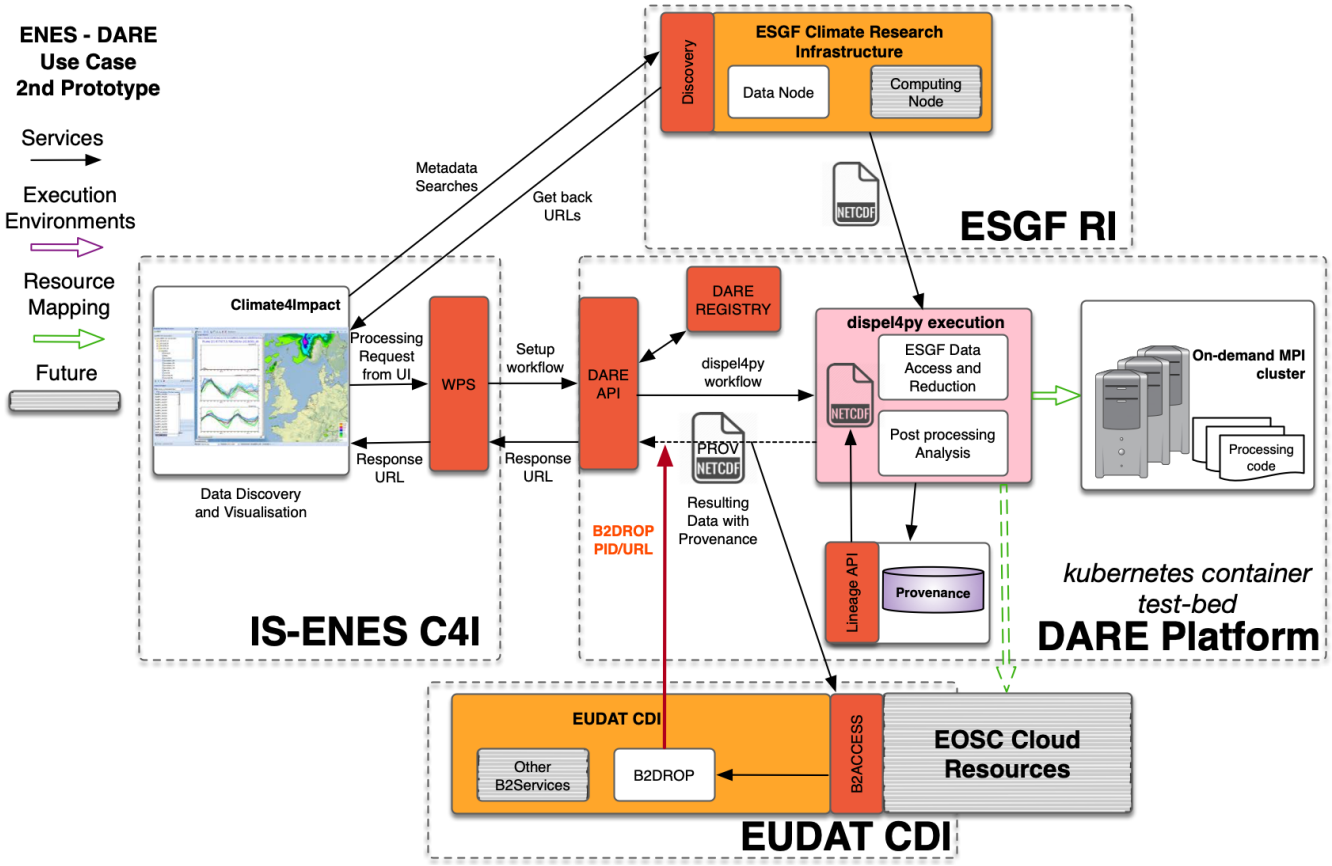


Figure 2: *Version 2 of the DARE Climate Use Sketch*

This updated version of the Use Case is the result of the development of the architecture and DARE components since the beginning of the project. It has evolved significantly from the first sketch and the use of external e-infrastructures is much more detailed and precise. It can be noted that EUDAT and EOSC are now explicitly a part of the use case sketch.

2.2 Overview of External Components Integration

In D7.1, a list of possible components of external architecture that can be used within this use case context, is defined:

- ESGF Climate Research Infrastructure (RI)
- EUDAT CDI Services: B2NOTE, B2SAFE, B2SHARE, B2DROP, B2HANDLE, B2FIND, B2STAGE, B2ACCESS, GEF
- IS-ENES CDI C4I
- EGI FedCloud

For authorization and authentication systems:

- EUDAT B2ACCESS
- ESGF OpenID
- EGI Certificates
- EOSC OpenID Connect

In the updated schematic presented in the previous section, we can see that the implementation is using the following components:

- IS-ENES CDI C4I
- EUDAT CDI Service B2DROP
- ESGF RI Data Nodes

- and of course DARE Components: API, Registry, dispel4py, lineage/provenance system
- No authentication system is yet in place in this first prototype, but placeholders are there.

2.3 Overview of DARE API Integration

The DARE components are deployed into the DARE kubernetes container testbed: testbed.project-dare.eu. This testbed is used for the development and evaluation phases.
The workflow of the Use Case implementation 1st prototype can be described as:

1. The end user access the specific WPS on the C4I front-end
2. The modifiable parameters are displayed
    a. Only textboxes input form is provided, and they are auto-generated from the WPS DescribeProcess
3. The user start the processing
4. The WPS is being executed
    a. Decode input parameters
    b. Setup the workflow in a dictionary representation
    c. Write a JSON file representing the workflow
    d. Get an auth token from the DARE API (empty/null username and password for this prototype)
        i. auth_token = F.login(REG_USERNAME, REG_PASSWORD, D4P_REGISTRY_HOSTNAME)
    e. Create a DARE workspace
        i. workspace_url, workspace_id, status = F.create_workspace("", WORKSPACE_NAME, "", creds)
    f. Create a Processing Element (PE) in the workspace
        i. pe_url = F.create_pe(desc="", name=PE_URL_NAME, conn=[], pckg=WORKSPACE_NAME.lower(), workspace=workspace_url, clone="", peimpls=[], creds=creds)
    g. Create the PE Implementation by sending the python dispel4py code that has functions to support the Use Case
        i. impl_id = F.create_peimpl(desc="", code=open(PWD+'/combineMultipleScenario.py').read(), parent_sig=pe_url, pckg=WORKSPACE_NAME.lower()+"_impl", name=PE_URL_NAME+"_impl", workspace=workspace_url, clone="", creds=creds)
    h. Upload the JSON workflow file to the workspace
        i. os.system('zip -r test_scenario_1.zip test_scenario_1.json')
        ii. resp = F.myfiles(token=F.auth(), creds=creds)
        iii. upload_path, exec_path = F.create_new_upload_path(json.loads(resp), UPLOAD_PATH)
        iv. F.upload(token=F.auth(), path=upload_path, local_path='test_scenario_1.zip', creds=creds)
    i. Submit the execution using python requirements
        i. F.submit_d4p(impl_id=impl_id, pckg=WORKSPACE_NAME.lower(), workspace_id=workspace_id, pe_name=PE_URL_NAME, token=F.auth(), creds=creds, reqs='https://gitlab.com/project-dare/WP7_IS-ENES_Climate4Impact/raw/master/requirements.txt', inputfile=exec_path, target='simple', n_nodes=1, no_processes=1)
    j. Send the output files to a EUDAT B2DROP account (in this prototype the credentials were hardcoded)

The python code being submitted and executed by dispel4py includes what is necessary to track provenance and lineage. It also includes all functions necessary for the calculations as well as doing the input/output using specific climate domain libraries.

2.4 Integration with the Climate Research Infrastructure

The integration with the Climate Research Infrastructure (RI) is not completed yet. The C4I input form for the execution parameters is still rudimentary. Also, the input files are not yet accessed through the ESGF Data Nodes but are rather prepared in advance and made accessible through the EUDAT B2DROP Service. Finally, the output file is not transferred yet back to the C4I user space, but is only uploaded to B2DROP.
But for the evaluation, since it is targeting the software developers of the climate RI and especially C4I, the focus for this first prototype was on the WPS itself with the integration of the DARE components. This is the main part software developers will have to deal directly with, when building the data processing services for the climate RI. This is where the DARE platform will be the most visible.
In this prototype, the WPS itself is executed within C4I, so the interface between the climate RI C4I and the DARE components take place within the WPS. This interface enables the deployment of the calculations away from the C4I server.

3. Training Event and Evaluation Results

The evaluation has taken place in the Netherlands and was organized by KNMI, on June 17th, 2019. The reader is encouraged to read D8.4 for further details on the organization and agenda of the training. Here, the focus will be on the results.

3.1 Training Attendees

The event was by invitation only, and some people involved in the DARE project were also present but did not fill the survey form, except for writing suggestions. There were also remote attendees. The total number of attendees was 16, with 4 people remotely (this list is also shown in D8.4):
Remote attendees

- From DARE: 1 software developer from DARE, and 1 from the seismology domain
- From IS-ENES: 2 software/infrastructure developers

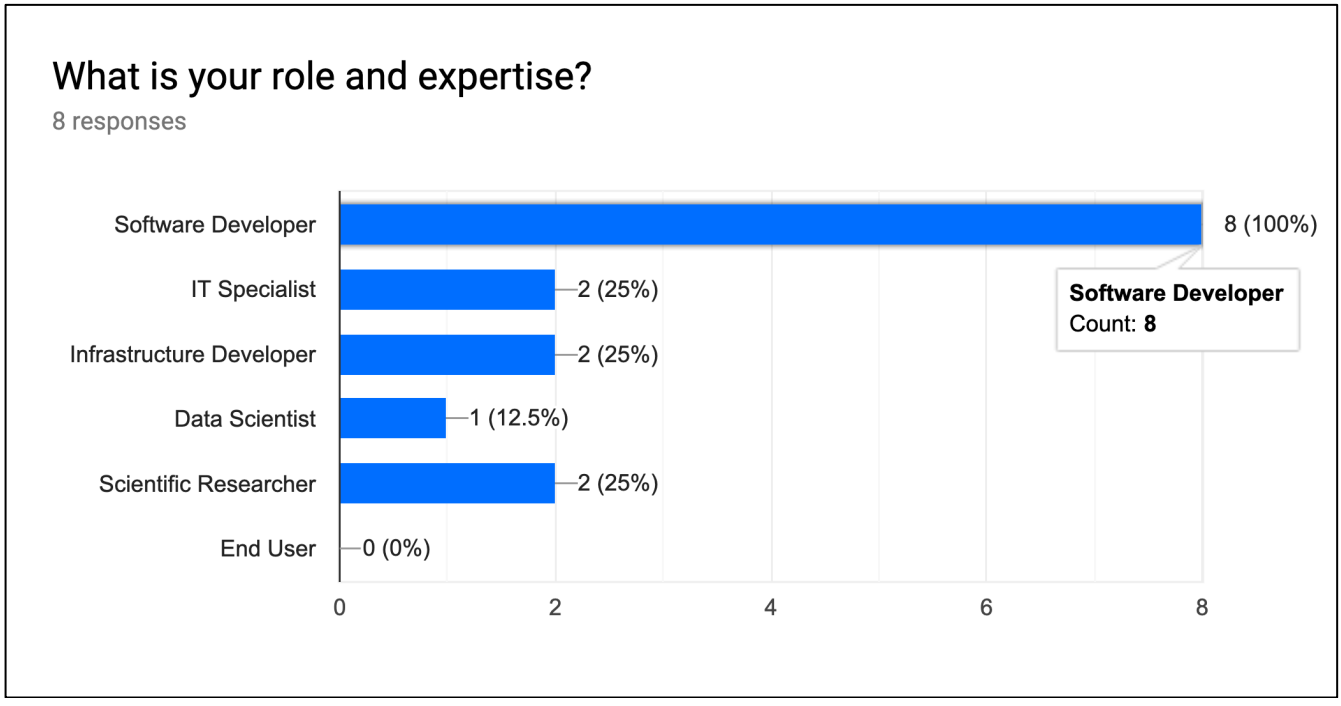
Attendees in room

- From DARE: 3 software developers
- From IS-ENES: 9 software/infrastructure developers

Figure 3: *Picture of the evaluation training, taken while participants are filling the survey form. Reproduced from D8.4.*

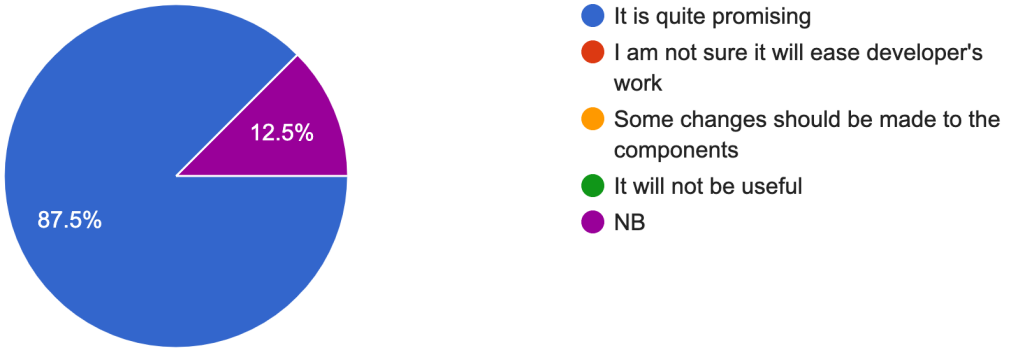
3.2 Evaluation Results

The survey form is accessible here: https://forms.gle/hyYUDj7P7Cfp2Mkc9 (take time to **open it** since in the output graphs some choices are ***truncated or not shown***). 8 people filled the form, with only 1 being from DARE (consider this 12.5% part in the results). The raw results are as follows:



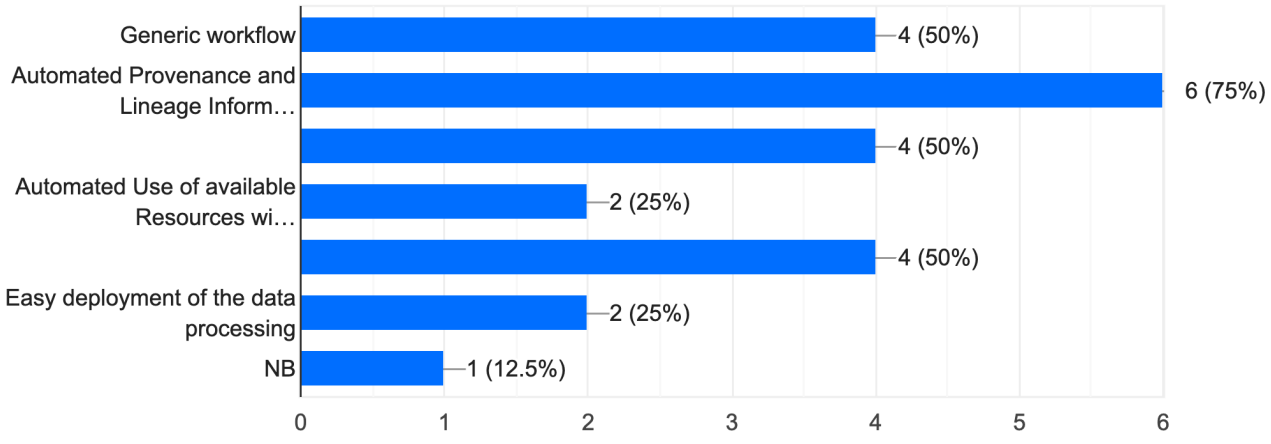
As a prototype, what do you think about the DARE Platform infrastructure approach

8 responses



What are the most useful features

8 responses



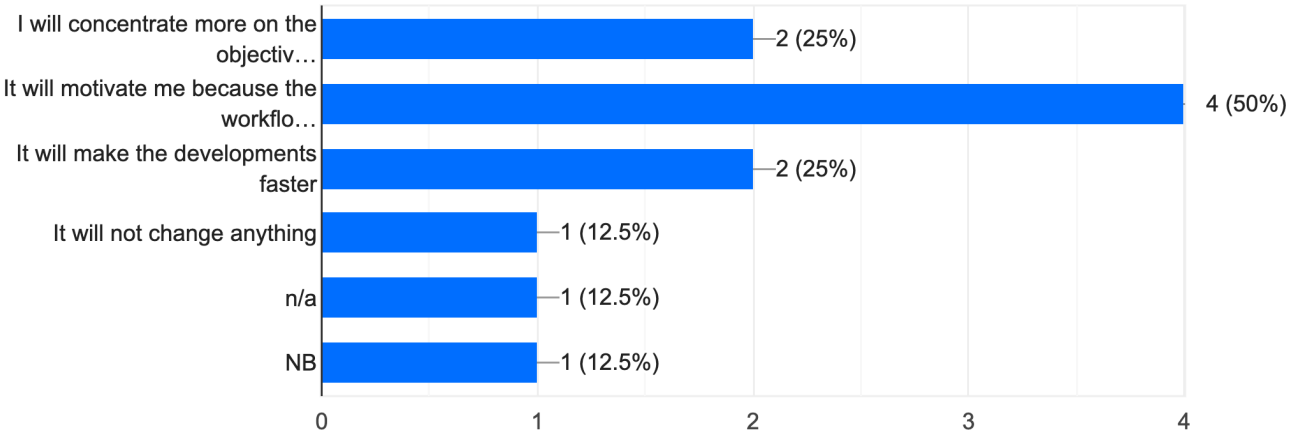
Which other features would you like to be available

8 responses

Abstract away the dependencies on dedicated infratstructures like EUDAT
OpenDAP usage,
Error tracking/management during the workflow execution
Parametrizable workflows from web user interface, ability to save output locally or into another backend instead into B2DROP
Link to dataset catalogs
Data Catalogue with also references to community vocabularies
don't know
cancel a workflow

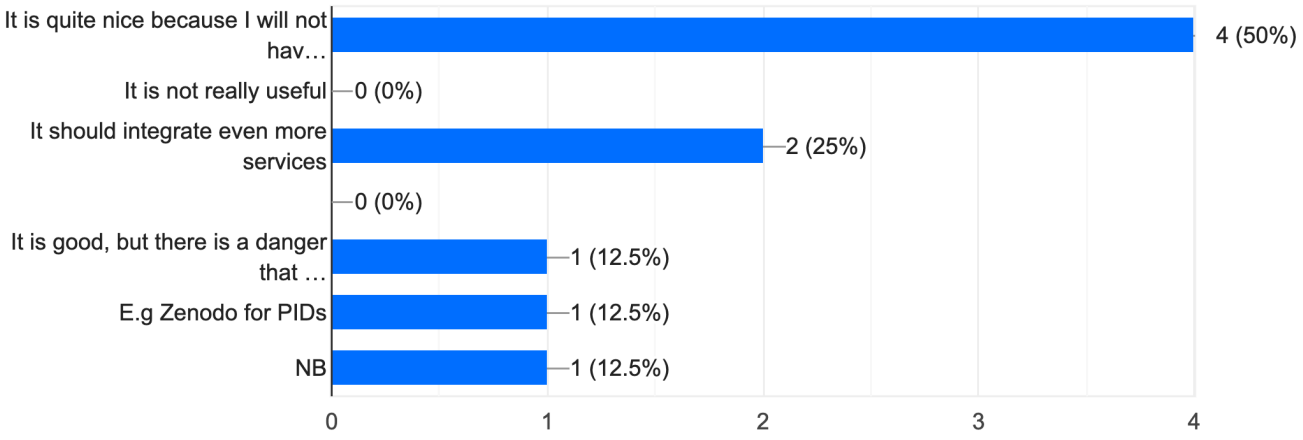
In what ways do you think the DARE Platform could change the way you develop workflows?

8 responses



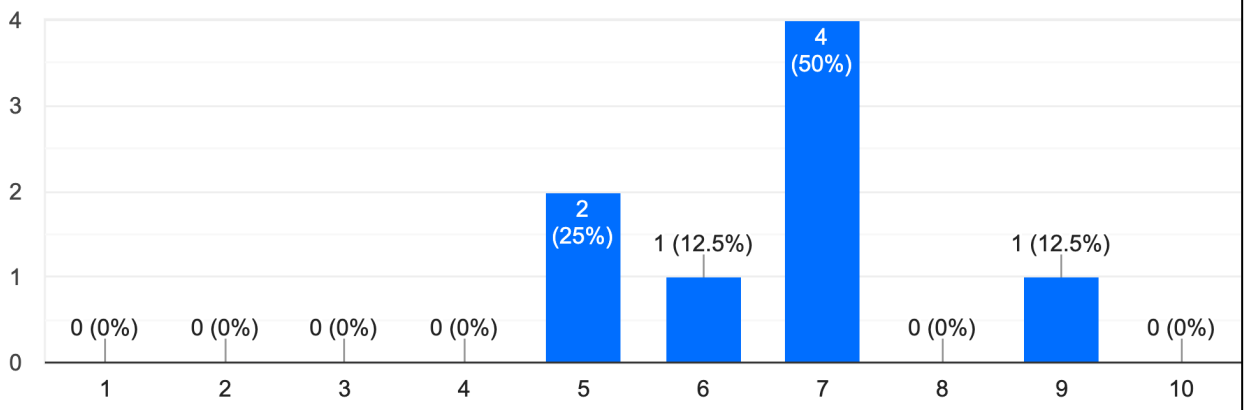
What do you think about the integration with available e-infrastructure Services (EUDAT, ESGF, ...)

8 responses



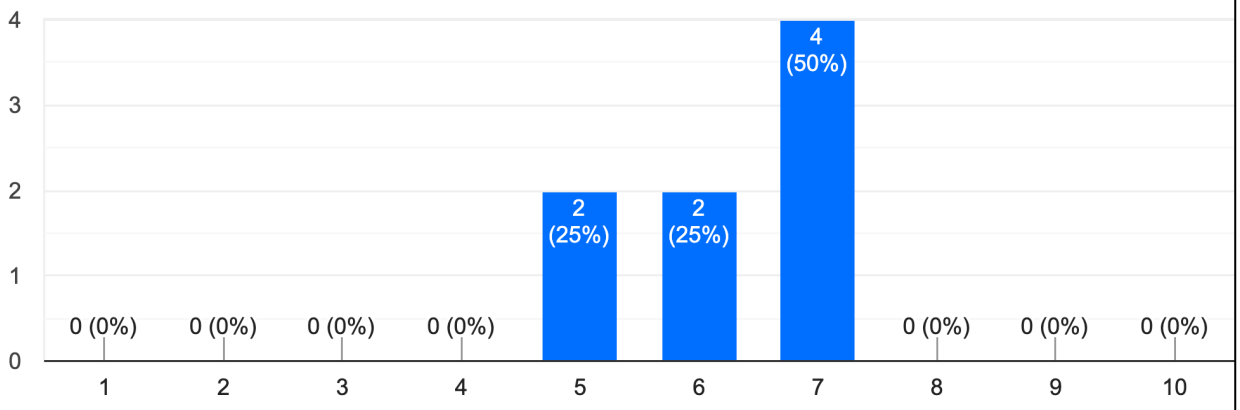
Is the DARE Platform duplicating similar efforts or it is unique and interesting?

8 responses



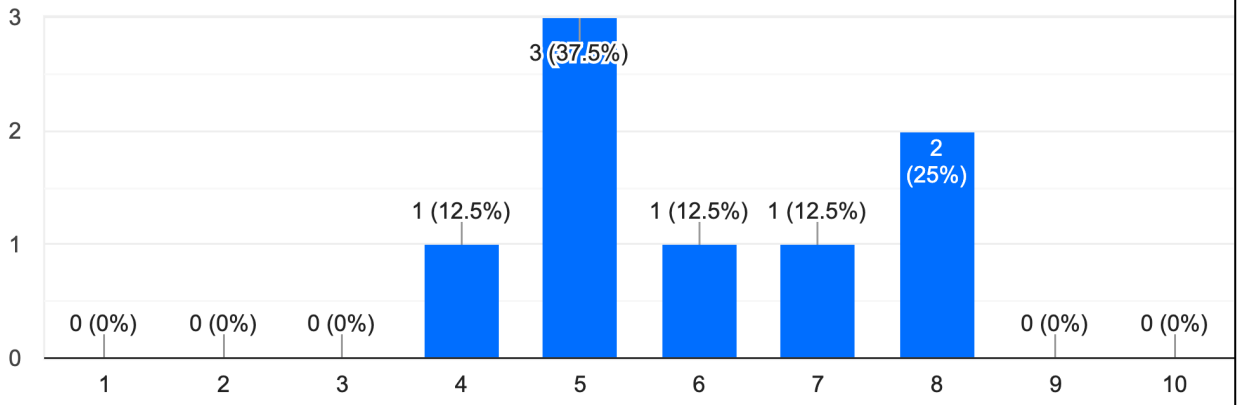
What is the usefulness of the DARE Platform based on this first Prototype

8 responses



How likely you consider using the DARE Platform when it will be operational?

8 responses



Please add your comments and suggestions. This is important to have as much information as possible to report!

8 responses

- I hope that Dare can be deployed by ourselves on our own managed cloud infrastructure or Amazon AWS.
- Would it be possible to cancel running workflows? Sometimes someone can start a too big unintend workflow.

I wonder when there will be a shift from data-files to data-as-a-service.
The current approach will be still shifting / moving around large climate-data between the storage-servers and processing units inside the Kubernetes cluster(s). Of course this is already much better than the end-users downloading the complete (climate) datasets.
Object-store (S3 on AWS) and NetCDF?
This current approach is not a really true native cloud solution.
We all dream about "processing close to the data". But reality is still shifting the data around...

It would be useful to have the possibility to restart a workflow from a specific point, not from the initial point, maybe in case of error.

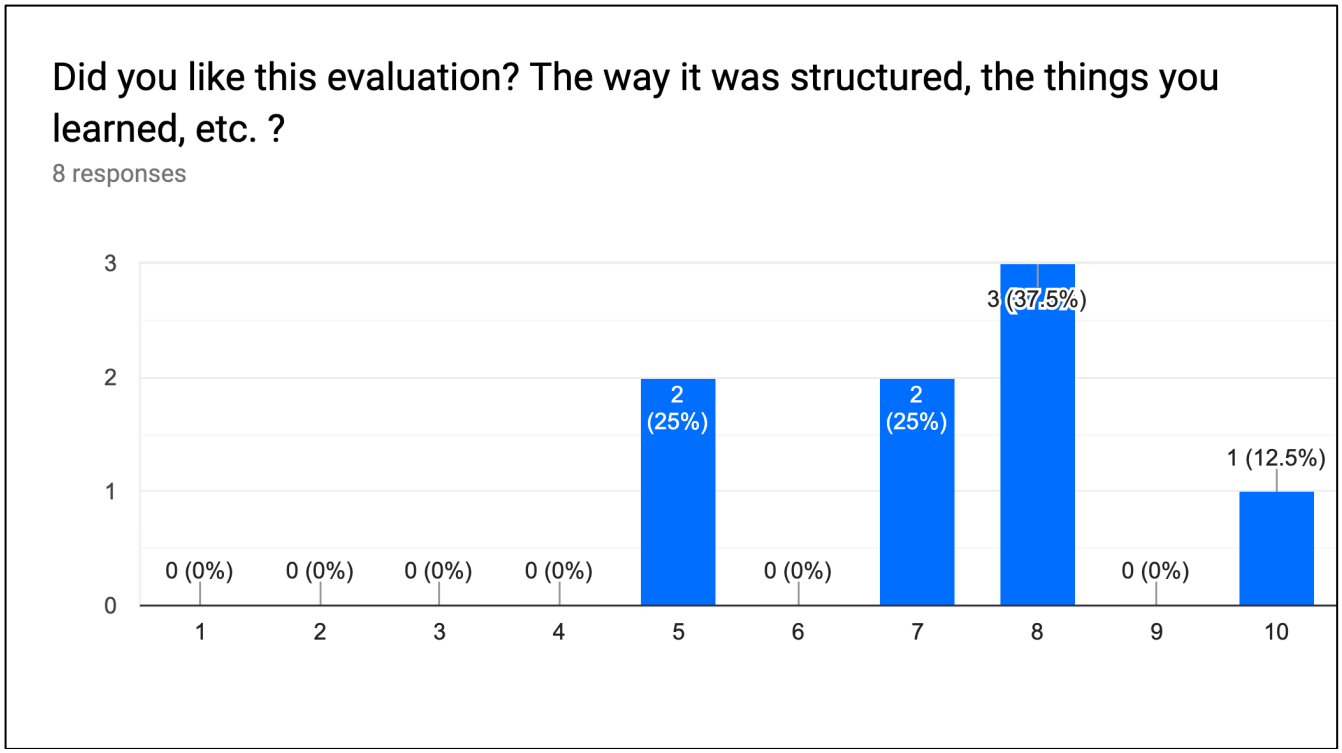
Metadata provenance is also being developed in other projects (metaclip.org), parallelism implemented in dispel4py maybe could be reused from Dask (similiar to what Pangeo is doing)

In the end more domains, not only climate and seismo, should be attached

Focus on which provenance and lineage metadata are required (reference values, quality test etc) and how to handle linkage from external datastore and catalogues. (PIDs). Improve the JSON description to be produced by a C4I user information. Support data streams in ICCLIM.

As a developer I'm mostly interested in the realization of DARE than in the usage.

I've just started on IS-ENES so I don't really know the requirements



For the analysis, each question will be discussed one by one, then an overall analysis will be done.

What is your role and expertise?

We can see in the results that the attendees were all software developers, which was the target. In addition, some were also IT specialists, infrastructure developers, data scientists, and scientific researchers. No end user were present, but this was on purpose.

As a prototype, what do you think about the DARE Platform infrastructure approach?

Except for one (a DARE representative), all of them responded that the DARE platform was quite promising. This is reassuring about our DARE approach.

What are the most useful features?

All of the listed features were seen as useful, but the most important one is the automated lineage and provenance features, along with generic workflows, API that hides underlying complexity as well as container-based technology. This is not a surprise since in the Climate Science domain, the lineage and provenance information is currently really lacking.

Which other features would you like to be available?

- Abstract away the dependencies on dedicated infrastructures like EUDAT.
  - This is what we want to achieve in the end, but the user will always have to input dedicated infrastructures' authorization credentials (token, etc.), unless they are homogenised within one e-infrastructure, for example EOSC.
- OpenDAP usage
  - This is already the case, so we should make it more explicit.
- Error tracking/management during the workflow execution
  - This is something we really need to have.

- Parametrizable workflows from web user interface, ability to save output locally or into another backend instead into B2DROP
    - This is on our development plan, except for the last point.
- Link to dataset catalogs
- Data Catalogue with also references to community vocabularies
    - Those two are linked. The data catalogue and community vocabularies should be on the front-end, so it is more an IS-ENES development than a DARE one.
- Cancel a workflow
    - Yes, definitely this is needed.

In what ways do you think the DARE Platform could change the way you develop workflows?
For 50% of the attendees, the most important aspect is that: it will motivate by the fact that workflows will be easily shareable. This is linked to reproducible science and the data life cycle, and this clearly shows that this is really a strong need in the community. At the moment there is no agreed workflow system used for data analysis, so it is not possible to share workflows among colleagues and researchers.

What do you think about the integration with available e-infrastructure Services (EUDAT, ESGF, ...)
50% said that it is quite nice because they will not have to develop interfaces to those e-infrastructures. This is exactly one of the main objectives of DARE. Also, nobody said that it was not really useful, which is reassuring. Two attendees also said that it should integrate even more services. This shows that this concept is quite interesting for developers.

Is the DARE Platform duplicating similar efforts or it is unique and interesting?
62.5% if the people gave 7/10 or more. This means that there is still work to do to be sure that DARE is filling a gap without duplication. The strong coordination with IS-ENES will help in this regard.

What is the usefulness of the DARE Platform based on this first Prototype?
50% of the people gave 7/10, while 25% gave 6/10 and 25% gave a 5/10. This is a bit surprising because it was more positive in the previous question and also in the first one. The reason could also be that the prototype is not advanced enough to show all the benefits. For example, we did not show yet how it would accelerate the processing for very large data volumes.

How likely you consider using the DARE Platform when it will be operational?
Mixed results, with 50% that gave 6/10 and more. After the evaluation has finished, two people would have already liked to begin developing using the DARE platform components, even if the platform is just at its prototype level. We hope that when we will show even more benefits the adoption rate will be higher.

Free suggestions and comments

- I wonder when there will be a shift from data-files to data-as-a-service. The current approach will be still shifting / moving around large climate-data between the storage-servers and processing units inside the Kubernetes cluster(s). Of course, this is already much better than the end-users downloading the complete (climate) datasets. Object-store (S3 on AWS) and NetCDF? This current approach is not a really true native cloud solution. We all dream about "processing close to the data". But reality is still shifting the data around…

    - Those are valid concerns, and some will be addressed. Calculations will be further delegated to the climate RI Computing Nodes as much as possible in the next prototypes, when the computing nodes will be ready. Those computing nodes will live very near the storage and the data nodes. For the object vs file approach, the DARE platform will adapt to the climate RI, where an object-like approach could be taken in the near future. These goals will also be facilitated by improving the data part of the DARE knowledgebase, planned to take place soon (D2.3).
- It would be useful to have the possibility to restart a workflow from a specific point, not from the initial point, maybe in case of error.
    - This is something DARE needs to support.
- Metadata provenance is also being developed in other projects (metaclip.org), parallelism implemented in dispel4py maybe could be reused from Dask (similar to what Pangeo is doing)
    - This must be evaluated in the next round of design and developments. It has been noted and it will be taken into account. However, Dask is already planned to be implemented in the data processing software that dispel4py will trigger.
- In the end more domains, not only climate and seismology, should be attached
    - Yes, definitely. This will be possible when we can show that for those two domains DARE has strong benefits. Dissemination will be important, as are plans for cross-domain webinars to be planned soon within the 2nd reporting period.
- Focus on which provenance and lineage metadata are required (reference values, quality test etc) and how to handle linkage from external datastore and catalogues. (PIDs). Improve the JSON description to be produced by a C4I user information. Support data streams in ICCLIM.
    - The provenance and lineage content will need to be taken care of during the next development phases to add something even more useful for the end users. The linkage to external datastore and catalogues (PIDs) will need to be done in the next development phases too. The JSON description of the workflow should probably be replaced by a CWL representation. For ICCLIM, we agree that it would be better to be able to support datastreams, but the related developments may take place in IS-ENES3.
- I hope that DARE can be deployed by ourselves on our own managed cloud infrastructure or Amazon AWS. Would it be possible to cancel running workflows? Sometimes someone can start a too big unintended workflow.
    - Cancelling workflows is definitely a requirement, which can be accommodated by the same mechanisms that handle pausing and error reporting.

Did you like this evaluation? The way it was structured, the things you learned, etc.?

75% of attendees gave 7/10 or more. The evaluation structure and organization was appropriate for the group of experts that took part. The aspects of the organization of the evaluation is analyzed in D8.4.

Overall, the results are quite positive, and it helps to plan for the next developments phases. It will be very interesting to compare those results and the feedback with the next evaluation that will take place later in the project.

4. Next Development Phases

At the beginning of the DARE project, the levels of development of the Use Case implementation were defined as (Level 1 in green, Level 2 in yellow, Level 3 in red):

Interface	Details	Themes
DARE (dispel4py) and ESGF Data Nodes	Enable the DARE Platform to download data from ESGF Data Nodes, with proper authentication	Workflow Authentication
DARE (dispel4py) and ENES CDI C4I	Enable C4I to trigger the execution of workflows using the DARE API, and retrieve the results along with provenance/lineage.	Workflow Provenance Authentication
PE Mapping Functions	Mapping of dispel4py PEs to calculation/processing functions (e.g. icclim python package). Add more provenance custom information.	Workflow Provenance
C4I GUI Wizard	Develop and Implement a Wizard on the C4I front-end to design and execute workflows	Workflow
DARE API and ESGF Computing Nodes	Delegate calculations to the ESGF Computing Nodes before accessing data on the Data Nodes	Workflow Provenance Authentication

With this evaluation feedback and suggestions, the following modified plan is proposed, with Level 1 (green) being already completed.

DARE API and B2DROP	Enable the DARE Platform to read input data or store results into EUDAT B2DROP Service	Workflow Authentication
DARE API and ENES CDI C4I	Enable C4I to trigger the execution of workflows using the DARE API, and retrieve the results along with provenance/lineage.	Workflow Provenance Authentication
PE Mapping Functions	Mapping of dispel4py PEs to calculation/processing functions (e.g. icclim python package). Add more	Workflow Provenance

	provenance custom information.	
DARE API and ESGF Data Nodes	Enable the DARE Platform to download data from ESGF Data Nodes, with proper authentication	Workflow Authentication
DARE API and C4I	Properly describe more complex workflows	Workflow
C4I icclim processing backend	Develop and implement a more efficient parallel processing (xarray & dask)	Workflow
Climate Domain Use Case	Make Use Case totally generic. Add more custom provenance and lineage.	Workflow Provenance
DARE API and ESGF Computing Nodes	Delegate calculations to the ESGF Computing Nodes before accessing data on the Data Nodes	Workflow Provenance Authentication
C4I GUI Wizard	Develop and Implement a Wizard on the C4I front-end to design and execute workflows	Workflow
C4I and DARE API	Implement proper workflow cancellation, restart and error tracking management	Workflow

This new development plan for the next phases is taking place from July 1st, 2019, and will be discussed among the architecture task force.

5. General Conclusions

The Climate Science Domain Pilot first prototype has been evaluated in a special training session that took place in Utrecht on June 17th, 2019. 16 attendees were present, including remote participation. The targeted users are software developers of the Climate Research Infrastructure, and all of the participants had this expertise. So, it was a success to be able to gather together a significant amount of people within the targeted expertise.

The overall feedback is quite positive, and the evaluation helped to identify better what should be the focus for the next development phases. One of the conclusions is that the interest is very high as DARE is really filling a needed gap in the infrastructures' services. If suggestions and feedback are taken into account with proper dissemination and a future second evaluation, the DARE platform will be fully integrated and use within the Climate Research Infrastructure. This will also help to disseminate to other scientific domains that have similar requirements.