8.3 Marine (CC)

- Ambition
- User stories
  - Argo user stories - introduction
  - SeaDataNet user stories - introduction
- Use cases
- Architecture & EOSC-hub technologies considered /assessed
  - Argo use cases
- SeaDataNet use cases
- Requirements for EOSC-hub
  - Technical Requirements
  - Capacity Requirements
    - Argo use cases:
- Validation plan

Short description	Marine CC (Task 8.3)
Type of community	Competence Centre
Community contact	Thierry Carval (Argo) Dick Schaap (SeaDataNet)

Ambition

The ocean experts are now converging in the estimation of integrated indicators such as global warming. However these indicators, based on interpolation of unevenly distributed observations, do not describe consistently the climate change. To better understand the ocean circulation and climate machinery, data scientists need to directly access the original observations otherwise diluted in spatial synthesis.

Original observations are published by Research Infrastructures (Argo, EMSO, ICOS…) and data aggregators (SeaDataNet, Copernicus Marine,…).

The Marine Competence Centre long term ambition is to push Ocean observations on EOSC infrastructure for data analytics. The work in the CC focuses on two areas:

1. Making Argo data more easily accessible for subsetting and online processing. IFREMER and its partners work on this area.
2. Simplifying/harmonising the access to data that reside at SeaDataNet partners from cloud-based applications. MARIS and its partners work on this area.

User stories

Instruction

Requirements are based on a user story, which is is an informal, natural language description of one or more features of a software system. User stories are often written from the perspective of an end user or user of a system. Depending on the community, user stories may be written by various stakeholders including clients, users, managers or development team members. They facilitate sensemaking and communication, that is, they help software teams organize their understanding of the system and its context. Please do not confuse user story with system requirements. A user story is an informal description of a feature; a requirement is a formal description of need (See section later).

User stories may follow one of several formats or templates. The most common would be:

"As a <role>, I want <capability> so that <receive benefit>"

"In order to <receive benefit> as a <role>, I want <goal/desire>"

"As <persona>, I want <what?> so that <why?>" where a persona is a fictional stakeholder (e. g. user). A persona may include a name, picture; characteristics, behaviours, attitudes, and a goal which the product should help them achieve.

Example:

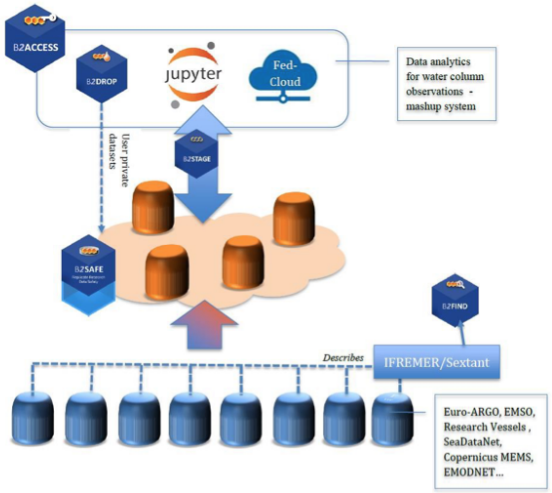
“As provider of the Climate gateway I want to empower researchers from academia to interact with datasets stored in the Climate Catalogue, and bring their own applications to analyse this data on remote cloud servers offered via EGI.”

Argo user stories - introduction

The Marine community produces diverse types of data (typically time-series data). They wish to store those data in files and make these files easily browsable and accessible by researchers. To maximise ease of use the files should be made available to users via a Dropbox-like system that makes relevant data files visible for each user in his/her ‘personal folder’. The users should be able to define patterns that define what kind of data they are interested in (location, time period, provider network, etc.) and the system should perform pattern matching to decide whether or not to make a particular incoming file (or set of files) visible for a given user. Such pattern matching can be CPU-intensive when we scale up to many users, many files files with complex data records. Depending on the community the source of data can be a single instrument (site), or can be multiple collection/production sites. In the latter case the data originating from multiple locations should be brought onto common formats and must be described with metadata in a coherent fashion.

The Argo activity of the Marine CC is testing (See Figure below)

- a combination of B2Find, B2Safe and B2Stage for the data management part (storage and transfer)
- a Jupyter, B2Access, EGI Cloud combination for user exposure. (data subscription and access)



No.	User stories
Argo user stories	
US1	A data provider should be able to link its data production instruments into the 'back-end' of the Marine CC setup and become a data provider for the CC users.
US2	A scientists should be able to browse the connected data source networks (e.g. Argo, EMSO, SeaDataNet, etc.) and define preferences for the data records he/she is interested in. The system should make matching records visible in his/her personal access folder.
US3	A user should be able to access his/her personal data access folder via a Jupyter system and perform data analytics on the data.

SeaDataNet user stories - introduction

The current workflows of the SeaDataNet are based on a pre-cloud architecture. Many operations happen asynchronously and in batch mode. In order to better serve the Marine community, we want to provide fast and scalable access to the datasets. To improve the current workflow users should be able to take advantage of the improved access and availability of the cloud. A user should be able to store their data on a Dropbox-like environment, However, still, be able to process and analyze them using both legacy/desktop software created during previous SeaDataNet projects and new cloud-based computing services. Furthermore, users should be able to discover data relevant to their needs using (semi) real-time discovery tools. Instead of preparing datasets for download for each user request, Data providers should be able to have their data stored on the cloud and provide access to users that have been granted permission. A data provider should be able to fix partial errors within a dataset during import, without having to re-upload the complete dataset.

No.	User stories
SeaDataNet user stories	
US4	As a user, I want to be able to use my legacy/desktop software to process and analyze data stored on the cloud.
US5	As a data provider, I want to only have to update erroneous files during import to only transfer the data required once.
US6	As a user, I want to be able to access my requested data through cloud computing tools within my cloud environment.
US7	As a user, I want to be able to find relevant datasets available within the cloud environment in (semi) real-time.

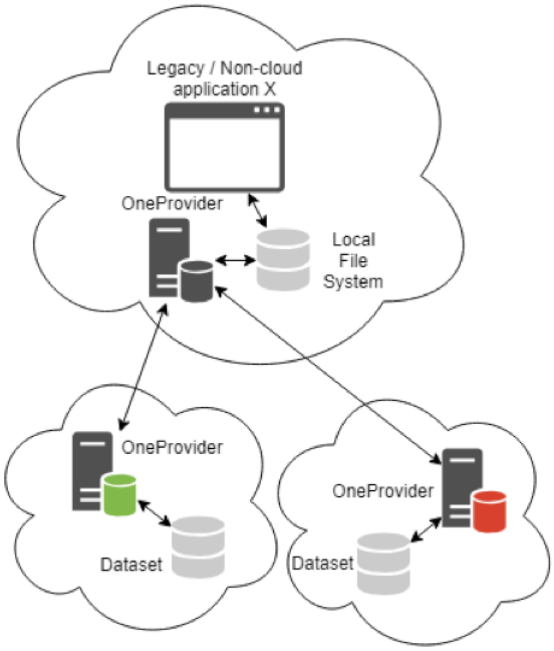
Use cases

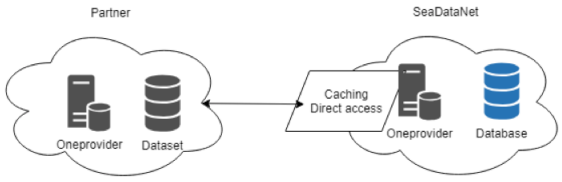
Instruction

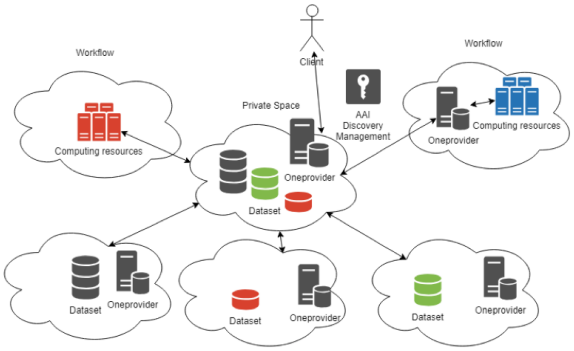
A use case is a list of actions or event steps typically defining the interactions between a role (known in the Unified Modeling Language as an actor) and a system to achieve a goal.

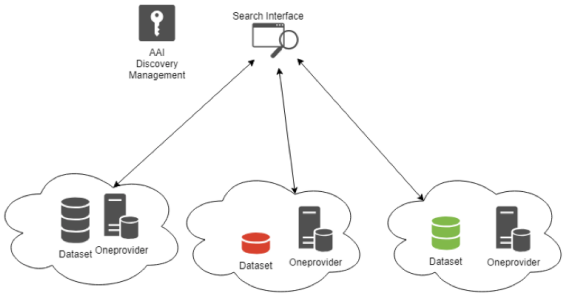
Include in this section any diagrams that could facilitate the understanding of the use cases and their relationships.

Step	Description of action	Dependency on 3rd party services (EOSC-hub or other)
Argo use cases		
UC1	• Data discovery and subsetting-subscription service on Argo observations.	
UC2	• DIVA data-interpolating variational analysis on Argo floats oxygen data, running on a Jupyter notebook.	

UC3	• Data scientist manages his workspace within JupyterHub : save and share notebooks, run codes on the datasets pushed by Resarch Infrastructures on EOSC (such as Argo) and his individual datasets.	
SeaDataNet use cases		
UC4	Cloud migration for legacy applications: Many software applications used by the Marine Community were developed during multiple projects spanning many years. These applications have specific requirements regarding File system operations. The most common assumption is that files are available on local storage. To simplify the processes of migrating these (mostly desktop) applications to the cloud we would use Onedata as a file access layer providing seamless access to files distributed in the cloud environments.	

UC5	Reducing redundant data transfer: Processing datasets from partners require that the files first be transferred to a staging area and then processed. However, these operations are not always successful. In the case of quality control, certain datasets may be rejected and have to be revised before being submitted again. This means that some complete datasets are transferred multiple times before being accepted. Using the direct access provided by Onedata to these files we can process only the required amount of data.	

UC6	Virtual space for user data storage and delivery: After a user searches for specific data he sends a request for a subset of datasets, a process is started to collect the datasets from many partners. This collection is an asynchronous process. The process can take weeks to collect all the requested files from the partners. This process is dependent on the resources available at the partners. We intend to use the features of Space and privileges management provided by Onedata to streamline these processes. We would provide the end users access to his requested files through a shared space. Ideally, such a space can be used to make his requested files available for further processing in a cloud environment.	

UC7	Interface for distributed search using metadata queries: In order to find specific files in a distributed environment, we use proprietary search indexes. These indexes are inflexible and only allow querying of predetermined fields. This increases the time required to process and index all available datasets. With the current search interface, we process changes daily. However, the use of datasets within workflows would benefit from up to date information on the available datasets. To extend the discovery capabilities of a cloud application we can leverage the advanced metadata querying functionalities of the Onedata platform.	

Architecture & EOSC-hub technologies considered /assessed

Argo use cases

B2SAFE: synchronize every day Argo data from Ifremer to B2SAFE

B2DROP: as an input for data scientists individual datasets

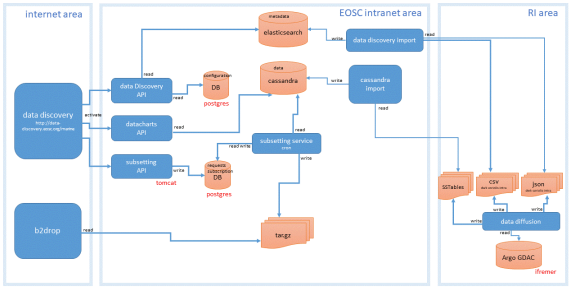
B2ACCESS: the user (data scientist) identification service

JupyterHub: the data analytics platform on datasets (Example: DIVA analysis on a Jupyter Notebook reading Argo data)

Data subscription web GUI and API to query

- Cassandra: the nosql data base for high performance query on data
- Elasticsearch: the for high performance queries on metadata

EOSC hub marine data discovery and subsetting GUI



SeaDataNet use cases

The use cases consider to evaluate the EGI DataHub service (OneData technology) in the above presented 4 architectural scenario.

Requirements for EOSC-hub

Technical Requirements

Requirement ID	EOSC-hub service	GAP (Yes/No) + description	Link to requirement ticket	Source Use Case
RQ1			EOSCWP1 0-41 - Jira .	
RQ2			EOSCWP1 0-77 - Jira .	

Capacity Requirements

Argo use cases:

EOSC-hub services	Amount of requested resources	Time period	Link to requirement ticket
B2SAFE	100go for Argo data	From 2018 (OK)	
B2DROP	EOSC hub data scientist user default account on B2DROP		
B2ACCESS	100 users should be able to access the services	From 2020	
JupyterHub	EOSC hub data scientist default account on Jupyter Hub	From 2018 with Cineca for DIVA analysis	

Host the data subscription web GUI with its Cassandra and Elasticsearch databases	Ongoing request for capacity with IN2P3 Alternative possibilities : CSC or Cineca	From mid 2019	

SeaDataNet use cases:

EOSC-hub services	Amount of requested resources	Time period	Link to requirement ticket
EGI DataHub	The 4 SeaDataNet use cases can be run on a testbed consisting of 3 sites: 2 as data providers, 1 as cloud compute provider. The sites should scale to the following level: The average number of files requested and processed is 500. Each file has a size of tens of KB but occasionally some larger files that average 500 MB are processed. A request is usually for ease of transfer and is usually between 50 to 100 MB.	From 2018 (OK)	EOSC WP 10-77 - Jira .

Validation plan

Not yet defined.