D6.5 Final report on the maintenance and integration of common services

Lead Partner:	MPG
Version:	1
Status:	Under EC review
Dissemination Level:	Public
Document Link:	https://documents.egi.eu/public/ShowDocument?docid=3640

Deliverable Abstract

This document contains a combined report on the maintenance and integration of common services of the EOSC Hub Service Catalogue[1] for the final project year. The work described within this document; focusing on service provisioning, maintenance and integration; contributes to the EOSC-hub Key Exploitable Results 4,5 and 8.

[1] For detailed description of the common services see the report of deliverable D6.1 First release of common services software (https://confluence.egi.eu/display/EOSC/D6.1%09First+release+of+common+services+software)

EOSC-HUB RECEIVES FUNDING FROM THE EUROPEAN UNION'S HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO. 777536.

COPYRIGHT NOTICE

DELIVERY SLIP

Date	*Name*	*Partner/Activity*	*Date*
From:	John Alan Kennedy	MPCDF/WP6	23/03/21
Moderated by:	Sjomara Specht	EGI Foundation/WP1	
Reviewed by:	Shaun de Witt Pavel Weber	UKAEA KIT	18/02/21 28/02/21
Approved by:	AMB		

DOCUMENT LOG

Issue	*Date*	*Comment*	*Author*
v.0.1	23/09/2020	Initial draft	Abdulrahman Azab (SIGMA2 - UiO), Miguel Caballer (UPV), Michele Carpene (CINECA), Enol Fernández (EGI), John Alan Kennedy (MPCDF), Germán Moltó (UPV), Olivier Rouchon (CINES), Heinrich Widmann (DKRZ)
v.0.2	26/01/2021	Reviewers comments added from 1st review phase	John Kennedy (MPCDF) Pavel Weber (KIT) Shaun de Witt (UKAEA)
v.0.3		Deliverable completed for 2nd phase of review	John Kennedy (MPCDF)
v.0.4	01/03/2021	2nd Phase Review completed	John Kennedy (MPCDF) Pavel Weber (KIT) Shaun de Witt (UKAEA)
v.1	23/03/2021	Final document	John Kennedy (MPCDF)

TERMINOLOGY

https://wiki.eosc-hub.eu/display/EOSC/EOSC-hub+Glossary

Terminology/Acronym	***Definition***
API	Application Programming Interface
CKAN	Comprehensive Knowledge Archive Network
CMD	Cloud Middleware Distribution
DOI	Digital Object Identifier
IaaS	Infrastructure as a Service
IdP	Identity Provider
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
PaaS	Platform as a Service
CaaS	Computing as a Service
PAM	Pluggable Authentication Module
PID	Persistent Identifier
RCD	Research Community Dashboard
REST	REpresentational State Transfer
SAML	Security Assertion Markup Language
VM	Virtual Machine
VO	Virtual Organization
WebDAV	Web Distributed Authoring and Versioning
WMS	Workload Management System
YAML	Yet Another Markup Language

Contents

Executive summary

This document contains a combined report on the maintenance and integration of common services of the EOSC-hub Service Catalogue[2] for the final project year. The work described within this document; focusing on service provisioning, maintenance and integration; contributes to the EOSC-hub Key Exploitable Results 4 (Internal Services in the Hub Portfolio), 5 (External Services in the EOSC Service Portfolio) and 8 (Interoperability and integration guidelines).

This deliverable covers the services in 6 categories, reflecting the different thematic areas of WP6.

- **Discovery and access:** The services in this category enable consistent interfaces to services in the area of the discovery and the access of digital resources. It will create a common access and discovery layer for depositing, exchanging, retrieving and staging data and metadata.
    - ***Services Provided:*** EGI Datahub, B2FIND, B2STAGE, B2DROP
- **Federated Compute:** Focuses on the corrective, adaptive and perfective maintenance of the EGI Cloud Compute, Cloud Container and High-Throughput Compute services. It aims to satisfy the emerging needs from the user communities and to support the most recent technology evolution.
    - ***Services Provided:*** EGI Compute Cloud, EGI Cloud Container, EGI Workload Management, EGI Online Storage, Advanced IaaS, EGI High Throughput Compute
- **Processing and Orchestration:** Focuses on the maintenance and integration of orchestration services with the Cloud Compute and Cloud Container services. This allows to build complex virtual infrastructures based on standards.
    - ***Services Provided:*** TOSCA for HEAT, Infrastructure Manager, PaaS Orchestrator, Future Gateway
- **Data and Metadata Management**: Integration and maintenance of the EOSC-hub common repository services and policy-driven data management/stewardship services with particular regard to registered data.
    - ***Services Provided:*** B2HANDLE, B2SAFE, B2SHARE, B2NOTE
- **Preservation:** The integration of certified Trusted Digital Repository (TDR) in the catalogue, resulting in a sustainable long-term data preservation service: the European Trusted Digital Repository (ETDR).
    - ***Services Provided:*** eTDR
- **Sensitive Data:** Focusing on integration activities for services with data whose access is restricted either by national or European regulations or by other confidentiality policies (e.g. business confidentiality).
    - ***Services Provided:*** TSD, ePouta

[2] For detailed description of the common services see the report of deliverable D6.1 First release of common services software
(https://confluence.egi.eu/display/EOSC/D6.1%09First+release+of+common+services+software)

WP6 followed a use-case driven workflow both with respect to individual services which solve core specific use-cases and the integration activities where multi-service compound use-cases were addressed. Seven community defined integration projects were defined to highlight integration activities and drive development. These have been addressed leading to the integration of numerous services to provide high level solutions. The direct outcomes of the WP6 activities are a set of common core services which can be used off the shelf by European research communities as well as numerous multi-service integration solutions that solve high level compound use-cases and pave the way for future integration activities.

1 Introduction

This document presents a global summary of the WP6 activities during the final period of the EOSC-hub project. The activities highlighted here build upon and extend the previous activities which are reported in deliverables D6.1-D6.4[3]. The deliverable comprises two major sections, section 3 which reports on the status of the common services which have been maintained and developed throughout the EOSC-hub project and section 4 which details the use-case driven integration activities. This deliverable combines two previous strands of deliverables and reports on activities which have been undertaken since the release of deliverables D6.3 and D6.4. Section 3 reports on the common service activities undertaken since D6.3 while section 4 reports on integration activities since D6.4.

The core services in section 3 are grouped into six thematic areas covering 'Discovery and Access', 'Federated Compute', 'Processing and Orchestration', 'Data and Metadata Management', 'Long Term Data Preservation' and finally 'Sensitive Data Services'. For each of these thematic areas the individual services are introduced, providing information about the service, integration opportunities, and future plans.

The integration activities described in Section 4 detail the global integration activities, starting with a description of the use-cases which were used to define these activities and then providing a detailed description of the integrations undertaken per thematic area.

The document concludes with a Summary section.

[3] https://documents.egi.eu/document/3414, https://documents.egi.eu/document/3480, https://documents.egi.eu/document/3558 , https://documents.egi.eu/document/3643

2 Core Services

The core services are grouped into six thematic areas which are detailed in Sections 3.1-3.6. For each service information about the development and maintenance activities undertaken since the publishing of deliverable D6.3 is given in the service release notes. Additionally, integration opportunities, and future plans are highlighted.

2.1 Data Discovery and Access

The overall objective of the task T6.1 'Data Discovery and Access' is the establishment of the Common Discovery and Access Interoperability Layer through which end-users can find, localize, transfer and re-use data resources within EOSC-hub for their own scientific purposes.

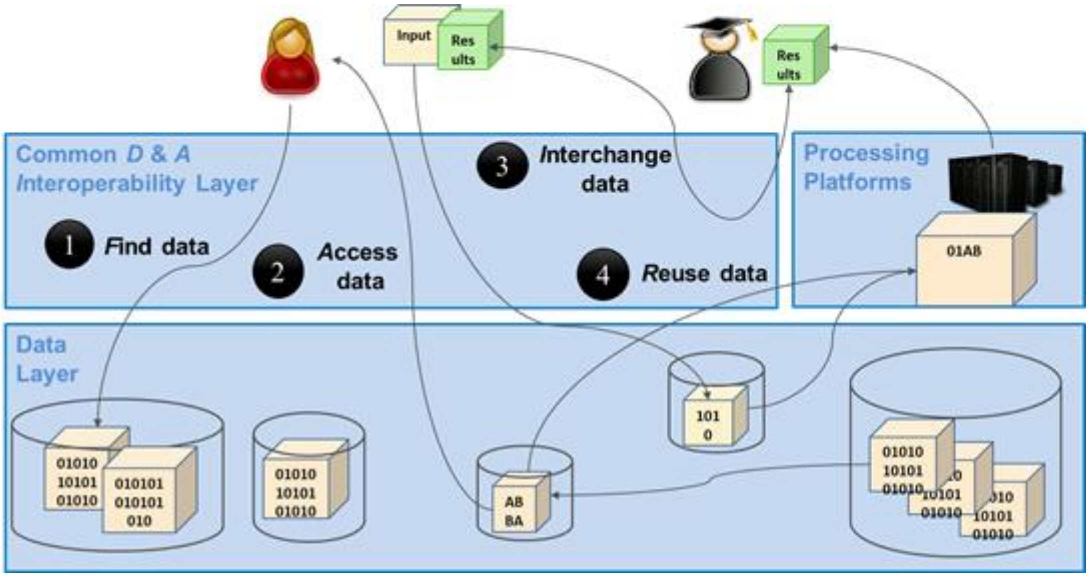


Figure 1. The Common Discovery & Access Interoperability Layer enabling FAIR data management

As shown in figure 1 above, end-users should be able to manage and use research data in a FAIR way, which means in particular:

1. [F] search for distributed data in EOSC-hub and beyond
2. [A] retrievability of metadata and data by open protocols
3. [I] interoperable sharing and publishing of research output following open standards and domain agnostic tools and interfaces
4. [R] reuse, exchange and staging of data

Regarding data discovery, the metadata service B2FIND is intended to play the role of the central search indexing tool of EOSC-hub. For this the service is extended and enhanced to cover data from EOSC-hub storage services

To allow seamless data access, easy data transfer between EOSC hub storage repositories should be investigated and established. B2DROP serves to easily exchange data with other researchers and to keep it synchronized and up to date. Within the last project period the main effort is put on the upgrade of the underlying software, Nextcloud, and the integration with B2SAFE and Datahub. B2SAFE based storage was successfully added as external storage to B2DROP via WebDAV. This procedure has been documented and can be deployed for communities on request. However, issues regarding the authentication token lifetime and network load during data transfers need to be considered, on a case-by-case basis, for production deployments.

The integration of Datahub has been halted due to differences in the protocols for the data exchange.

The concrete realisation and implementation of the Common Discovery and Access Interoperability Layer followed a roadmap comprising workpage activities (WPAs) describing the integration of pairs of common services within T6.1. but also, with partner services from T6.4 and by using AAI tools of WP5:

- WPA 6.1.3 & 4: Integration of EGI Datahub with B2FIND (indexing and discovery of EGI data resources)
- WPA 6.1.5: Get B2SAFE data collections indexed by B2FIND (see Herbadrop use case in section 4.1.3. of this report)
- WPA 6.1.6: Interoperability between EGI datahub and B2SAFE (data transfer between storage media of both services))
- WPA 6.1.7: B2STAGE integration with B2SHARE (retrieve processed data and store in B2SHARE)
- WPA 6.1.8: B2STAGE integration with B2DROP (prepare input data for B2STAGE / retrieve and store - small sized - data)
- WPA 6.1.9: B2DROP integration with EGI datahub

To achieve this, an integration plan was defined, in the initial phase of the project, which identifies and specifies the work plan activities WPA 6.1.N. An overview of this integration plan is shown in the figure 2.

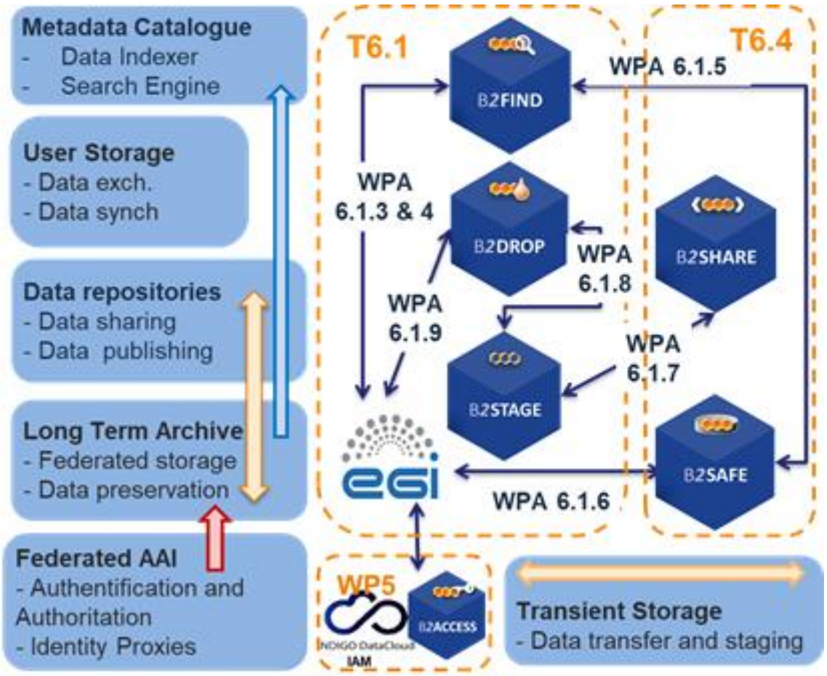


Figure 2. The updated integration architecture of T6.1 - including services from T6.4 and AAI tools from WP5

The figure includes not only the integration of common services from T6.1, but also the storage services B2SHARE and B2SAFE from T6.4 ‘Data and Metadata management’ and the AAI tools from WP5. The latter are important for seamless authentication and authorisation for service and data access.

There are two additional WPAs that go beyond the common services of WP6 and are related to the joint activities milestones JAM.1.9 and JAM.1.11 within the cooperation between EOSC and OpenAIRE (see https://wiki.eosc-hub.eu/display/EoscHubOpenAIRE/JA1+Service+Integration):

- WPA 6.1.11: Integrate EGI DataHub with OpenAIRE Research Community Dashboard by adapting to OpenAIRE guidelines
- WPA 6.1.12: Integrate EUDAT B2FIND/B2SHARE with OpenAIRE Research Community Dashboard by adapting to OpenAIRE guidelines for data providers

In the following subsections a detailed description of each service in the thematic area ‘Data Discovery and Access’ is provided including a description, detailed release notes, integration activities and future plans. Finally, the achieved results are summarised in the ‘Summary’ section.

2.1.1 EGI DATAHUB

2.1.1.1 Service description

Service/Tool name	EGI DataHub
Service/Tool url	https://datahub.egi.eu
Service/Tool information page	https://onedata.org
Description	Onedata is a globally distributed storage solution, integrating storage services from various providers using possibly heterogeneous underlying technologies, such as NFS or other POSIX-compliant filesystems as well as Ceph, S3, GlusterFS, WebDAV and OpenStack SWIFT and provides to client's interfaces based on CDMI, REST API and virtually mounted POSIX filesystem. Onedata architecture consists of two major components: Oneprovider and Oneclient. The former is installed within a data center and provides a unified interface to multiple filesystems used in the center. Servers can scale to thousands of instances in order to improve performance. The main functional components include: • Onezone - the federation and authentication service, each Onezone instance (e.g., EGI DataHub) provides a single-sign on to a network of connected storage providers. • Oneprovider - is the main data management component of Onedata, which is deployed in the data centers and is responsible for provisioning the data and managing transfers. • Oneclient - provides access to the virtual filesystem on a VM or host directly via a Fuse mountpoint.
Value proposition	• Discovery of data via a central portal. This will include a search mechanism plus a rating system which may be based on, e.g., number of accesses, based on ElasticSearch engine which indexes the data and metadata managed by the service, with search and discovery user interface integrated into the Onezone service, • Access to data conforming to required policies which may be: 1) unauthenticated open access; 2) access after user registration or 3) access restricted to members of a Virtual Organization (VO). This access may be via a GUI (e.g., a webpage) or an API (e.g., programmatic access to the data) • Replication of data from data providers for resiliency and availability purposes. Replication may take place either on--demand or automatically. Replication will

	require the introduction of a file catalogue to enable tracking of logical and physical copies of data. • Access to data from the AppDB to enable VOs to associate appropriate data with matching Virtual Appliances • Authentication and Authorization Infrastructure (AAI) integration between the EGI DataHub and with other EGI components and with user communities existing infrastructure • File catalogue to track replication of data: logical and physical file
User of the service/tool	Users or communities with data intensive applications or interested in open data publishing/sharing
User Documentation	https://onedata.org/#/home/documentation/doc/user_guide.html
Technical Documentation	https://onedata.org/#/home/documentation/doc/admin_guide.html
Product team	ACC CYFRONET-AGH
License	MIT/Apache 2.0
Source code	https://github.com/onedata
Testing	Onedata platform, which is the technological solution used for EGI-DataHub undergoes continuous integration and functional tests, constituting over 500 test cases.

2.1.1.2 Release notes

During the reporting period most of the work was focused on preparing a new major release (20.02.4) and ensuring backward compatibility with Oneprovider instances from the previous major release already in production (19.02.X). The new release supports significant upgrades in the data access performance, stability and user interface, a complete changelog can be found at https://github.com/onedata/onedata/blob/develop/CHANGELOG.md. The upgrade to release 20.02.4 was performed on 2nd December 2020, and subsequently in January 2021 it was further upgraded to release 20.02.5.

2.1.1.3 Integration activities and opportunities

During the reporting period EGI DataHub has been integrated with the final version of B2FIND. Existing records in EGI DataHub have been cleaned in order to remove any redundant or test records, and the records were reindexed by B2FIND.

2.1.1.4 Future plans

The future plans are focused on further evolution of EGI DataHub functionality and adoption of new use cases.

2.1.2 B2FIND

2.1.2.1 Service description

Service/Tool name	B2FIND
Service/Tool url	http://b2find.eudat.eu
Service/Tool information page	https://eudat.eu/catalogue/B2FIND
Description	B2FIND provides a discovery portal which allows users to find data collections within an international and inter-disciplinary scope. It is based on a comprehensive metadata catalogue of research data collections stored in EUDAT data centres as well as metadata that are steadily harvested from community specific repositories. Harmonization of the metadata descriptions collected from heterogeneous sources enables not only the presentation in a consistent form, but as well the faceted search across scientific domain boundaries. The backbone of B2FIND is its ingestion process that consists of three steps. First the metadata records - provided by various research communities - are harvested. Afterwards the raw metadata records are converted and mapped to unified key-value dictionaries as specified by the B2FIND schema. The main challenge here is the divergence in metadata standards and schemas; these differences correspond to the specific communities´ needs. To assure and improve metadata quality this mapping process is accompanied by a) iterative and intense exchange with community representatives, b) usage of controlled vocabularies and community specific ontologies and c) formal and semantic mapping and validation. Finally, the mapped and checked records are uploaded as datasets to the catalogue, which is based on CKAN, an open-source data portal software that provides a rich RESTful JSON API and uses Apache SOLR for indexing.
Value proposition	• Allows users to search and browse datasets via keyword searches • Supports faceted, geospatial and temporal metadata searches • Results displayed in user-friendly format and listed in order of relevance • Access to the scientific data objects is given through references provided in the metadata • Available for all interested scientific Communities, Research Infrastructures and Data Providers

User of the service/tool	B2FIND is openly accessible to everyone
User Documentation	https://eudat.eu/services/userdoc/b2find http://b2find.eudat.eu/guidelines/index.html http://b2find.eudat.eu/guidelines/providing.html http://b2find.eudat.eu/guidelines/harvesting.html http://b2find.eudat.eu/guidelines/mapping.html
Technical Documentation	https://github.com/EUDAT-B2FIND
Product team	DKRZ
License	GNU Affero General Public License version 3 (AGPLv3)
Source code	https://github.com/EUDAT-B2FIND/ckanext-b2find https://github.com/EUDAT-B2FIND/md-ingestion https://github.com/EUDAT-B2FIND/ckanext-oaipmh-server https://github.com/EUDAT-B2FIND/ckanext-timeline
Testing	B2FIND offers several test machines that are used for testing a) metadata ingestion of new integrated scientific communities, b) improvements of the ingestion workflow and c) improvements of CKAN. A training test instance is offered, as well as a training module: https://github.com/EUDAT-Training/B2FIND-Training

2.1.2.2 *Release notes*

Since the entire complex of metadata management has increased in importance and developed further in recent years, there was a growing need for enhanced modularisation of the core B2FIND ingestion code. The main aspect is that the ingestion software lacks a flexibility that is required for integrating Communities which expose metadata a) in a non-standardized way using b) non standardized metadata prefixes. Based on this reception the whole B2FIND ingestion code was refactored within the last report period and as new version 3.0.0 has been in October 2020, basic features are now:

- clean, modular python-based code in the backend that enables faster and more flexible metadata integration according to research communities' needs on several complexity levels
- integrated metadata 'reader' for common standards as DataCite, DublinCore, ISO19139 (INSPIRE) and FGDC.
- option to combine different harvesting endpoints for one Community with specific mapping for each endpoint, thus combining different metadata 'reader' for common standards as well as community specific mapping issues within one Community
- new test environment for pytest

The new release for B2FIND ingestion software is openly accessible in GitHub: https://github.com/EUDAT-B2FIND/md-ingestion/releases/tag/v3.0.0

Detailed change history can be tracked at:

https://github.com/EUDAT-B2FIND/md-ingestion/commits/master

Additionally, an expanded Metadata schema has been developed and published as B2FIND Metadata schema v2.0, including now further metadata elements `Size`, `Version`, `Funding Reference` and `Instrument` in accordance with metadata schemas of DataCite and OpenAIRE. B2FIND´s metadata schema requires 7 metadata elements as mandatory (`Community`, `Title`, `Publisher`, `Publication Year`, `Discipline`, `OpenAccess` and at least one `Identifier`) whereas `Community` refers to the Data Provider (or Research Infrastructure) B2FIND harvest from, for `Discipline` a closed vocabulary exists that is based on a classification for scientific disciplines used by re3data and `OpenAccess` is "true" per default, if not specified otherwise within `license` or `rights` information. A further 17 metadata elements are recommended or optional. B2FIND metadata schema description (as well as a crosswalk to DataCite and OpenAIRE metadata schemas) is here: http://b2find.eudat.eu/guidelines/mapping.html

The XSD schema definition is openly accessible here:

http://b2find.eudat.eu/schema/b2f/2.0/meta.xsd

Furthermore, an OAI-PMH extension for CKAN was developed and deployed, which enables harvesting of B2FIND metadata records in several ways, thus widening the visibility of scientific outcome via further aggregators as e.g., OpenAIRE. The following OAI commands have been implemented: ListSets, ListMetadataFormats, ListRecords, ListIdentifiers and GetRecord. B2FIND offers its records with DataCite as a common standard metadata schema (with prefix 'oai_datacite', to be seen here: http://eudatmd2.dkrz.de/oai?verb=ListRecords&metadataPrefix=oai_datacite) and with B2FIND metadata schema (using metadata prefix 'oai_b2f', to be seen here: http://eudatmd2.dkrz.de/oai?verb=ListRecords&metadataPrefix=oai_b2f).

The OAI-PMH extension for CKAN is openly accessible in GitHub:

https://github.com/EUDAT-B2FIND/ckanext-oaipmh-server

Detailed change history can be tracked at:

https://github.com/EUDAT-B2FIND/ckanext-b2find/commits/master

Finally, the whole software stack has been set up on new hardware, which is a Blade Server with 2x8 core CPU, 256 GiB, 2x480 GiB SSD on CentOS 7.8. For the productive machine, an upgrade has been done to CKAN 2.8.4. Therefore, all already existing Communities had to be re-ingested, and several new Communities were integrated. B2FIND now represents meta/data from 36 Communities, displaying more than 1.1 mio metadata records via an openly accessible search portal with extensive search options to enhance precision and recall. The new machine has been opened up to the public on October the 16th: https://www.eudat.eu/news/eudat-unveils-new-improved-b2find-30

2.1.2.3 Integration activities and opportunities

Three major integration projects were completed during the reporting period:

- The integration of EGI-datahub within B2FIND is fostered and enhanced, see details above.

- The EOSC-Nordic project (Discovery and re-use of Nordic community specific data in EOSC) is finalized and in the related deliverable the receipt for harvesting metadata from Nordic repositories to EOSC metadata catalogue B2FIND is described.

- The indexing of data resources within B2SFAFE was achieved within the Herbadrop use case, by harvesting data collections from an elastic search portal rather than directly from iRODS. This generalized approach was deployed for the Herbadrop use case but paves the way for similar approaches by other communities.

Furthermore, OpenAIRE is recently enabled to harvest metadata from B2FIND. So, B2FIND is now compliant with the guidelines of OpenAIRE (Joint activity JA1.9, see above) and looks forward to taking advantage of the added value services of OpenAIRE.

2.1.2.4 Future plans

The focus for the future is to maintain the service and continuously integrate new “Communities” (Data Provider, Research Infrastructures) in B2FIND, for which the new machine and the new software stack offers a basic foundation now. From a technical point of view some concrete goals for the next year are:

- enabling the harvesting of triples, thus integrating Linked Data in B2FIND
- integrating several vocabularies for linked data, e.g., DCAT
- developing and integrating further metadata standards as metadata ‘reader’, e.g., DDI (for Social Sciences) and/or DDI-CDI (as an interdisciplinary standard) or crawling as well from schema.org sources
- further development on the software stack (md ingestion code, oai-pmh-extension for CKAN, CKAN extension for B2FIND, timeline-search) in order to offer a ‘dockerized’ B2FIND to be reusable
- collaboratively develop a generic classification of research areas for research data (clara.science) with partners (re3data, Datacite)
- collaboratively develop and integrate PIDs for Instruments
- improve and extend the validation of metadata by implementation of FAIRisizer tools as FUJI from FAIRsFAIR

2.1.3 B2STAGE

2.1.3.1 Service description

Service/Tool name	B2STAGE
Service/Tool url	https://www.eudat.eu/b2stage
Service/Tool information page	https://github.com/EUDAT-B2STAGE/B2STAGE-GridFTP https://github.com/EUDAT-B2STAGE/http-api
Description	The B2STAGE service allows data transfer into and out of EUDAT data nodes and supports the assignment of unique Persistent IDentifiers (PID) to staged data. EUDAT exposes two protocols for staging data: GridFTP and HTTP-API. The B2STAGE GridFTP service (via the EUDAT Data Storage Interface) is aimed at supporting large data transfer and a large number of files. The key functionality of the B2STAGE GridFTP service is transferring relevant data sets between HPC centers and EUDAT in order to store them, process them and, possibly, move the results back. Data could also be already stored in one or more EUDAT data centers, as a result of the Safe Replication activity. Eventually, output data are identified through a Persistent Identifier (PID). The B2STAGE HTTP-API service is a set of RESTful endpoints that allow access to both B2SAFE data and metadata and it is aimed at small to medium datasets. This service offers programmatic access to data and thus allows for smooth integration of such data into other applications and data services.
Value proposition	• Transfer large data collections from EUDAT storage facilities to external HPC facilities for processing • Ingest computation results onto the EUDAT infrastructure • Access stored data sets through associated PIDs • In conjunction with B2SAFE, replicate community data sets, ingesting them onto EUDAT storage resources for long-term preservation.
User of the service/tool	B2STAGE GridFTP is mainly aimed for power users who need to access data in B2SAFE and move them to compute sites. B2STAGE HTTP-API is mainly conceived for community developers who want to integrate data and features from B2SAFE into their community-specific applications.
User Documentation	https://github.com/EUDAT-B2STAGE/B2STAGE-GridFTP https://github.com/EUDAT-B2STAGE/http-api#user-guide

Technical Documentation	https://github.com/EUDAT-B2STAGE/B2STAGE-GridFTP https://github.com/EUDAT-B2STAGE/http-api#get-started https://github.com/EUDAT-B2STAGE/http-api#administration
Product team	CINECA
License	HTTP API: MIT License
Source code	https://github.com/EUDAT-B2STAGE/B2STAGE-GridFTP https://github.com/EUDAT-B2STAGE/http-api
Testing	B2STAGE http-api can be tested by accessing the online prototype available at: https://b2stage-test.cineca.it/api/status. Protype dedicated instructions: https://github.com/EUDAT-B2STAGE/http-api/blob/master/docs/prototype.md

2.1.3.2 Release notes

Versions released during the reporting period (1.1.1 and 1.1.2) enforced the integration efforts toward B2SHARE and upgraded the underlying dependencies (in particular PostgreSQL, Python, nginx, Flask, ApiSpec and Webargs). A new version (1.1.3) is currently under development. Detailed change history can be tracked at: https://github.com/EUDAT-B2STAGE/http-api/releases

2.1.3.3 Integration activities and opportunities

B2STAGE is natively integrated with B2SAFE and allows to transfer data from and to the bound B2SAFE instance. B2STAGE is able to resolve and create PIDs on B2HANDLE by leveraging the irules integrated in B2SAFE. The authentication mechanism implemented in B2STAGE is allowed to communicate with B2ACCESS OAUTH 2 endpoint to let the user to authenticate.

itself by using B2ACCESS credentials. A PoC of integration between B2SHARE and B2STAGE has been implemented to create a linking between objects indexed in B2SHARE and stored in B2SAFE by letting the user to download resources, identified by PIDs, through B2STAGE.

2.1.3.4 Future plans

The current plan is to continue to maintain the service by continuously upgrading the underlying dependencies. Additionally, bug fixes, in particular for security issues, will continue to be provided. B2STAGE is currently in use for the SeaDataCloud project and deployed in the production environment that will be maintained for the next three years. A follow-up of SeaDataCloud, as well as derived projects, is also possible.

2.1.4 B2DROP

2.1.4.1 *Service description*

Service/Tool name	B2DROP
Service/Tool url	https://b2drop.eudat.eu
Service/Tool information page	https://eudat.eu/services/b2drop
Description	B2DROP is an easy-to-use, user-friendly and trustworthy storage environment which allows users to synchronise their active data across various desktops and to easily share this data with collaborators. The service is targeted at all European researchers whose institute does not host such storage or researchers who need to share their data across institutional boundaries. B2DROP is based on Nextcloud. The service is intended for the long-tail[4] and volatile data which can change and are still subject to active research, e.g., drafts of research papers. Beside the functionalities of sync and share services, B2DROP has advanced functionalities with several extensions. The B2SHARE bridge extension allows users to publish the final documents on B2SHARE. The users select files, the metadata schema and the publication name within B2DROP and the publication draft, containing the supplied information is created in the background. The user finishes the draft on B2SHARE and receives a persistent identifier for the publication.
Value proposition	• Sharing a file with local users • Share publicly across B2DROP instances • Accessing shared artefacts • Desktop synchronization • Mount B2DROP using the WebDAV client • Publishing files to B2SHARE
User of the service/tool	Users who need to share/access small data sets from their desktops
User Documentation	https://eudat.eu/services/userdoc/b2drop
Technical Documentation	https://eudat.eu/services/userdoc/b2drop
Product team	FZ Juelich
License	GNU Affero General Public License v3.0

[4] Data which may be small in volume 10s of GBs but that may serve many researchers

Source code	https://github.com/EUDAT-B2DROP
Testing	Internal continuous integration test suite.

2.1.4.2 Release notes

During the report period two major updates of the software stack were undertaken. Through these updates a hardening of the security was made, and the service performance was enhanced by several database optimizations. Beside personal information from AAI the group membership information is retrieved from the AAI, too. Users will be added to the groups which are provided during the login.

2.1.4.3 Integration activities and opportunities

The integration with B2SAFE and Datahub has been investigated. B2SAFE storage could be added as external storage to B2DROP. In this case the B2SAFE service is added via WebDAV protocol to B2DROP. The procedure to integrate this service was documented and a production deployment is possible for communities on request.
The integration of Datahub has been halted because Datahub and B2DROP do not offer the same set of protocols for the data exchange. If this situation changes in the future the integration activities can be re-started.

2.1.4.4 Future plans

The graphical user interface and the integration with B2SHARE service (B2SHAREbridge) will be adjusted for every release of the underlying software. These are, on the one hand, security changes and, on the other hand, changes for a better usability of the service.
Additionally the functionality and the usability of the B2SHAREbridge will be continuously updated. These changes will offer users an easier workflow from collaborative working on documents to the publication of the results. As one part of these improvements, the implementation to support multiple B2SHARE servers at the same time is ongoing.
Furthermore, the development team will continue investigating which integrations of other services and apps are useful and possible in an easy way for the users.

2.1.5 Summary

The metadata service B2FIND is extended and enhanced to index data originated from EGI-datahub, B2SHARE and B2SAFE. Within the last project period incremental harvesting from all three services could be fostered and enhanced.

Regarding data access, seamless data access and transfer in and between the storage services EGI datahub and B2SAFE was enabled. To realise this the B2STAGE service and AAI proxies and aspects of WP5 (‘Federation Services’) were used.

Regarding the service B2DROP within the last project period the main effort is put on the upgrade of the underlying software, Nextcloud, and the integration with B2SAFE and Datahub. B2SAFE based storage was successfully added as external storage to B2DROP via WebDAV. This procedure has been documented and can be deployed for communities on request. However, issues regarding the authentication token lifetime and network load during data transfers need to be considered, on a case-by-case basis, for production deployments.

The integration of Datahub has been halted due to differences in the protocols for the data exchange.

Concerning the two activities of the cooperation between EOSC-hub and OpenAIRE, the following has been achieved:

- WPA 6.1.11: Integrate EGI DataHub with OpenAIRE Research Community Dashboard by adapting to OpenAIRE guidelines - EGI DataHub conforms to the OpenAIRE Guidelines for Data Archives[5] with respect to support for OAI-PMH interface, which allows metadata harvesting from EGI DataHub service. Furthermore, it supports automatic DOI minting during open data publishing. With respect to DataCite Metadata Schema, a proxy service has been implemented which converts DC records in EGI DataHub metadata into DataCite format automatically.
- WPA 6.1.12: Integrate EUDAT B2FIND/B2SHARE with OpenAIRE Research Community Dashboard by adapting to OpenAIRE guidelines for data providers - B2FIND installed within the last project period, an OAI-PMH CKAN extension, which allows OpenAIRE to harvest from B2FIND. Furthermore, Metadata Schemas of both services - OpenAIRE and B2FIND - are based on DataCite Metadata Schema and so fulfils the OpenAIRE guidelines. Because B2FIND operates as an aggregator of B2SHARE data sources on behalf of OpenAIRE, this activity is also finalized w.r.t. B2SHARE.

Overall, all the integration activities described above and in the work plan have been completed within the reporting period.

2.2 Federated Compute

The Federate Compute task covers those services providing resources for the execution of user applications as virtual machines (Cloud Compute), as containers (Cloud Container Compute), or as jobs (HighThroughput Compute) on the distributed e-Infrastructure. Users needing tighter control on the resources and how these are allocated should use Cloud Compute, users with existing containerised applications following a cloud-native approach are better served with Cloud Container Compute, and for those users with the need to run parallel computing tasks at scale that can be modelled as traditional jobs in a batch system, High Throughput Compute will better meet their needs.

[5] https://guidelines.openaire.eu/en/latest/data/index.html

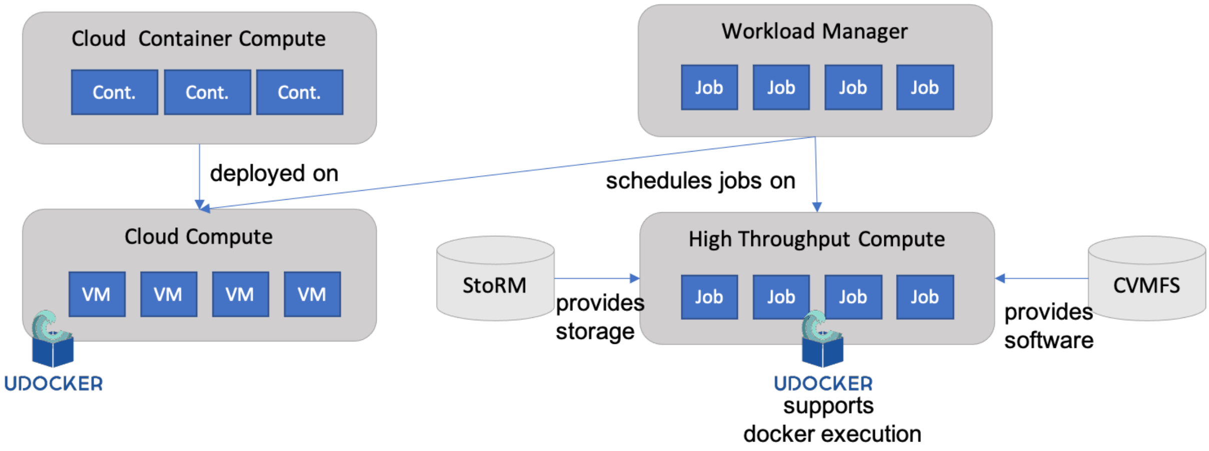


Figure 3. The components of the Federated Compute architecture including dependencies.

These services are complemented and integrated with Workload Management, StoRM (Online Storage), CVMFS and Advanced IaaS to provide advanced features on top of the basic computing power. Workload Manager provides users with an automated distribution of tasks across different computing services. Online Storage offers access to files and objects from the Virtual Machines, containers or jobs. CVMFS offers a software distribution system so the user applications are available in the distributed infrastructure. Advanced IaaS offers the possibility to easily execute containerised applications on systems without native docker support and without administrative privileges as available in the EGI High-Throughput Compute service. A high-level view of this architecture is given in figure 3.

2.2.1 EGI Cloud Compute

2.2.1.1 Service description

Service/Tool name	Cloud Compute
Service/Tool url	https://www.egi.eu/services/cloud-compute
Service/Tool information page	https://www.egi.eu/services/cloud-compute
Description	Cloud Compute gives you the ability to deploy and scale virtual machines on-demand. It offers guaranteed computational resources in a secure and isolated environment with standard API access, without the overhead of managing physical servers. Cloud Compute offers the possibility to select pre-configured virtual appliances (e.g., CPU, memory, disk, operating system or software) from a catalogue replicated across all EGI cloud providers. With Cloud Compute you can:

	• Execute compute- and data-intensive workloads (both batch and interactive) • Host long-running services (e.g., web servers, databases or applications servers) • Create disposable testing and development environments on virtual machines and scale your infrastructure needs • Select virtual machine configurations (CPU, memory, disk) and application environments to fit your requirements • Manage your Cloud Compute resources in a flexible way with integrated monitoring and accounting capabilities
Value proposition	Manage IaaS resources in a multi-cloud environment in a flexible way with single-sign on and integrated image catalogue.
User of the service/tool	Research Communities, individual researchers, VREs, SME/Industry, Resource providers
User Documentation	https://docs.egi.eu/users/cloud-compute/
Technical Documentation	https://docs.egi.eu/users/cloud-compute/federation/
Product team	Partners involved: CESNET, CSIC, EGI.eu, GRNET, IASA, SRCE
License	Apache 2.0
Source code	https://github.com/EGI-Foundation/documentation https://github.com/EGI-Foundation/VMI-endorsement https://github.com/EGI-Foundation/cloud-info-provider https://github.com/IFCA/caso https://github.com/IFCA/keystone-voms https://github.com/the-cloudkeeper-project/cloudkeeper https://github.com/the-cloudkeeper-project/cloudkeeper-one https://github.com/the-cloudkeeper-project/cloudkeeper-os https://github.com/the-rocci-project/keystorm https://github.com/ARGOeu/nagios-plugins-fedcloud https://github.com/IASA-GR/appdb-core https://github.com/IASA-GR/appdb-is-publisher https://github.com/the-oneacct-export-project/oneacct-export https://github.com/goat-project
Testing	All products in CMD (Cloud Middleware Distribution) are verified against EGI's current Quality Criteria. Reports are available in EGI document DB under “Software Provisioning and Technology” category. Installation and configuration notes are available in the EGI repository. Images are tested automatically by the AppDB/SECANT integration.

2.2.1.2 Release notes

The EGI Cloud Service is composed of several related but independently developed components. The following sections describe the main release notes for each of them:

CMD-OS (Components of Cloud Compute for OpenStack):

- cloud-info-provider 0.12.x releases during this period include: a complete automation of the release process for generating artifacts used in the deployment of the component (.deb & .rpm packages and docker containers); improved python 3 support; various fixes found in production and the roll-out to production of the operation of the component centrally. Release details are available at https://github.com/EGI-Foundation/cloud-info-provider/releases. An automated git-based operation of the component was also introduced using GitHub Actions CI/CD features at the https://github.com/EGI-Foundation/fedcloud-catchall-operations repository. Changes in the repository once approved via Pull Requests are deployed in the infrastructure automatically.
- cASO 1.4.x releases introduced several more fixes for the long-running VM accounting issues found in production. A cASO v2.0.0 release introduces support for floating IP accounting. Details available at https://github.com/IFCA/caso/releases.
- A new accounting probe framework (goat, https://github.com/goat-project) was developed to improve accounting features of the Cloud providers, supporting reporting of VM usage[6], floating IP usage and storage usage to the central EOSC-hub accounting repository.
- A new command line interface tool, named *egicli*, was developed to facilitate the use of the infrastructure. The tool is distributed as a python package at PyPi, with complete release notes and available at https://github.com/EGI-Foundation/egicli.

AppDB:

The AppDB development focused on the adaptation to the infrastructure changes introduced by CMD components and AAI. Complete list of changes is available at the AppDB Changelog at https://wiki.appdb.egi.eu/main:about:changelog:

- Finalise move to GLUE schema 2.1 and AMS as information transport and deprecation of GLUE schema 2.0 and BDII.
- New information system implementation that to provides complete information on the infrastructure, including endpoints supporting a REST API along with a Swagger UI interface (http://is.marie.hellasgrid.gr/rest/), GraphQL queries (http://is.marie.hellasgrid.gr/graphql) and a complete GraphQL UI to test queries (http://is.marie.hellasgrid.gr/tools/graphiql)
- Migration of VMOPs to the new information system and OpenID Connect for authentication. This change removes the need of having X.509 certificates supported in the infrastructure, which are deprecated now for authenticating users.
- Improved Continuous Delivery features for Virtual Appliance images.
- Introduced a security dashboard for controlling the VA available in the infrastructure.
- Introduced endorsement dashboard to allow users to endorse specific VA versions

[6] https://docs.egi.eu/users/cloud-compute/federation/#accounting

EGI images:

As with previous periods, all the images provided by EGI in AppDB were updated by using the latest packages versions from upstream distribution. The creation mechanism was also improved to rely on the Continuous Delivery features of AppDB. CentOS 6 and Ubuntu 14.04 images were removed as they are EOL'ed upstream. New images for Ubuntu 20.04 and CentOS 8 were introduced.

Monitoring:

A new monitoring probe for the cloud-info-provider was introduced into the EGI monitoring system (details at https://github.com/ARGOeu/nagios-plugins-fedcloud)

Documentation:

A complete revision and reorganisation of the service documentation was performed during this period. Both user and provider documentation are now centrally available under https://docs.egi.eu/. An automated build system for the documentation with a Pull-Request review system for incorporating changes was developed. Documentation now includes better description of the interaction of the EGI Cloud providers with the different components of the infrastructure and covers configurations commonly requested by providers (like multiple OpenID Connect AAI systems integrated into a single provider).

2.2.1.3 Integration activities and opportunities

During the EOSC-hub the EGI Cloud Compute has been adapted to: (i) allow the integration of multiple IdPs at a single provider of the infrastructure, thus facilitating the use from different communities relying on different EOSC-hub AAI implementations; (ii) provide accounting records for floating IP records and storage usage to the central accounting repository; (iii) use Argo Messaging System as a transport for information and deprecating legacy interfaces based on the BDII; and (iv) use of GlueSchema 2.1 so resource discovery is easier for any components interacting with the infrastructure.

Besides these lower-level integration activities, the EGI Cloud Compute is a flexible service that allows users to run almost any kind of workload. This makes it possible to integrate with multiple services at the user-level and has been successfully used by services like Infrastructure Manager, PaaS Orchestrator, EGI Workload Manager, or DODAS to run user applications. It is also a major contributor to the Early Adopter Programme use cases, hosting several of the platforms developed under that activity.

2.2.1.4 Future plans

With the cloud-info-provider now fully centralised, the cloudkeeper and accounting will be also moved to a similar setup, facilitating the integration of new providers. GPGPU usage will be improved with better information using Glue 2.1 and accounting. AAI integration will also be improved with account de-provisioning. Support for quota-related information will also be covered in future releases of the components.

2.2.2 EGI Cloud Container

2.2.2.1 *Service description*

Service/Tool name	Cloud Container Compute
Service/Tool url	https://www.egi.eu/services/cloud-container
Service/Tool information page	https://www.egi.eu/services/cloud-container
Description	Cloud Container Compute (in beta phase) gives you the ability to deploy and scale Docker containers on-demand. It offers guaranteed computational resources in a secure and isolated environment with standard API access, without the overhead of managing the operating system. Main characteristics: • On-demand provisioning • Lightweight environment for maximised performance • Standard interface to deploy on multiple service providers • Interoperable and transparent • Removes friction between development and operations environments.
Value proposition	Deploy, manage, and scale containerized applications using Kubernetes on EGI Cloud providers.
User of the service/tool	Research Communities, individual researchers VREs, SME/Industry, Resource providers
User Documentation	https://docs.egi.eu/users/cloud-container-compute/
Technical Documentation	https://docs.egi.eu/users/cloud-container-compute/
Product team	EGI Foundation, UPV
License	Apache 2.0
Source code	https://github.com/grycap/ansible-role-kubernetes https://github.com/EGI-Foundation/VMI-endorsement
Testing	Images are tested automatically by the AppDB/SECANT integration. GitHub Actions integration for Kubernetes: https://github.com/grycap/ansible-role-kubernetes/actions?query=workflow%3ACI

2.2.2.2 *Release notes*

The EGI Cloud Container Compute has seen a major redesign being now capable of deploying scalable Kubernetes clusters on the cloud providers of the federation for the execution of any container-based applications. This development is based on the EC3 component and Infrastructure

Manager (task T6.3) for deployment and configuration of resources. This kind of managed deployment of Kubernetes is a feature commonly requested by several use cases.

The grycap.kubernetes ansible role includes several new features and fixes:

- Remove the usage of the insecure port (8080) in local connections as it has been removed in Kubernetes version 1.20.
- Deploy nfs-client-provisioner without helm as it has been removed from helm repos.
- Enhance NVIDIA docker support in CentOS VMs.
- Update CNI drivers and remove deprecated ones (Romana).
- Add support to use Cert-Manager to automatically generate Let's Encrypt valid certificates.
- Updater helm to version 3.

2.2.2.3 Integration activities and opportunities

The EGI Cloud Container Compute now is capable of leveraging all major cloud providers, including the EGI Cloud Compute based on OpenStack thanks to the multi-cloud support provided by EC3. The service is focused in providing a working kubernetes experience for all kinds of users and integrates the container orchestration system with commonly used features in research-oriented clouds like OpenID based authentication, NFS and OpenStack for volumes, GPGPUs.

The increased popularity of Kubernetes makes the service an attractive platform for deployment of any kind of applications.

2.2.2.4 Future plans

Now that the EGI Cloud Container service provides automated deployment of Kubernetes for users, the work will focus on maintenance and continuous improvement of the recipes used for the deployment of the clusters and adding new features as requested by the users.

2.2.3 EGI Workload Management

2.2.3.1 Service description

Service/Tool name	EGI Workload Manager
Service/Tool url	https://dirac.egi.eu/DIRAC
Service/Tool information page	https://www.egi.eu/services/workload-manager/
Description	The EGI Workload Manager service to schedule and manage centrally thousands of computational tasks on cloud and HTC. The service is based on the DIRAC Interware framework and is composed of several components. It allows users to submit jobs, monitor their status and retrieve the results. Users can also use storage resources to store and register their data. User communities can connect their specific computing and storage resources to

	enlarge the total available capacity. Users interact with the DIRAC services both using GUI and API. The component of the DIRAC service are: • DIRAC WMS services (multiple high-performance machines) • DIRAC DB (MySQL) server (high performance, large memory machine) • DIRAC REST server (medium sized machine) • DIRAC Web server (low CPU, high memory machine) • DIRAC configuration server, one central master and multiple distributed slaves (low CPU, high memory machine) • DIRAC data manager service (low CPU, high memory machine)
Value proposition	• Retain communities and attract new communities into the dirac.egi.eu service • Provide more flexible VRE simplifying access to EGI resources and accelerate the ability of researchers to undertake excellent science with DIRAC technical innovations • Transfer DCI skills and know-how to other medium and big communities and resource providers in the context of EGI
User of the service/tool	Users who need to deploy parallel applications in federated infrastructures or Cloud platforms.
User Documentation	https://docs.egi.eu/users/workload-manager/
Technical Documentation	https://dirac.readthedocs.io/en/latest/DeveloperGuide/Overview/index.html?highlight=workload%20manage
Product team	CNRS
License	GNU General Public License v3.0
Source code	https://github.com/DIRACGrid/DIRAC
Testing	Internal continuous integration test suite: https://travis-ci.org/DIRACGrid/DIRAC

2.2.3.2 Release notes

The EGI Workload Manager service based on the DIRAC Interware framework[7] has continued the maintenance and evolutionary work. The service was regularly updated to follow the evolution of the DIRAC software to include new types of resources and new computing elements. In particular, resources of the OSG infrastructure were connected in order to increase the capacity used for the COVID-19 related research.

The S3 protocol-based storage was deployed at CYFRONET and connected to the DIRAC service. A special S3 storage plugin was developed to allow a seamless integration into the DIRAC Data Management System.

The resources tagging mechanism was applied in order to increase priorities of the COVID-19 related tasks and allow resources providers to monitor their contributions into these studies. The work on integration with the EOSC-hub AAI infrastructure continued in a close collaboration with the Check-In service developers.

2.2.3.3 Integration activities and opportunities

The EGI Workload Management service integrates various types of computing resources into a single widely distributed system with a single common access point. Among resources accessible through the DIRAC services there are HTC, cloud and HPC computing resources. Both resources integrated into the EGI infrastructure and specific computing and storage resources of the user communities can be used together in the same workflows. Users access the services in a secure way with the X509 based certificates. Integration with the OAuth based AAI systems is developed to enable integration with the EGI Check-In service and other similar services, e.g., INDIGO AIM. The DIRAC user interface is integrated with the EGI Jupyter Notebooks service to enabled access to the EGI Workload Manager

The development of the new DIRAC component (RucioFileCatalog) allows integration with the Rucio Data Management service and allows access to the Rucio managed storage systems from the user jobs managed by the DIRAC services.

2.2.3.4 Future plans

Further plans include various developments that will enrich the service and facilitate its access:

- The client Python API will be upgraded to allow using of the Python3 interpreter
- The new client/service protocol based on HTTP and OAuth based tokens will provide standardized access to the DIRAC functionality. Together with the REST interface it will allow DIRAC connections in a language neutral way
- Integration with the EOSC-hub AAI service will allow using the user community policies applied to the payloads managed by DIRAC

The EGI Workload Manager service hosting will be moved from the CYFRONET cloud system to the IN2P3 Computing Center, CNRS. This will allow integration of the service with the similar France-

[7] http://diracgrid.org/

Grilles DIRAC service. This will help optimization of the service operations as well as reach new user communities.

2.2.4 StoRM - EGI Online Storage

2.2.4.1 Service description

Service/Tool name	EGI Online Storage/StoRM
Service/Tool url	https://www.egi.eu/services/online-storage/
Service/Tool information page	https://www.egi.eu/services/online-storage/
Description	EGI Online Storage allows you to store data in a reliable and high-quality environment and share it across distributed teams. Your data can be accessed through different standard protocols and can be replicated across different providers to increase fault-tolerance. Online Storage gives you complete control over the data you share and with whom. The service is powered by StORM (STOrage Resource Manager): a light, scalable, flexible, high-performance, file system independent, storage manager service (SRM) for generic disk-based storage system, compliant with the standard SRM interface version 2.2.
Value proposition	Easily share and organise your data, and control the data you share
User of the service/tool	Researchers, VREs, SME/Industry
User Documentation	http://italiangrid.github.io/storm/documentation.html
Technical Documentation	http://italiangrid.github.io/storm/documentation.html
Product team	INFN
License	Apache 2.0
Source code	https://github.com/italiangrid/storm
Testing	Quality Criteria Report: https://documents.egi.eu/document/3335 Stage Roll out Report: https://documents.egi.eu/document/3335

2.2.4.2 Release notes

In the reporting period, StoRM has produced 4 releases, detailed below.

StoRM 1.11.16 (https://italiangrid.github.io/storm/2019/10/02/storm-v1.11.16-released.html):

- introduces the support for the CKSUM command, so that an ADLER32 checksum is returned if already known for a file, or computed on the fly and stored in an extended attribute;
- introduces configurable support for Conscrypt in StoRM WebDAV that improves TLS performance for Java applications by delegating the handing of cryptographic operations to boringssl (the Google fork of OpenSSL);
- fixes StoRM WebDAV start-up failure due to an unreachable OpenID Connect provider and some minor configuration issues;
- fixes error description when a SRM mkdir path contains non existing intermediate directories.

StoRM 1.11.17 (https://italiangrid.github.io/storm/2019/12/17/storm-v1.11.17-released.html):

- This release introduces the support for CentOS 7 for StoRM WebDAV and StoRM GridFTP.

StoRM 1.11.18 (https://italiangrid.github.io/storm/2020/08/07/storm-v1.11.18-released.html):

- This release introduces the support for CentOS 7 for all StoRM components.
- Bug fixes for several outstanding bugs: fixes errors on published storage space occupancy in case multiple storage area shares the same VO; fixes not published WebDAV endpoints when latest logic is used; fixes not dropped Authorization header in WebDAV TPC redirects; fixes leaked file descriptors when Conscrypt is enabled on StoRM WebDAV; sets correctly HTTP content-length for large files; fixes errors on transferred files through GridFTP that leave empty files with an adler32 checksum for a non-empty file; fixes KillMode on GridFTP systemd unit.
- Implemented Improvements: fixes wrong ERROR log messages when file does not exist on srmRm requests; changes the way info provider checks if Backend is running; introduces a Background DU Service (disabled by default) that periodically updates the storage space info for non-GPFS storage areas (read more info here); adds Date and thread pools metrics in the metrics logged info; updates spring boot to 2.2.6 release for StoRM WebDAV; adds SystemD support for StoRM Backend and StoRM Frontend.

StoRM 1.11.19 (https://italiangrid.github.io/storm/2020/10/29/storm-v1.11.19-released.html):

- fixes a bug introduced with StoRM v1.11.18 about the final update of the status on database of a srmPtG or srmBoL requests;
- introduces new metrics in storm-backend-metrics.log;
- fixes a storm-webdav bug about the returned body content in case of HEAD requests.

2.2.4.3 Integration activities and opportunities

During the reporting period, StoRM WebDAV component has improved its support for token-based authentication and authorization by introducing an internal OAuth authorization server that can be used to issue tokens to the clients authenticated with VOMS credentials. It also supports OpenID connect authentication and authorization on storage areas, so it can be configured to use INDIGO IAM or EGI Check-in services.

2.2.4.4 Future plans

Keep up with the maintenance and evolution of the StoRM codebase and its integration with EOSC-hub services. In more detail, we have already started working on

- a deep clean-up of the StoRM frontend code in order to avoid memory leaks
- the refactoring of the Backend database connection pool
- enhancements to the Third-Party Copy support on the WebDAV service
- support for CentOS 8

2.2.5 Advanced IaaS

2.2.5.1 Service description

Service/Tool name	Advanced IaaS
Service/Tool url	https://github.com/indigo-dc/udocker
Service/Tool information page	https://github.com/indigo-dc/udocker/blob/master/SUMMARY.md
Description	udocker is a user-level tool that enables users to run containerized applications without needing root privileges on the machine, and without having to install any additional system software. udocker offers a subset of the Docker functionalities aimed at supporting unprivileged execution. udocker has a command line interface similar to Docker and can be used to pull containers from Docker hub or other Docker like repositories and execute them without requiring the presence of Docker. udocker can also import standard directory tarballs making possible the execution of non-Docker containers. udocker can be easily transferred to the target Linux system by the end user and simply executed there without any need of compilation or system wide installation, therefore does not requiring the system administrator intervention. It relies on several user level mechanisms to provide a chroot-like environment to execute applications and services, but without requiring privileges. udocker integrates several drivers (execution modes) to offer the best possible environment for the execution of containers without requiring administrator privileges.
Value proposition	• Running Linux containers on machines without Docker or any other container runtime locally installed • Running Linux containers on machines without root privileges • Running Linux containers without having to compile additional software. • Using Linux containers in batch, interactive or cloud systems including grid, HTC and HTP clusters • Running containers in systems where kernel features needed to run containers are unavailable or disabled • Enables execution of applications with GPGPUs. • Empower users to execute their applications everywhere through container encapsulation
User of the service/tool	Users with complex applications and software dependencies who need to run applications or services encapsulated in containers, but which do not have administrator privileges on the execution nodes. This includes computing clusters many of which may not be willing to

	support conventional container execution due to security concerns and related policies. udocker is also being used successfully to support containers execution in cloud computing namely to exploit function-as-a-service.
User Documentation	https://github.com/indigo-dc/udocker/blob/master/doc/user_manual.md
Technical Documentation	https://github.com/indigo-dc/udocker/blob/master/SUMMARY.md
Product team	LIP
License	Apache License 2.0
Source code	Stable release https://github.com/indigo-dc/udocker Python 3.x latest release https://github.com/indigo-dc/udocker/tree/devel3 Python 2.x latest release https://github.com/indigo-dc/udocker/tree/devel
Testing	Functional, integration and unit tests available at: https://github.com/indigo-dc/udocker/tree/master/tests https://github.com/indigo-dc/udocker/tree/devel3/tests https://github.com/indigo-dc/udocker/tree/devel/tests The udocker team uses a Jenkins CI system to perform code style verification, execution of tests and identification of potential security issues.

2.2.5.2 Release notes

Three latest releases of the Python 2.x version were performed under the development branch (1.1.5, 1.1.6 and 1.1.7). These include latest developments and fixes which were also pushed to the latest releases of the Python 3.x version under the devel3 branch. Information regarding the individual releases can be obtained from the changelogs available at the GitHub repository. The introduced developments and fixes can be briefly described as follows.

The porting of udocker to Python 3.x was finished enabling the continued support of udocker beyond the Python 2.x end of support. The complete Python 3.x version that also supports Python 2.7 is currently available from the devel3 branch and will be soon released as stable, enabling python 3 to become the default production version. The PTRACE execution engine was improved to enable dynamic detection of the SECCOMP filtering implementation. This provides better support for the P1 accelerated execution mode across different kernels, fixing among other problems with the backporting of SECCOMP kernel patches. The processing of ELF headers by the fakechroot execution engine was improved enabling the support of Java and a wider range of other applications under the F3 and F4 execution modes. New shared preload libraries were added to the fakechroot engine supporting the latest Fedora, Ubuntu and Alpine releases. Several fixes were introduced to address minor issues.

2.2.5.3 Integration activities and opportunities

The support for rootless execution was improved both through the aforementioned developments and furthermore by integrating crun (https://github.com/containers/crun) into the fakechroot engine thus extending the support for OCI compliant tools. Therefore, the udocker Fn execution modes can now exploit both crun and runc as underlying execution engines with benefits in terms of performance and interoperability. The selection between both options is automatically performed by udocker, but the user can also force the selection. Also, in this context the support for OCI images was also improved. Further opportunities for exploitation and integration have been pursued in contact with projects, namely in EOSC-synergy to support more EOSC thematic services and Big-HPC to support big data and HPC applications.

2.2.6 CREAM - EGI High-Throughput Compute

2.2.6.1 Service description

Service/Tool name	CREAM
Service/Tool url	CREAMCE is a distributed service, with multiple endpoints.
Service/Tool information page	https://www.egi.eu/services/high-throughput-compute
Description	The CREAM (Computing Resource Execution And Management) Service is a simple, lightweight service that implements all the operations at the Computing Element (CE) level. The service is a basic component for a federated service-oriented architecture managing distributed processes (jobs). In order to guarantee the interoperability among different applications it implements a standard Web Service interface based on WS-I specification. The CREAM service accesses and operates local resource management systems. The current version of the application supports the following management systems: TORQUE, LSF, SLURM, HTCondor, Grid Engine (partially). For the Federated Compute package of EOSC-hub a new installation and configuration tool has been developed for the CREAM and the resource BDII services. The tool is based on the Puppet framework and replaces the old reference application (YAIM) for the Grid environment, and it is distributed as a Puppet module through the Puppet Forge portal: https://forge.puppet.com/infnpd/creamce

Value proposition	- Workload and data management tools to manage all computational tasks - Large amounts of processing capacity over long periods of time - Faster results for your research - Shared resources among users, enabling collaborative research
User of the service/tool	Users of large-scale data intensive computational parallel applications.
User Documentation	https://cream-guide.readthedocs.io/en/latest
Technical Documentation	https://cream-guide.readthedocs.io/en/latest
Product team	INFN
License	Apache Software License, v.2.0
Source code	https://github.com/italiangrid
Testing	The platform chosen for the continuous integration is Jenkins; the build system takes advantage of its advanced features, like pipeline processing and Docker container support, for compiling and testing the code efficiently. The system at the moment makes use of the Jenkins platform hosted at INFN; it will be completely integrated into UMD workflow as soon as the compatibility is certified.

2.2.6.2 Release notes

No release of the software component was done during the reporting period.

2.2.6.3 Integration activities and opportunities

The CREAM-CE will no longer be used as a job submission gateway for the EGI High Throughput Compute service as the development will be discontinued. Alternative implementations (HTCondor-CE and ARC) were identified during the EOSC-hub project and providers of the service have migrated to them ensuring the continuity of the service. Integration of these solutions with other WP6 services (e.g., EGI Workload Manager, EGI Online Storage) and central services (e.g., accounting, monitoring) is already in place.

2.2.6.4 Future plans

CREAM-CE is to be deprecated with the end of EOSC-hub and no further releases will be produced.

2.2.7 Summary

This task goal dealt with the corrective, adaptive and perfective maintenance of the Federated Compute Services to provide powerful and fully customisable computing solutions for EOSC users. These services are now available through the EOSC Portal and completely integrated with the EOSC AAI framework. Within the last project period, EGI Cloud Compute has improved its integration with core services (AAI, accounting, monitoring) and improved its documentation to facilitate usage and integration of new providers. EGI Cloud Container Compute has evolved to provide on-demand deployment of kubernetes matching the requirements of several user communities of EOSC. The EGI Workload Manager was extended to support more computing and storage endpoints and better integration with AAI. The StoRM server (EGI Online Storage) was upgraded to support CentOS 7 and introduced several fixes and enhancements. udocker saw several releases that introduced full support for Python 3 and new features to improve compatibility of applications. The CREAM component of EGI High Throughput Compute service was phased out during the last reporting period and alternative solutions were identified and rolled out into the production infrastructure.

2.3 Processing and Orchestration

This task focuses on the maintenance and integration of orchestration services with the Cloud Compute and Cloud Container services. This allows to build complex virtual computing infrastructures based on the OASIS TOSCA Simple Profile YAML standard[8] and integrate the INDIGO-DataCloud PaaS components as orchestrator for the EOSC-hub services.

Figure 4 provides an overview of the architecture and interrelation of the different components that are part of task T6.3 “Processing and Orchestration”. It also includes additional components (such as the Cloud Provider Ranker or the Monitoring system) that, even though they are not strictly included in T6.3 since they are not expected to be evolved in the context of EOSC-hub, they are part of the PaaS Orchestration layer.

[8] Crandall, John, and Paul Lipton. “OASIS Topology and Orchestration Specification for Cloud Applications (TOSCA) TC.” https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=tosca.

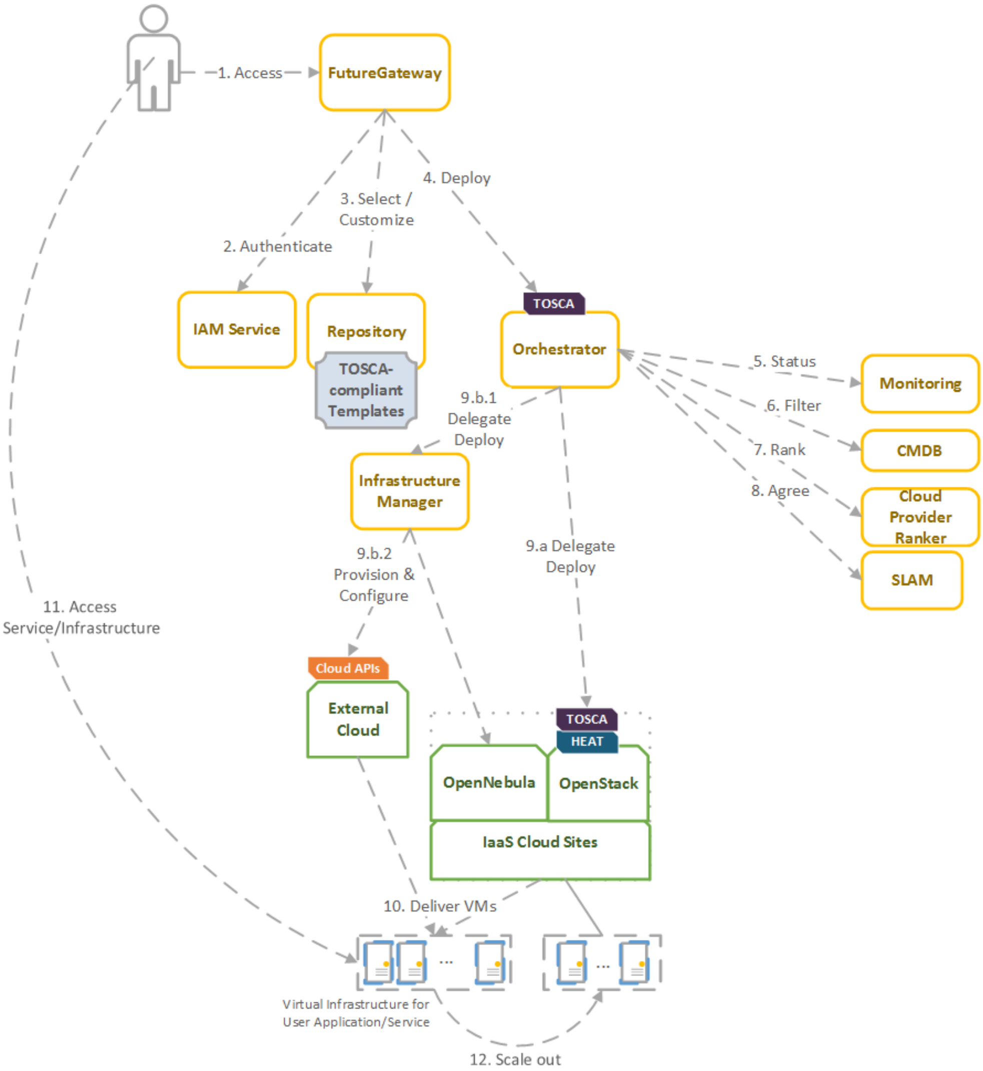


Figure 4. Architecture and workflow among the Processing and Orchestration services in EOSC-hub.

End users are expected to access the PaaS Orchestration layer via high-level graphical user interfaces, such as the portlets provided by the *FutureGateway*. Each portlet is a graphical representation of a set of TOSCA templates available in a defined repository. TOSCA stands for the Topology and Orchestration Specification for Cloud Applications and it is a YAML-based domain-specific language (DSL) to describe application architectures to be deployed on a Cloud. These templates are developed by some TOSCA expert users, specifically for each application supported by the platform. Therefore, end users only have to introduce a set of input values to deploy their applications. Also, the FutureGateway is responsible for performing the authentication with the *IAM service* and interacting with the *Orchestrator* by submitting the TOSCA template selected by the user, with his/her specific set of input values. Advanced end-users could also interact with the Orchestrator via the authenticated REST APIs provided.

Once the Orchestrator receives the TOSCA template it is in charge of interacting with different services in order to identify the most appropriate IaaS Cloud site on which to perform the execution. This decision depends on the monitoring state of the underlying Cloud sites (information managed by the *Monitoring* service), the SLAs (Service Level Agreements) agreed between the user and the sites (information managed by the *SLAM* service) and the availability of the VMIs (Virtual Machine Images) on each site (information managed by the *CMDB* service). With all of this information, the *CloudProviderRanker* service is employed in order to apply a set of rules in order to obtain a ranked list of Cloud sites.

The Orchestrator delegates on the *Infrastructure Manager (IM)* to perform the deployment on public Cloud sites (Amazon Web Services, Microsoft Azure, Google Cloud Platform and Open Telekom Cloud) or on other external Clouds managed by popular Cloud Management Platform (CMPs) such as OpenNebula and OpenStack. The IM can also be configured with single-site mode in order to provide a TOSCA-enabled endpoint and support local-site orchestration of complex application architectures. In order to achieve a similar functionality in OpenStack, the *HEAT Translator* component can be used in order to translate from a TOSCA template into a HOT template, the native language employed by the HEAT service in OpenStack.

The following sections provide an overview of the services involved in task T6.3 “Processing and Orchestration in EOSC-hub”. For each service, a brief description is included, and, in a separate section, the integration activities carried out in EOSC-hub for each service, during the last reporting period, are described.

2.3.1 TOSCA for HEAT

2.3.1.1 Service Description

Service/Tool name	TOSCA for Heat
Service/Tool url	https://wiki.openstack.org/wiki/Heat-Translator
Service/Tool information page	https://wiki.openstack.org/wiki/Heat-Translator
Description	TOSCA for Heat service translates OASIS TOSCA templates in order to produce valid HOT (Heat Orchestration Template)-compliant documents ready to be deployed by the OpenStack Orchestration program, Heat. The main component is the heat-translator command-line tool, which, in the case of incoming TOSCA templates, makes use of in-memory graphs provided by TOSCA Parser to obtain the resultant HOT.
Value proposition	Deployment of cloud topologies in OpenStack Heat described using OASIS TOSCA standard.
User of the service/tool	User communities interfacing directly with an OpenStack IaaS framework to deploy complex applications using TOSCA templates.
User Documentation	https://docs.openstack.org/heat-translator/latest/

Technical Documentation	https://docs.openstack.org/heat-translator/latest/
Product team	CSIC
License	Apache License 2.0
Source code	https://github.com/openstack/heat-translator https://github.com/indigo-dc/heat-translator
Testing	https://zuul.openstack.org/job/heat-translator-tox-py27-tp https://jenkins.indigo-datacloud.eu:8080/job/Pipeline-as-code/job/heat-translator/

2.3.1.2 Release notes

No release of the software component was done during the reporting period.

2.3.1.3 Integration activities and opportunities

As an accessory tool for OpenStack cloud frameworks, the TOSCA-for-HEAT was extended throughout the lifespan of the EOSC-hub project in order to cover the new features included in the TOSCA templates used by the project's thematic services, which required the orchestration capabilities provided by the OpenStack's HEAT service. However, as a result of the rapid evolution and extended usability of container orchestration technologies, the HEAT solution was increasingly less used in order to satisfy the requirements of the use cases in the EOSC-hub project. When it comes to complex deployments in OpenStack cloud providers, the maintenance costs of the TOSCA-for-HEAT translator have proven to be higher than leveraging such container orchestration solutions, deployed over standalone virtual machines that are spawned directly using the OpenStack's API.

2.3.1.4 Future plans

According to the previous section, the support for the TOSCA-for-HEAT component will be discontinued.

2.3.2 Infrastructure Manager

2.3.2.1 Service Description

Service/Tool name	Infrastructure Manager
Service/Tool url	REST API: https://appsgrycap.i3m.upv.es:31443/im Web portal: https://appsgrycap.i3m.upv.es:31443/im-web/ Dashboard: https://appsgrycap.i3m.upv.es:31443/im-dashboard/
Service/Tool information page	http://www.grycap.upv.es/im/index.php
Description	IM is a tool that orchestrates the deployment of complex and customized virtual infrastructures on multiple back-ends. The IM automates the deployment, configuration, software installation, monitoring and update of virtual infrastructures. It supports a wide variety of back-ends, including both public IaaS Clouds (Amazon Web Services, Microsoft Azure, etc.), on-premises Cloud Management Platforms (OpenNebula, OpenStack, etc.) and Container Orchestrators (Kubernetes), thus making user applications Cloud agnostic. In addition, it features DevOps capabilities, based on Ansible enabling the contextualization of the infrastructure at run-time by installing and configuring all the required software that may not be available in the Virtual Machine Images, thus providing the user with a fully functional virtual infrastructure. The IM provides both XML-RPC and REST APIs to enable high-level components to access its functionality. These APIs provide a set of functions for clients to create, destroy, and get information about the infrastructures. It also enables elasticity management, both horizontal (adding/removing nodes) and vertical (growing/shrinking the capacity of nodes). The cloud topologies can be defined using its native language called RADL (Resource and Application Description Language) or TOSCA OASIS standard (YAML version 1.0).
Value proposition	Deployment of complex and customized virtual infrastructures on multiple Cloud back-ends using standard specification languages.
Customer of the service/tool	Research Communities, VREs, individual researchers.
User of the service/tool	Research Communities, individual researchers.
User Documentation	https://imdocs.readthedocs.io/
Technical Documentation	https://imdocs.readthedocs.io/

Product team	GRyCAP - UPV
License	GPL 3.0
Source code	https://github.com/grycap/im
Testing	https://jenkins.i3m.upv.es/job/grycap/job/im-unit-master/

2.3.2.2 Release notes

In previous versions the IM has been improved to enable the creation of private networks, in this period it has improved network creation enabling wildcard (192.168.*.0/24) and it will select the first network address range available in the cloud provider.

To enhance the contextualization steps in some cloud providers where there are strong restrictions with public IPs or firewalls, the IM has been improved to enable the configuration of infrastructures where port SSH is closed by organization firewall or infrastructures without any VM with public IP.

Furthermore, to enhance the deletion process it has been enabled the option to delete an infrastructure asynchronously (also adding a new “deleting” state).

2.3.2.3 Integration activities and opportunities

The IM has been evolved in the framework of the EOSC-hub project. The first step was to test the authentication systems provided in the project. IM was already integrated with the INDIGO IAM (based on OpenID) and it has been successfully tested with the EGI Check-In system (also based on OpenID) without any code modification.

As requested by the DODAS Thematic service support for EC2 spot instances has been added. EC2 spot instances were already supported in the AWS IM connector but TOSCA support has been added. This required adding the “preemptible instance” property in the TOSCA compute node definition and the proper translation in the IM orchestrator core.

Regarding the extension of the TOSCA types, two main contributions have been made: Kubernetes and JupyterHub. A set of Kubernetes node types have been defined to allow the user to specify a TOSCA document describing a Kubernetes cluster. This enables the user, together with the Ansible roles commented in section 4.2.6, to automatically provision a fully functional Kubernetes cluster. This definitely eases the Kubernetes cluster as a Service functionality required in the ELIXIR CC.

The IM has been improved to enable the creation of private networks, in those sites where the policies enable the user to do it, instead of selecting from the ones previously created by the administrator. This enhances the security on the virtual infrastructures as the networks are not shared with other users/infrastructures. Also, the addition of static routes has been integrated in order to use some instances (Virtual Machines) as routers for all the nodes of the created network. These new functionalities are oriented to improve the creation of hybrid deployments among different sets of cloud providers, thus enhancing network connectivity.

The IM has included support to configure VMs using SSH proxy connections and tunnels in order to provide increased support for hybrid deployments using OpenStack and also public Cloud providers such as Amazon Web Services (AWS). Also, the contextualization process was modified to support Python 3 and to reduce the number of SSH logins in order to speed it up. A new JavaScript-based Software Development Kit (SDK) was created to increase adoption. Increased support for requesting GPUs in the OpenStack connector was included together with support for additional providers (Linode and Orange Cloud).

2.3.2.4 Future plans

Continue with the maintenance of the service and add or improve the integration with EOSC-hub services as EGI Cloud Compute, AppDB VMOps, AppDB IS, EGI Checkin, etc.

2.3.3 PaaS Orchestration system

2.3.3.1 Service Description

Service/Tool name	PaaS Orchestrator
Service/Tool url	https://indigo-paas.cloud.ba.infn.it/orchestrator
Service/Tool information page	https://indigo-dc.gitbooks.io/indigo-paas-orchestrator/content
Description	The Orchestrator is the core component of the INDIGO PaaS system and the entry point for the users' requests: it allows to coordinate the provisioning of virtualized compute and storage resources on Cloud Management Frameworks, both private and public (like OpenStack, OpenNebula, AWS, etc.), and the deployment of dockerized long-running services and batch jobs on Apache Mesos clusters. It receives the deployment requests, expressed through templates written in TOSCA (Simple Profile in YAML version 1.0), and orchestrates the deployments on the best available cloud sites. In order to select the best site, the Orchestrator implements a complex workflow: it gathers information about the SLAs signed by the providers with the user, the monitoring data about the availability of the compute and storage services, the location of the data requested by the user (if any). Hybrid deployments spanning multiple sites are supported. Using the PaaS Orchestrator and the TOSCA templates, the end users can exploit computational resources without any knowledge about the IaaS details. The end users can submit their request to the Orchestrator using different clients, e.g., the web dashboard or the command-line interface, or directly using the REST APIs. Anyway, the

	Orchestrator is the only service the users need to interact with for creating, monitoring and deleting their deployments. In order to coordinate the complex deployment workflow, the Orchestrator relies on the following INDIGO PaaS services: *CMDB* (Configuration Management Database) is a database system in INDIGO PaaS environment used by other platform components as a central point of technical information such as location, types and other metadata about sites, services and virtual machine images. CMDB provides a REST API for that purpose. The interface allows to manage schema of the entities stored in the database. The schema is enforced in operations such as CREATE, READ, UPDATE and DELETE. Security of these operations is ensured by different access levels for operators and other users as well as hierarchy of entity ownership. *SLAM* (SLA manager) is a component of the INDIGO PaaS system that handles negotiations between infrastructure customers and infrastructure providers. *Customer* is an entity representing a group of people (system users) sharing common interests and entitled to share computing and storage resources in the scope of the PaaS system. *Provider* is an entity delivering computing and storage resources for customers, represented by systems users with provider role given. SLAM allows users to create SLA drafts that express customer requirements for the underlying infrastructure such as computing volume restrictions, public IP restrictions for computing resources, as well as time and volume restrictions for storage resources. Each SLA is created for service available on a given site. The SLA draft is created by customer representative and subjected to provider's approval. The Provider is able to either agree or deny customer's proposal, as well as to reflect their own suggestions for the given restrictions. The customer reviews these suggestions. The process of negotiations continues until the consensus is met between the customer and the provider. Once SLA is signed it is made available for other systems to consume (Orchestrator). SLAM also allows to reflect customer preferences for the priority for the use of resources - in case specific sites or services should be used in favour to others. In order to access SLAs as well as customer preferences REST API was made available. *CPR* (Cloud Provider Ranker) is a component of the INDIGO PaaS system that ranks cloud provider services implementing a rule-based algorithm. It is a stateless microservice providing REST APIs to request the ranking. The ranking is performed considering SLAs targets and monitoring metrics for the different services; configurable rules are applied to the input data in order to

	normalize them and use specific weights. The CPR computes the scores for the different services and assigns a rank to each of them, thus returning an ordered list of cloud services. The Orchestrator consumes this service in order to obtain the ordered list of providers/services where it will try to schedule the deployment. The first of the list will be considered the "best" service for fulfilling the user request; the other ones will be considered in case of failure. *Monitoring* system is another component of the INDIGO PaaS system. Its architecture consists of three sub-components: the Zabbix server that stores the metrics, the Zabbix wrapper that provides a REST API interface to Zabbix to retrieve the metrics (e.g., this is used by the Orchestrator to get the monitoring data about the different services) and a set of Zabbix probes that collect the relevant metrics for the different cloud services.
Value proposition	high level orchestration of heterogeneous Cloud resources provisioning of virtualized compute and storage resources SLA negotiation
User of the service/tool	Users who want to deploy complex applications on multiple Cloud resources, with SLA support.
User Documentation	https://indigo-dc.gitbooks.io/indigo-paas-orchestrator/content/how_to_deploy.html
Technical Documentation	https://indigo-dc.gitbooks.io/indigo-paas-orchestrator/content REST APIs docs: https://indigo-dc.github.io/orchestrator/restdocs
Product team	INFN, Reply
License	Apache License 2.0: https://github.com/indigo-dc/orchestrator/blob/master/LICENSE
Source code	https://github.com/indigo-dc/orchestrator (main component) https://github.com/indigo-dc/cmdb https://github.com/indigo-dc/slam https://github.com/indigo-dc/CloudProviderRanker https://github.com/indigo-dc/monitoring
Testing	https://jenkins.indigo-datacloud.eu:8080/job/Pipeline-as-code/job/orchestrator

2.3.3.2 Release notes

In the latest stable release (2.4.0) the Orchestrator workflow has been improved in order to better manage failures: the retry mechanism is now fully supported for all the types of deployments, including containerized applications sent to Mesos clusters and, potentially, batch jobs sent to HPC sites through QCG Gateways [QCG].

Moreover, the mechanism has been extended in order to address the timeout in the deployment creation/update.

The user can specify:

- the maximum time for the single trial at each provider;

- the overall maximum time for the deployment creation (including the possible retries).

The support for hybrid deployments has been improved and consolidated as well. The CMDB has been extended to also publish the networking configuration and capabilities of the sites (e.g., if the site provides public IPs, if the site allows to create self-service private networks, etc.). This information is used by the Orchestrator in order to decide how to perform the deployment: the Orchestrator selects the sites to be used for the deployment depending on the scenario (with or without auto-provisioned L2 networks) and coordinates the allocation of the compute and networking resources accordingly.

With the previous version of the Orchestrator, the users could submit deployments to public cloud providers (e.g., Amazon EC2 or Azure) providing their credentials in the TOSCA template. This approach has been improved from the security point of view supporting the integration with a Secrets Manager, based on Hashicorp's Vault [VAULT]: it is used to securely store and access sensitive data like user credentials. The user credentials are no longer passed in the template as they are dynamically retrieved by the Orchestrator from the Vault.

The HPC plugin to manage the interaction with the QCG REST Gateway [QCG] exposed by HPC sites is now in its final shape; the Cloud Info Provider has been enhanced in order to collect relevant information about the QCG and SLURM services (as reference implementation) to be published in the CMDB and consumed by the Orchestrator; a new Monitoring probe has been implemented in order to monitor the QCG endpoint and collect metrics and health information to be consumed by the Orchestrator.

Finally, a Web Dashboard for the PaaS Orchestrator is now available with the aim of providing a user-friendly and immediate interface that does not require any technical knowledge about TOSCA and the PaaS.

[QCG] https://apps.man.poznan.pl/trac/qcg-computing

[VAULT] https://www.vaultproject.io/

2.3.3.3 Integration activities and opportunities

The Orchestrator is being extended to support EGI Checkin as Identity Provider along with the INDIGO IAM. This feature will be released soon and will allow federation, under the INDIGO PaaS, for IaaS sites that are integrated only with EGI check-in and do not support INDIGO IAM.

Moreover, due to the increasing need to support deployments on kubernetes, a dedicated plugin for this use-case is under development. It will allow the users to submit deployment requests to the federated kubernetes clusters in the same way as Mesos clusters are already managed.

Finally, the availability of the Orchestrator dashboard has greatly increased the interest for the PaaS that is now being used by different use-cases, including DODAS, PolicyCLOUD and OpenRiskNet.

2.3.3.4 Future plans

Continue with the maintenance of the service and integration with EOSC-hub services.

2.3.4 Future Gateway

2.3.4.1 Service Description

Service/Tool name	Future Gateway
Service/Tool url	https://futuregatewayframework.github.io
Service/Tool information page	https://github.com/FutureGatewayFramework/fgDocumentation
Description	FutureGateway is a set of components enabling end-users to access functionality such as authentication/authorization or deployment management. It exposes a REST interface to a service called fgAPIServer. The REST calls can be used to Create, Read, Update or Delete (CRUD) entities such as tasks to be executed on remote computing resources. For high-level actions (e.g., new task submission), fgAPIServer delegates to APIServerDaemon, a backend responsible for direct interaction with different middleware's. This solution is flexible to allow largely different scenarios to be handled. On one hand, fgAPIServer has been shown to work seamlessly with mobile applications, Java-based scientific workflows and with direct web-browser REST call generation. At the same time, APIServerDaemon handles single job submission as well as Virtual Machine deployment in a private cloud.
Value proposition	An easy-to-use frontend for complex backend scenarios such as: remote job submission or virtual machine and container deployment.
User of the service/tool	Research communities, individual researchers

User Documentation	https://github.com/FutureGatewayFramework/fgDocumentation/blob/master/usage.md
Technical Documentation	https://github.com/FutureGatewayFramework/fgDocumentation/blob/master/installation.md
Product team	INFN Catania, PSNC
License	Apache 2.0
Source code	https://github.com/FutureGatewayFramework https://github.com/indigo-dc/indigo-parent https://github.com/tzok/fg-docker-compose https://github.com/tzok/eosc-fgapiserver https://github.com/tzok/fgAPIServer https://github.com/tzok/grid-and-cloud-engine
Testing	A set of tests is available in the main repository: https://github.com/FutureGatewayFramework/fgAPIServer/tree/master/tests Additionally, the Docker Compose / Kubernetes version of the service has been validated with integration tests from the following repository: https://github.com/indigo-dc/indigoclient

2.3.4.2 Release notes

The FutureGateway released updated versions of its components, Docker images and configurations. The releases fixed issues present in the codebase or in the building process. In particular, security bug fixes were applied to the REST API component (fgAPIServer) and Maven configuration was updated for the backend component (APIServerDaemon). Also, the documentation and tutorials, mainly for Kepler scientific workflows, were improved.

In particular, the *FutureGateway* branch of the `tzok/grid-and-cloud-engine` fork has been updated to ignore iRODS dependencies. The iRODS Java artefacts are no longer available in the version required by JSaga library. FutureGateway did not use iRODS adapters to JSaga, so this removal of dependency has no impact on the functionality of this service.

Also, the `tzok/fgAPIServer` fork has been released under version v0.0.10.1. It contains several bug fixes. The most important is a security fix, which prevents information leak in one of the REST calls.

Finally, the `tzok/eosc-futuregateway` has been released under version 2.1. It contains build procedures and running instructions for a containerized version of the FutureGateway service. The 2.1 version includes fixes mentioned above.

2.3.4.3 Integration activities and opportunities

No integration activities in the reporting period.

2.3.4.4 Future plans

Continue maintenance and support for integration with other EOSC-hub services.

2.3.5 Summary

The PaaS Orchestrator is available through the EOSC Portal and a production instance for EOSC-hub is actively maintained. The integration with EGI Check-In now allows requests from users authenticated through this AAI framework. Additional enhancements, such as a Kubernetes plugin, the network selection in tenants and an enhanced visual improvement of the dashboard has been achieved. The Infrastructure Manager is also available through the EOSC Portal and a production instance is actively maintained. Enhancements to support the use cases have been implemented in the last reporting period, such as hybrid deployments across cloud, a new JavaScript-based Software Development Kit and additional support for GPUs in the OpenStack connector. The FutureGateway has undergone several releases for bug fixing and security improvements in the REST API, including better documentation especially for the Kepler scientific workflows. The goal of the task, which aimed at integrating the aforementioned components with the Cloud Compute and Cloud Container services has been achieved.

2.4 Data and Metadata Management

The EOSC-hub Data and Metadata management services provide a set of policy-driven data management/stewardship services, with particular regard to registered data (i.e., data associated to a persistent identifier). These services, which are described in the following sections, support FAIR data principles allowing datasets to be stored in a geographically distributed repository and enable the association of persistent identifiers with the data, making the data set's physical location independent from the logical references pointing to it. The identifier is globally resolvable, and the data set is replicated in multiple copies, which are tracked in the metadata associated with the identifier. Data can be published, and community specific metadata may be associated with it, this metadata can then be harvested and indexed by a discovery service to make the data findable. Data can also be annotated, manually or programmatically via API. And last, but not least, data are curated through a set of policies that each data manager can define.

2.4.1 B2HANDLE

2.4.1.1 Service description

Service/Tool name	B2HANDLE
Service/Tool url	Not applicable as it is a distributed service. Current nodes visible at https://dp.eudat.eu/operations/B2HANDLE
Service/Tool information page	https://eudat.eu/services/b2handle
Description	B2HANDLE is a distributed service, designed to contribute to data persistency by maintaining opaque, globally unique persistent identifiers (PIDs). PIDs are used in other user-facing e-infrastructure services such as B2SAFE and B2SHARE to reliably identify data objects over long periods of time, possibly beyond object lifetime, and thereby provide easy, stable references for use by service components and end-users. B2HANDLE identifiers bear metadata (kernel information), which provides the critical information about an identified object for use by other services. The B2HANDLE service offers management of identifier namespaces (Handle prefixes), supports object policies, and maintains stable business workflows to provide a reliable, trustworthy and scalable service. B2HANDLE also offers the Central PID Catalog, which enables reverse-lookups on B2HANDLE PIDs and simple searches/filtering. For this, B2HANDLE maintains the necessary software components. To simplify automated management of PIDs, B2HANDLE provides a set of Python libraries (b2handle, pyhandle). B2HANDLE also includes the HRLS component to provide reverse-lookups and searching, which is essential to run the Central PID Catalog.
Value proposition	B2HANDLE facilitates persistent identification of research assets, particularly data, independent from and throughout changes of current location or ownership. It provides a trustworthy, scalable, reliable service. Usage of PID Profiles, the Central PID Catalog and the B2HANDLE support libraries can facilitate efficient automated data management at early stages of the data life cycle.
Customer of the service/tool	Research infrastructures, disciplinary or generic, and their hosting institutions/service providers
User of the service/tool	Infrastructure services: end-users interested in identifying research assets before publication phases.
User Documentation	https://www.eudat.eu/services/userdoc/b2handle https://eudat.eu/services/userdoc/b2handle-for-communities https://eudat.eu/services/userdoc/b2handle-for-end-users

Technical Documentation	http://eudat-b2safe.github.io/B2HANDLE http://eudat-b2safe.github.io/PYHANDLE https://github.com/EUDAT-B2SAFE/B2HANDLE-HRLS
Product team	SURFsara, GRNET, SNIC/KTH, DKRZ
License	Source code available under Apache 2.0 Open Source license: https://github.com/EUDAT-B2SAFE/B2HANDLE/blob/master/LICENSE
Source code	https://github.com/EUDAT-B2SAFE/B2HANDLE https://github.com/EUDAT-B2SAFE/PYHANDLE https://github.com/EUDAT-B2SAFE/B2HANDLE-HRLS
Testing	Continuous integration testing is configured for B2HANDLE via Jenkins (GRNET)

2.4.1.2 Release notes

- Made pre-release available supporting Python3
- New server setup and numerous bug fixes were performed
- Updates to the required docker files to support the tests in various environments
- Updates to documentation

2.4.1.3 Integration activities and opportunities

- In the last year, the integration with the DataHub has been completed, remarkably interesting work and many demos have been done.
- Initiated collaboration with FAREA project[9]
- Integration with B2SHARE is in progress
- Added integration for the B2HANDLE code in the Eudat Gitlab: gitlab.eudat.eu.

2.4.1.4 Future plans

- Planning to upgrade the Handle server now ongoing.
- Discussion is still ongoing about how metadata and pid records may be improved (Low priority)
- Planning with respect to the maintenance of the currently independent b2handle / pyhandle code bases: it would be better to concentrate on one generic code generic base (pyhandle) and to extend this to include EUDAT/b2handle extensions. Yet this has implications on the currently operationally deployed b2handle instances.

[9] https://www.project-freya.eu

2.4.2 B2SAFE

2.4.2.1 Service description

Service/Tool name	B2SAFE
Service/Tool url	Not Applicable: it is a distributed service. See the number of instances listed in the status report here: http://avail.eudat.eu/lavoisier/status_report-site?accept=html
Service/Tool information page	https://www.eudat.eu/services/b2safe
Description	B2SAFE is a highly available multi-purpose service that allows community repositories to implement data management policies on their research data that is distributed across multiple administrative domains. The following components are part of B2SAFE. The service core: it offers functionality for the long-term data preservation. The main feature is the function to replicate data sets across different data centres in a safe and efficient way while maintaining all information required to easily find and query information about the replica locations. The information about the replica locations and other important information is stored in a PID registry (B2HANDLE). The B2SAFE Service is implemented as a package on top of iRODS, providing a set of iRODS rules and scripts. Data transfer: B2SAFE supports different transport protocols in combination with B2STAGE. It provides endpoints for those protocols and enable third-party transfers. Data Policy Management: data management policies are abstract descriptions of data management, stewardship and curation tasks. These policies, which are stored in a database and handled via a Data Policy Manager (DPM), translate into concrete operational tasks that are event-triggered and executed within the B2SAFE rule engine for instance when data is ingested, stored and must be curated or transferred.
Value proposition	For the communities who need to guard against data loss; a customer facing service that allows data replication and policy-driven data management using different storage services provided by geographically distributed centres in the EUDAT CDI.
User of the service/tool	Teams of researchers or single users

User Documentation	https://eudat.eu/services/userdoc/b2safe https://eudat.eu/services/userdoc/configure-b2safe http://eudat.eu/services/userdoc/using-b2safe https://eudat.eu/services/userdoc/joining-b2safe
Technical Documentation	https://github.com/EUDAT-B2SAFE/B2SAFE-core/wiki https://github.com/EUDAT-B2SAFE/pam-oauth2 https://github.com/EUDAT-B2SAFE/B2SAFE-DPM
Product team	eudat-safereplication@postit.csc.fi
License	https://github.com/EUDAT-B2SAFE/B2SAFE-core/blob/master/LICENSE
Source code	https://github.com/EUDAT-B2SAFE/B2SAFE-core https://github.com/EUDAT-B2SAFE/pam-oauth2 https://github.com/EUDAT-B2SAFE/B2SAFE-DPM
Testing	A test suite, which is executed manually is available here: https://github.com/EUDAT-B2SAFE/B2SAFE-core/tree/master/scripts/tests

2.4.2.2 Release notes

During 2020 the new B2SAFE-core 4.3.0 pre-release has been released. Moreover, porting to the new KIT Devops infrastructure based on GitLab has been concluded (https://gitlab.eudat.eu). The new release is publicly available now in fact at: https://gitlab.eudat.eu/b2safe/B2SAFE-core/-/releases. Docker image as well as rpm files have been published on the new artifact repository based on JFrog at Sara:

- RPM: https://artie.ia.surfsara.nl/ui/repos/tree/General/Eudat-RPM-Production-Public%2FCentos%2F7%2Firods-4.2.8%2Fmaster%2Fnoarch%2FPackages%2Firods-eudat-b2safe-4.3.0-0.noarch.rpm
- Docker images: https://artie.ia.surfsara.nl/ui/repos/tree/General/eudat-docker-public

Porting of all B2SAFE related components is going to be concluded in the near future.

The migration to the new GitLab-runner infrastructure at KIT will be completed (old Sara installation will be removed).

2.4.2.3 Integration activities and opportunities

Integration of B2SAFE-core in the new gitlab.eudat.eu repository has been completed. Currently the process of migrating the GitLab-runner daemons from the Sara cluster to KIT is underway. B2SAFE-core code has been refactored and cleaned-up before the pre-release, old branches have been pruned and new versions tagged in the gitlab.eudat.eu repository. New merges include Python3 porting, bug fixing and integration of pid microservices features. Now working hard on the integration of the b2safe-b2share iBridge component (which, at the time of writing, is close to completion) into the B2SAFE-core master branch.

Collaboration with the Compbiomed community has started as the B2SAFE-core team is actively supporting them, moreover a use case document has been created including a description of the data synchronization mechanism between sites.

2.4.2.4 Future plans

In the near future the development team will focus on completing the setup of the gitlab.eudat.eu Devops environment and starting to test the integration of b2safe-b2share iBridge. Finally, the final B2SAFE--core release will be produced and made publicly available. Collaboration with Compbiomed will move ahead and other communities are expected to join.

2.4.3 B2SHARE

2.4.3.1 Service description

Service/Tool name	B2SHARE
Service/Tool url	https://b2share.eudat.eu
Service/Tool information page	https://www.eudat.eu/services/b2share
Description	Repository for shareable digital objects to improve your data sharing and publishing and guarantee long-term persistence of your locally stored data. Increased findability and discoverability.
Value proposition	For the individual researchers who do not have adequate facilities for storing, preserving and sharing data, B2SHARE service is a customer-facing service which provides a safe repository for scientific data and an easy way to share it in the research community. For research communities that want to centrally manage publications and enforce policies like review workflows.
User of the service/tool	Individual researcher, scientific community, scientific institution (e.g., university)
User Documentation	https://eudat.eu/services/userdoc/b2share-usage https://www.eudat.eu/b2share-training-suite https://github.com/EUDAT-Training/B2SHARE-Training
Technical Documentation	https://github.com/EUDAT-B2SHARE/b2share https://www.eudat.eu/b2share-training-suite
Product team	CSC - IT-center for Science
License	GPLv2

Source code	https://github.com/EUDAT-B2SHARE/b2share
Testing	Training instance of B2SHARE where users can try and test out functionality provided by B2SHARE is available at: https://trng-b2share.eudat.eu Automatic unit test suite can be found at: https://github.com/EUDAT-B2SHARE/b2share/tree/master/tests

2.4.3.2 Release notes

Version 2.1.5 has been released which improves user experience and optimizes development and builds of new releases.

Notable updates include:

- Updated B2NOTE widget

2.4.3.3 Integration activities and opportunities

Effort has been put into integrating the updated B2NOTE widget and making sure (the number of) annotations of records and files are better visible to the user in landing pages of the web interface.

2.4.3.4 Future plans

Work will continue on improving B2NOTE integration, updating the root metadata schema and expanding community possibilities. With these updates many of the community requests can be complied with.

2.4.4 B2NOTE

2.4.4.1 Service description

Service/Tool name	B2NOTE
Service/Tool url	https://b2note.eudat.eu
Service/Tool information page	https://eudat.eu/catalogue/B2NOTE
Description	Data annotation service to enrich the description of shared/published datasets without modifying the underlying metadata schemata. Service allows to add semantic tags coming from existing semantic resources (controlled vocabularies, thesauri, ontologies), user-defined keywords and comments that could be used to curate, enrich and retrieve data elements of relevance. Improve reusability and findability. The service is offered in two forms: 1. Widget integrated in web pages of repository services

	2. Stand-alone application
Value proposition	For the individual researchers who are reusing shared/published data for their scientific project, B2NOTE service is a customer-facing service which integrates with data repositories and data services user interface to provide a simple tool for annotating data with semantic tags, keywords or comments and reusing these annotations to retrieve and aggregate relevant data elements across heterogeneous and distributed data repositories and services. For research communities that want to foster the reuse of their datasets, extend the metadata description with semantics, curate datasets.
User of the service/tool	Individual researcher, scientific community, scientific institution (e.g., university)
User Documentation	User manual included in the Widget making it directly accessible from any service which embeds the widget.
Technical Documentation	https://e-sdf.github.io/b2note-docs
Product team	e-Science Data Factory
License	MIT licence
Source code	https://github.com/EUDAT-B2NOTE/b2note
Testing	Sandbox instance at: https://b2note.bsc.es Integrated with the training instance of B2SHARE where users can try and test out functionality provided by B2NOTE is available at: https://trng-b2share.eudat.eu Automatic unit test suite can be found at: https://travis-ci.org/EUDAT-B2NOTE/b2note

2.4.4.2 Release notes

The current major release is v3.x that is a complete rewrite of the original v1 using up-to-date technologies and practices and providing additional features and improved user experience. v2 was an intermediate experimental version that was released as development.

2.4.4.3 Integration activities and opportunities

B2NOTE Widget is integrated in the B2SHARE service and OpenAIRE Explore and Dashboard. Thanks to the IFRAME integration, the integration work is relatively easy at the side of service providers with no maintenance needs (automatic updates). The Widget communicates with the hosting service through JavaScript messaging, which enables login token exchange and event notifications. As such, it is possible to automatically log in the user into the Widget. The B2NOTE API then enables the service to make queries about annotations, e.g., to indicate existing annotations on their page.

2.4.4.4 Future plans

B2NOTE sustainability will be ensured by e-Science Data Factory where it is being turned into a product branded as “Semaphora” -- various use-cases and segments are explored, as well as new deployment targets (browser plugins).

2.5 Data preservation

This task aims at integrating certified Trusted Digital Repository (TDR) in the EOSC-hub catalogue, resulting in a sustainable long-term data preservation service: the European Trusted Digital Repository (eTDR).

2.5.1 eTDR

2.5.1.1 Service description

Service/Tool name	eTDR
Service/Tool url	https://www.cines.fr/en/europe/eudat-cdi/etdr/
Service/Tool information page	https://marketplace.eosc-portal.eu/services/etdr-european-trusted-digital-repository
Description	eTDR is a service that ensures digital information remains findable, accessible, interoperable and reusable over time. This includes capacity and resource planning, as well as application of long-term preservation techniques and technologies. It also combines policies, processes and actions to ensure access to "born-digital" and re-formatted data, regardless of the challenges of technological changes or failures (metadata, file format, media).
Value proposition	The service relies on mature, “ready-to-use” EUDAT services, which provide data transferring functionalities. These include B2SAFE, B2HANDLE, B2FIND.
Customer of the service/tool	Communities willing to preserve large collections of datasets over time, with no limit of duration.
User of the service/tool	Institutions representing a research community
User Documentation	https://drive.google.com/drive/folders/1MZehVTZ9aGc5s0xdkqovVEY7TLEDy5Mi
Technical Documentation	N/A
Product team	CINES
License	N/A

Source code	N/A
Testing	N/A

2.5.1.2 *Release notes*

No specific release notes are available from the eTDR internal services which is provided as a hosted service by CINES. The Interfaces to the services which are seen by EOSC-hub users (B2SAFE, B2HANDLE, B2FIND) are detailed elsewhere in this document.

2.5.1.3 *Integration activities and opportunities*

In close consultation with SURF, DANS has worked on the integration of signposting[10] in B2HARE. Signposting is an approach to make the scholarly web more friendly to machines. It uses *Typed Links* as a means to clarify patterns that occur repeatedly in scholarly portals. For resources of any media type, these typed links are provided in *HTTP Link headers*. For HTML resources, they may additionally be provided in *HTML link elements*.

The signposting module implements part of signposting, level 2. The module creates the linkset endpoint based on the metadata of the dataset. The linkset can be added to the HTTP header. With signposting implemented it will become possible to deposit a single dataset at any repository which implements signposting. Via a pull request, the dataset can be pulled to the archive for processing and depositing.

With the Proof of Concept of the signposting module in B2SHARE DANS has shown that DANS is able to implement signposting in B2SHARE. The basic principles of signposting have been implemented, i.e., adding signposting via linkset link to the landing page.

Since B2SHARE will upgrade to a new version in 2021 some adjustments, which fall under the responsibility of B2SHARE, are required when the pull request is accepted. Additionally, after the upgrade to Invenio 3, the B2SHARE development team will need to create an “archive to DANS” button to enable this functionality. These actions have been described in the report ‘Signposting in B2SHARE’[11].

2.5.1.4 *Future plans*

The business model for the CINES instance of eTDR has been published:

https://docs.google.com/document/d/1uCcSzkuPBBXse-ulU-uJ-9zdfAi-5W0P/edit?rtpof=true. DANS will conduct a similar exercise. Besides, also DANS started to change its technical infrastructure in 2020 and will continue to do so in 2021. The implementation of the signposting module in B2SHARE will lead to a new business model.

[10] https://signposting.org/

[11] https://docs.google.com/document/d/1ThsKbZValDQMzHtopJ0h_VvqHkD0E5lKqdT1rbfAjc/edit?ts=602106a1

2.6 Sensitive Data Services

2.6.1 TSD

2.6.1.1 Service description

Service/Tool name	Services for Sensitive Data (TSD)
Service/Tool url	https://www.uio.no/english/services/it/research/sensitive-data
Service/Tool information page	https://www.uio.no/english/services/it/research/sensitive-data/about/
Description	TSD offers storage capability, computing infrastructure, analysis / visualization platforms and web-based data collection tools suitable for running complex research projects in a secure IT-infrastructure.
Value proposition	Desktop with secure storage and software to enable the collection and analysis of sensitive data.
Customer of the service/tool	Researchers, Research group, Business, Research organisations
User of the service/tool	Researchers dealing with sensitive data
User Documentation	https://www.uio.no/english/services/it/research/sensitive-data/use-tsd/usermanual/
Technical Documentation	https://www.uio.no/english/services/it/research/sensitive-data/about/description-of-the-system.html
Product team	Department of Research Support Services https://www.usit.uio.no/english/about/organisation/rc/rss/index.html
License	Parts available under BSD/MIT license
Source code	https://github.com/unioslo/tsd-file-api https://github.com/unioslo/tsd-api-client https://github.com/unioslo/pg-iam https://github.com/unioslo/pypg-iam https://github.com/unioslo/tsd-s3cmd
Testing	https://www.uio.no/english/services/it/research/sensitive-data/use-tsd/administrative-tasks/administer-users/#apply-for-a-test-user

2.6.1.2 Release notes

In close collaboration with CSC, who are planning to launch a similar service, a SecureB2SHARE pilot service has been made available at UiO. The TSD variant will implement the common design principles of the SecureB2SHARE framework, but adapt it for use with TSD APIs, features and regulations.

- Integrates TSD and B2SHARE services
- B2SHARE provides metadata catalogue
- Sensitive datasets are kept in TSD secure data storage
- Users can request access via a web form with submission to TSD, provided by UiO's Nettskjema service
- Authorisation decisions are made by data owners in a web portal accessible within TSD
- Users that have had their requests granted will be able to access sensitive datasets via the TSD publication system.

2.6.1.3 Integration activities and opportunities

UiO/TSD will continue collaborating on the development of further integrations for SecureB2SHARE.

2.6.1.4 Future plans

Following user feedback from the pilot service, work towards getting SecureB2SHARE into production.

2.6.2 ePouta

2.6.2.1 Service description

Service/Tool name	Secure and cost-effective cloud computing for processing sensitive data (ePouta)
Service/Tool url	https://research.csc.fi/epouta
Service/Tool information page	https://research.csc.fi/cloud-computing
Description	This service provides infrastructure as a-service for running analysis on sensitive data. The ePouta Virtual Private Cloud service allows customers to provision virtual machines and storage resources directly to their own internal networks. It provides an easy-to-use admin web interface and a programmable API for managing virtual machines, networks and storage. CSC ePouta is targeted for sensitive data processing.
Value proposition	IaaS for sensitive data processing combined with local secure data storage.

Customer of the service/tool	Researchers, Research organisations, Research group, Providers
User of the service/tool	Researchers dealing with sensitive data
User Documentation	https://docs.csc.fi/cloud/pouta/
Technical Documentation	Not available publicly. Please contact the named contact person for more technical details of the service.
Product team	CSC
License	N/A
Source code	N/A
Testing	N/A

2.6.2.2 Release notes

During the reporting period most of the ePouta related work has focused on drafting integration plans for SecureB2SHARE pilot service. Note that a similar SecureB2SHARE pilot service utilizing TSD services (See section 7.1.1) has been implemented jointly by CSC and UiO during the EOSC-hub project. SecureB2SHARE utilizing ePouta will share common design principles with the implementation utilizing TSD services but will utilize ePouta specific APIs for authorization management and data access.

Currently SecureB2SHARE pilot service utilizing ePouta has been planned in the following way.

- SecureB2SHARE pilot service integrates ePouta and B2SHARE services.
- Dedicated B2SHARE instance will be used to store non-sensitive metadata of sensitive dataset.
- ePouta will be used to store the actual sensitive dataset and possible sensitive metadata of the dataset.
- Users will discover and request access to the dataset from SecureB2SHARE by filling in an access request form.
- Authorization decisions will be made in ePouta, utilizing existing REMS tool developed by CSC.
- Users that have been given authorization to access a sensitive dataset will access the dataset through remote desktop connection provided by ePouta. Alternatively authenticated users can use an API token to access the dataset.

2.6.2.3 Integration activities and opportunities

SecureB2SHARE plans drafted during the EOSC-hub project will be utilized in a new project, where SecureB2SHARE (and as such ePouta) will be further integrated to existing Galaxy portal service provided by CINECA.

2.6.2.4 Future plans

Development of SecureB2SHARE pilot service utilizing ePouta will continue in future projects. In addition to implementing and offering the pilot service, other activities aiming to make the pilot service into a more production ready must be done. For example, deployment process of ePouta utilizing SecureB2SHARE will be further adjusted, and risk analysis will be performed.

3 Global Integration activities

This section highlights the global, high-level, use-cases which have been addressed during the EOSC-hub project and reports on activities which were undertaken during the last phase of the project, after the release of deliverable D6.4. Each use-case is community driven although the integration activities themselves which are derived from these use-cases are valid for several communities. By addressing such integration issues compound solutions are provided for projects which reduce the workload on the end users, and which can be recycled for other communities/projects. This approach to integration activities, addressing high level cross-domain issues, brings much value to the EOSC-hub services. These activities ensure that EOSC-hub is more than the sum of its parts and greatly simplify the use of EOSC-hub Services for the communities and end-users alike. The driving use-cases are defined in Section 4.1, these use-cases were reported in D6.4 and are included here for the sake of completeness, they have been used to define much of the integration activities. The description of the use-cases is followed by a detailed description of the integration activities per thematic area in section 4.2. The information reported here focuses on the period since deliverable D6.4 was released, roughly corresponding to the last project year.

3.1 Integration Use Cases

The following use cases provided requirements for the common services. Those requirements have been collected according to a common template and stored in the EOSC-hub wiki (https://wiki.eosc-hub.eu/display/EOSC/Community+requirements+DB). Following the initial collection of the use-cases WP10 performed an analysis of each use-case and together with WP6 developers a series of meetings was held with the community representatives to define appropriate solutions. Each use-case is tracked via the EOSC JIRA system.

The use cases related to data management activities are collected from mainly five sources:

- Thematic Services.
- Competence Centers.
- Communities already using EUDAT/EGI/Indigo services.
- New communities entering EOSC.
- Low hanging fruits identified by EOSC-hub service providers about the integration among common or federated services.

This last point is the only one not directly related to user requirements. It encompasses integration activities, which can offer new features, like the interoperability between two data services, potentially interesting for the users and achievable by the service developers with a limited effort.

3.1.1 ECAS: Perform analysis on remote large volume climate data

The ENES Climate Analytics Service (ECAS) offers scientific users a set of tools to perform data analysis experiments on large volumes of multidimensional data, using parallel processing workflows on remote systems without needing to download data.

3.1.1.1 Integration of B2DROP to store and exchange output of ECAS output

B2DROP can be used within different parts of ECAS.

The first integration aspect of ECAS with B2DROP concerns the Ophidia workflow framework. The framework was extended by a custom operator which stores the workflow output in the B2DROP account of the user. To use this operator, the user creates an app password within B2DROP and stores this in the ECAS user space. The credentials and the files to be uploaded to B2DROP are then configured as part of calling the operator. To upload the files to the B2DROP space of the user, the B2DROP space is mounted locally using the WebDAV protocol.

The second part of ECAS that is integrated with B2DROP is JupyterHub. JupyterHub is a web-based framework for execution of Jupyter notebooks on remote resources. The deployment of JupyterHub within ECAS was configured and extended to access two different kinds of B2DROP storage. The first one is a shared space for all ECAS users and does not need further authentication of the users, thus facilitating easy exchange of files between ECAS users. The second storage is the user's private B2DROP space. This space is only visible and accessible for the owner after authentication. To use the private B2DROP space, the user creates an app password within B2DROP and stores the credentials in the environment file of the notebooks. Akin to the Ophidia operator, the B2DROP space is mounted locally via the WebDAV protocol. For the usage of the B2DROP space the graphical interface was extended with two buttons. The buttons are called "Share" and "Move". The "Share" button copies the notebook to the shared space and the notebook is shared with all other ECAS users. The "Move" button copies the notebook to the private or the shared B2DROP space, specified by the user. Instead of moving the notebooks, the users can create them directly within the B2DROP space, too.

For more information see deliverable D7.2 chapter 4.

Service: B2DROP, B2ACCESS and IAM

Resource Providers: DKRZ and CMCC

Resources: allocation of B2DROP storage depending on use case

Use case requirements:

- B2DROP account (optional, only if sharing with other ECAS users is desired)
- Input data must reside in any data source supported by ECAS, e.g., B2DROP or community store (OpenDAP)
- Output data must not exceed user quota of B2DROP

3.1.1.2 Integration of EGI FedCloud resources with ECAS

The EC3 AoD platform has been extended to enable exploiting the EGI Cloud Compute service to deploy on demand ECAS elastic clusters. In that way a user can deploy an ECAS cluster following a simple wizard where the main parameters (CPU and memory of the nodes, number of nodes of the cluster, etc.) can be selected. Finally, the actual interaction with the infrastructure is delegated to the Infrastructure Manager (IM).

The integration of ECAS in the cloud-based resources provided by EGI allows users to easily deploy a full ECAS elastic cluster (composed of multiple nodes) in the cloud resources of the EGI Federation customized to their requirements. The EC3 service will take care of automatically installing and configuring the whole ECAS environment stack, including services and tools such as JupyterHub, PyOphidia, a rich set of data science Python libraries, the Ophidia HPDA framework, as well as a comprehensive set of Jupyter Notebooks for training. Furthermore, the internal elasticity manager (CLUES) automatically grows or shrinks the size of the cluster based on its workload.

Service: EC3, IM and FedCloud

Resource Providers: CMCC and FedCloud partner

Resources:

- Allocation of a set of VMs (from 2 to 12) with at least 2 CPUs and 4 GB or RAM in FedCloud sites.

Use case requirements:

- Launch ECAS virtual elastic cluster over FedCloud resources

3.1.2 Marine Competence Center use cases

The Marine Competence Center[12] shows interest in the integration of B2DROP and B2STAGE for the two use cases described below. The use cases are not dependent on each other, but rather complementary.

3.1.2.1 Processing measurement data and share processed data for collaborative analysis.

Regarding measurement data, coming from Argo floats. To establish a workflow consisting of the following phases: upload raw data, process the raw data, generate processed data, upload processed data, collaborate on the analysis of the processed data, possibly publish the final results and reports in relevant formats for open access.

The raw data is incrementally uploaded once a day, consisting of both new and corrected/curated data. Therefore, the raw data needs to be processed daily to distinguish any difference to the processed data. The range of the incremental raw data is one calendar month. The size of the monthly raw data is ~ 2 GB. The size of total raw data is ~ 300 GB.

Processing of the raw data is a time-consuming event and should be carried out in a batch or asynchronous manner. Therefore, the CC would like to have a notification when the process has finished.

The identified components for enabling these steps are: B2STAGE for uploading/transferring raw data to storage, B2SAFE for storage of raw data, Apache Spark for data intensive computation tasks, FedCloud for running the compute and analysis applications, B2DROP for collaborative analysis on shared data, Jupyter Notebooks for interactive analysis, B2ACCESS for authentication and B2SHARE for publishing the final result data and associated publications and reports.

[12] https://wiki.eosc-hub.eu/display/EOSC/T8.3+Marine

Service: B2STAGE, B2SAFE, FedCloud, B2DROP, B2ACCESS and B2SHARE

Resource Providers: CSC, CINECA, Jülich and FedCloud partner

Resources:

- Allocation of 300 GB storage in B2SAFE
- Processed data to share must not exceed user quota of B2DROP

Use case requirements:

- B2DROP and B2SHARE accounts
- Output data must not exceed user’s quota on B2DROP
- Optional: Notification of user when processing finished

3.1.2.2 User applications in a Virtual Research Environment

Regarding a virtual research environment to establish a web-based platform able to host a range of scientific applications. The application instances are launched per user-on-user demand. The scientific applications could be used to e.g., analyse the processed data in "*Use case 1*". Some applications are single-component, others are client-server where the server could be shared among users with the client being per user. Many of the applications are memory-bound, some requiring up to 8 CPU and 16 GB RAM per user. Yet other applications require as little as 1 CPU and 4 GB RAM. Most require no GPU-capabilities. The analysis is often interactive serial, rarely batch parallel, therefore traditional HPC/HTC computing cluster are not relevant. The duration of the sessions/runs are often not known beforehand, and the sessions should not be killed pre-emptively losing user data. Cloud-provided resources are most suitable for this kind of dynamic/elastic purposes.

The user's saved data should be accessible across applications in near real-time, as well as accessible from a central user interface for managing.

The identified components for enabling these could be: B2ACCESS for authentication, B2DROP for syncing the data between applications and collaborating on, FedCloud for providing the VRE infrastructure, Kubernetes for orchestrating the VRE system and application containers, and Jupyter Notebooks for common and interactive analysis.

For more information see deliverable D7.2 chapter 4.

Service: B2DROP, FedCloud, Kubernetes and B2ACCESS

Resource Providers: CSC, CINECA, Jülich and FedCloud partner

Resources: allocation of memory and compute power, depending on the application up to 8 CPUs and 4 GB RAM

Use case requirements:

- B2DROP account
- Data must be accessible across applications via a central user interface
- Output data must not exceed user’s quota on B2DROP

3.1.3 ICEDIG/Herbadrop use case: Digitisation infrastructure test on EUDAT

The Herbadrop use case comes from a data pilot in the EUDAT project aiming at 'an innovative approach to long-term preservation and analysis of digitised herbarium specimens'[13] and is now being treated in the EU-funded project ICEDIG[14]. The project milestone 'Digitisation infrastructure test on EUDAT' is described in detail in the pdf in the Appendix[15].

Herbadrop's archive comprises 27 TB of data volume on the B2SAFE instance at CINES. The objective of the ICEDIG data pilot is to develop the premise of the future ETDR (long-term European certified Trustworthy Digital Repository), in which CINES is involved. Thanks to services such as B2FIND or community portals in interaction with ETDR, FAIR data which is preserved and curated into the ETDR infrastructure would be accessible for non-profit users in CINES open data portal as well as it would be searchable and accessible via EUDAT B2FIND.

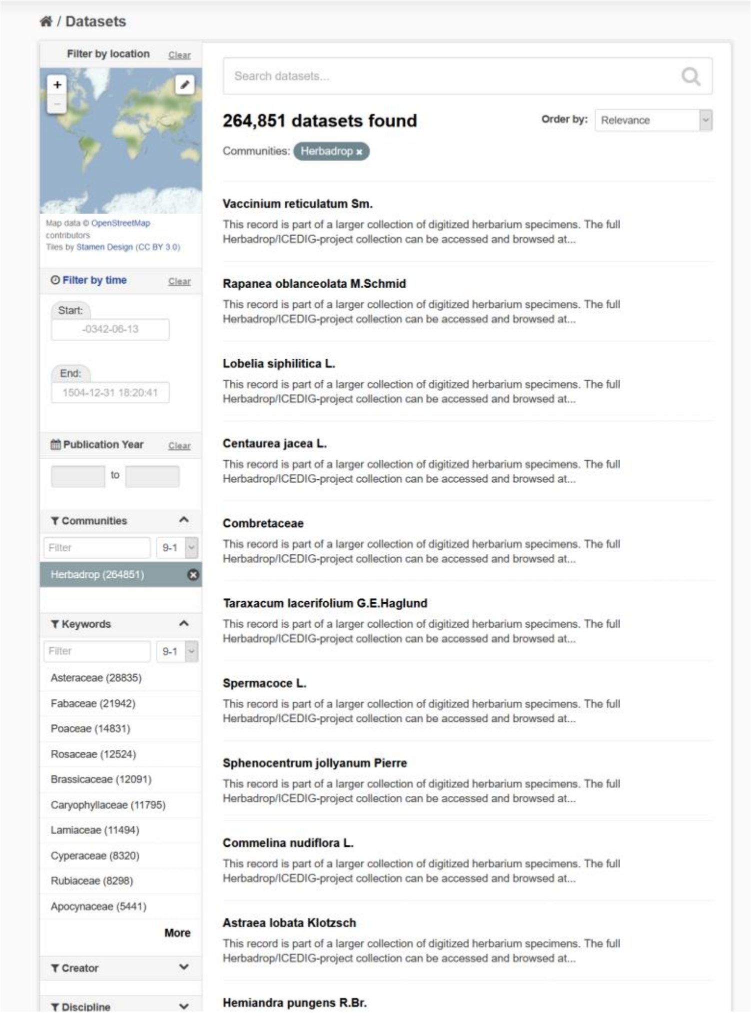
The ICEDIG architecture is split into a sequence of functions that processes one-step of the workflow. The image replication operation uses the EUDAT B2SAFE service. B2HANDLE is required for PID (Persistent Identifier) generation and then to guarantee data access through the B2FIND portal. The B2FIND portal and API provide users with advanced search functionalities and allow access to the data resources associated to the metadata found in the catalogue. EUDAT retrieves the metadata in Elasticsearch with our HTTP-API and feeds in pull mode the B2FIND portal for the herbarium images deposited on the ICEDIG platform. The access to data is then made possible through a webdav proceeding without authentication on the iRODS data node at CINES.

Herbadrop is visible in B2FIND as a Community, which means that a search request may start with showing all records that are offered by Herbadrop. To narrow down a search, e.g., for certain plant families, the facet <Keywords> may be used (figure 5). The detailed search result page offers a direct link to the institution maintaining the digital objects as visualised in figures 6 and 7.

[13] https://www.eudat.eu/herbadrop-an-innovative-approach-to-long-term-preservation-and-analysis-of-digitised-herbarium

[14] ICEDIG stands for "Innovation and consolidation for large scale digitisation of natural heritage" https://www.icedig.eu/

[15] Appendix: Milestone MS39_ ICEDIG_Digitisation infrastructure test on EUDAT_v1.pdf, see at https://wiki.eosc-hub.eu/download/attachments/26416995/Milestone%20MS39_%20ICEDIG_Digitisation_infrastructure_test_on_EUDAT_v1.pdf?api=v2



🏠 / Datasets

Filter by location Clear

Map data © OpenStreetMap contributors
Tiles by Stamen Design (CC BY 3.0)

Filter by time Clear

Start: -0342-06-13

End: 1504-12-31 18:20:41

Publication Year Clear

to

Communities

Filter 9-1

Herbadrop (264851)

Keywords

Filter 9-1

Asteraceae (28835)
Fabaceae (21942)
Poaceae (14831)
Rosaceae (12524)
Brassicaceae (12091)
Caryophyllaceae (11795)
Lamiaceae (11494)
Cyperaceae (8320)
Rubiaceae (8298)
Apocynaceae (5441)

More

Creator

Discipline

Search datasets...

264,851 datasets found

Order by: Relevance

Communities: Herbadrop ×

Vaccinium reticulatum Sm.
This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at...

Rapanea oblanceolata M.Schmid
This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at...

Lobelia siphilitica L.
This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at...

Centaurea jacea L.
This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at...

Combretaceae
This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at...

Taraxacum lacerifolium G.E.Haglund
This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at...

Spermacoce L.
This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at...

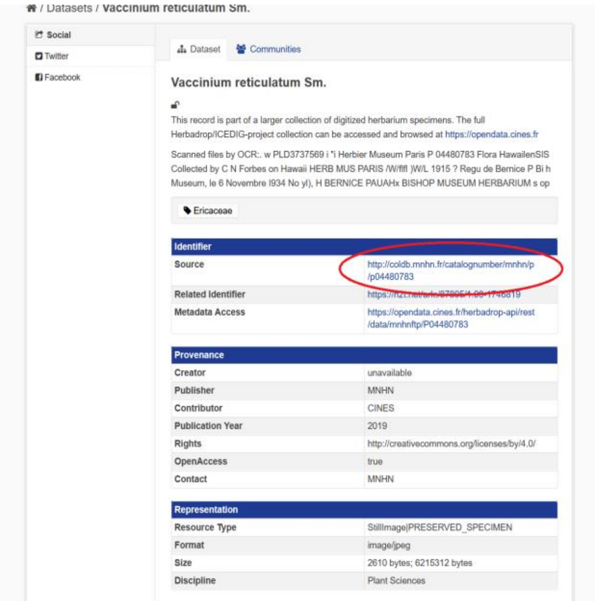
Sphenocentrum jollyanum Pierre
This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at...

Commelina nudiflora L.
This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at...

Astraea lobata Klotzsch
This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at...

Hemiandra pungens R.Br.

Figure 5: Datasets found for Community 'Herbadrop' and optionally narrowing down by choosing a plant family from facet Keywords



🏠 / Datasets / **Vaccinium reticulatum Sm.**

Social
Twitter
Facebook

Dataset Communities

Vaccinium reticulatum Sm.

This record is part of a larger collection of digitized herbarium specimens. The full Herbadrop/ICEDIG-project collection can be accessed and browsed at https://opendata.cines.fr

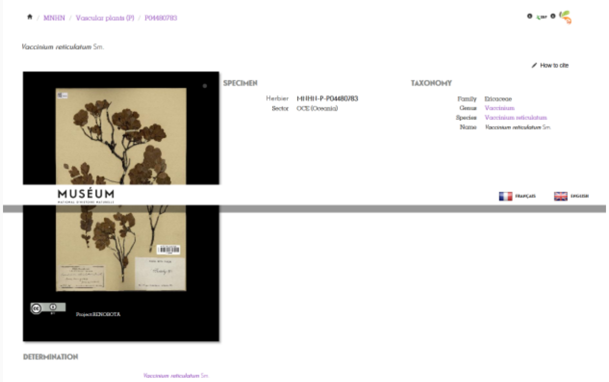
Scanned files by OCR:. w PLD3737569 i "i Herbier Museum Paris P 04480783 Flora HawailenSIS Collected by C N Forbes on Hawaii HERB MUS PARIS /W/flfl)W/L 1915 ? Regu de Bernice P Bi h Museum, le 6 Novembre I934 No yl), H BERNICE PAUAHx BISHOP MUSEUM HERBARIUM s op

Ericaceae

Identifier	
Source	http://coldb.mnhn.fr/catalognumber/mnhn/p/p04480783
Related Identifier	https://n2t.net/ark:/87895/1.90-1746819
Metadata Access	https://opendata.cines.fr/herbadrop-api/rest/data/mnhnftp/P04480783

Provenance	
Creator	unavailable
Publisher	MNHN
Contributor	CINES
Publication Year	2019
Rights	http://creativecommons.org/licenses/by/4.0/
OpenAccess	true
Contact	MNHN

Representation	
Resource Type	StillImage|PRESERVED_SPECIMEN
Format	image/jpeg
Size	2610 bytes; 6215312 bytes
Discipline	Plant Sciences



🏠 / MNHN / Vascular plants (P) / P04480783

Vaccinium reticulatum Sm.

How to cite

MUSÉUM
NATIONAL D'HISTOIRE NATURELLE

FRANÇAIS ENGLISH

Project RENOBOTA

SPECIMEN

Herbier	MHHN-P-P04480783
Sector	OCE (Oceania)

TAXONOMY

Family	Ericaceae
Genus	Vaccinium
Species	Vaccinium reticulatum
Name	*Vaccinium reticulatum* Sm.

DETERMINATION

Vaccinium reticulatum Sm.

Figure 6 and 7: As Identifier in this case by the field Source the direct link to the data resource is provided

Service: B2SAFE, B2HANDLE, B2FIND

Resource Providers: CINES

Resources:

- Allocation of 27 TB of data volume on the B2SAFE instance at CINES
- B2HANDLE prefix to register PIDs

Use case requirements:

- B2SAFE instance at CINES
- Access to Elasticsearch repository via HTTP-API by B2FIND
- Workflow for generation of PIDs and WebDav-URLs referring the iRODs collections

3.1.4 WeNMR use case

WeNMR is a worldwide e-Infrastructure for NMR spectroscopy and Structural biology. It is the largest Virtual Organization in the Life sciences and is supported by EGI.

Through integration with EGI DataHub, WeNMR users will be able to access a data space provisioned through EGI DataHub and its underlying platform Onedata, through the West-Life Virtual Folder. "Virtual Folder" provides a unified access mechanism to files stored in a variety of locations including the local file system and cloud storage facilities.

Oneprovider enables several means for integration with other services including: REST API, CDMI (Cloud Data Management Interface) and POSIX. The integration with Virtual Folder will be based on the POSIX Fuse mountpoint enabled by Oneclient command line tool.

Service: EGI-DataHub, “Virtual Folder”

Resource Providers: EGI, CYFRONET

Resources:

- Onezone and Oneprovider EGI-DataHub instances deployed at CYFRONET

Use case requirements:

- POSIX Fuse mount point enabled by Oneclient
- Transparent POSIX access to files on remote storages

3.1.5 CompBioMed data replication use case

CompBioMed is a European commission H2020 funded Centre of Excellence focused on the use and development of computational methods for biomedical application. The data-intensive workflows and distributed international partners involved in the project urges the use of proper data management solutions for handling the data. Safe data replication and large data transfer is one of the major requirements within the community. Initial activities focused on a use case to replicate data from BSC (Barcelona Supercomputing Centre) to SURFsara (Netherlands) and EPCC (UK) using the EUDAT B2SAFE service. Once the replication service is deployed and configured, it is expected that terabytes of data will be replicated between the HPC centers which facilitates large data exchange and access to valuable data for researchers in this community.

Service: EUDAT B2SAFE service

Resource Providers: BSC, SURFsara, EPCC

Resources: allocation of at least 24 TB storage at each of the HPC centers

Use case requirements:

- Data to be replicated is 3D finite element mesh (file format can be. vtk, .txt).
- The maximum size per file is 1.2 TB.
- The total data to be replicated is 24 TB.
- Two copies of replicas are desired, one on the compute facilities to run simulations and one on tape.
- The data owner assesses the replicas.
- Data will be downloaded by researchers
- Full access control to the data (i.e., read/write/Exec access)
- Data needs to be findable Potentially after publication and/or after the 3-year quarantine

3.1.6 DODAS use case

The Dynamic On-Demand Analysis Services (DODAS) is an open-source Platform-as-a-Service tool, developed and maintained by INFN, which allows to deploy software applications over heterogeneous and hybrid clouds. DODAS completely automates the process of provisioning, creating, managing and accessing a pool of heterogeneous computing and storage resources, thus drastically reducing the learning curve, as well as the operational cost of managing community-specific services running on distributed clouds. DODAS currently supports the on-demand deployment of:

- Batch system as a Service instances based on the HTCondor technology;
- Big Data analysis platforms providing Machine Learning as a service;

DODAS has already been integrated into the submission Infrastructure of the Compact Muon Solenoid[16] (CMS), one of the two biggest and general purpose experiments at the CERN Large Hadron Collider [17](LHC), and into the Alpha Magnetic Spectrometer[18] (AMS-02), an experiment hosted on the International Space Station, data analysis workflow.

One of the main architectural goals of DODAS Thematic Service is to provide a high level of modularity, a key to a generic applicability.

Being modular, the architecture provides the ability to easily customize the workflow depending on the community computational requirements. In this context the major EOSC-hub services adopted are:

- The PaaS Orchestrator which has the role of taking the requests related to application or service deployment coming from the user expressed using TOSCA, the OASIS standard to specify the topology of services provisioned in IT infrastructures. Based on the user

[16] https://home.cern/science/experiments/cms

[17] https://home.cern/science/accelerators/large-hadron-collider

[18] https://ams.nasa.gov/

requirements (typically expressed in the TOSCA template), the Orchestrator has the role to identify the best infrastructure (IaaS) for the deployment considering information about user's SLAs the availability and the health status of the IaaS services.

- The actual interaction with the infrastructure is delegated to the Infrastructure Manager (IM). This service is a key in the architecture as it is in charge to deploy complex and customized virtual infrastructures on different IaaS Cloud deployment, providing an abstraction layer to define and provision resources in different clouds and virtualization platforms. From the integration perspectives the TOSCA support provided by IM represent a key feature. Moreover, it eases the access and the usability of IaaS clouds by automating the VMI (Virtual Machine Image) provisioning including selection, deployment, configuration, software installation.
- The glue of the implemented flow is the Identity and Access Management service (IAM). IAM provides a layer where identities, enrolment, group membership, attributes and policies to access distributed resources, and mostly supports the federated authentication mechanisms. Identity and Access Management is provided through multiple methods (SAML, OpenID Connect and X.509) by leveraging on the credentials provided by the existing Identity Federations (i.e., IDEM, eduGAIN, EGI CheckIN). ESACO service is also part of the DODAS integrated service, and this is responsible for guaranteeing Cloud providers (such as Openstack based providers) with support of multiple OAuth2 Authorization Servers.
- The support to Distributed Authorization Policies and Token Translation Service in DODAS is implemented thanks to the WaTTS service (https://github.com/indigo-dc/tts) which guarantees selected access to the resources as well as data protection and privacy.

Service: Indigo-IAM, PaaS Orchestrator, WaTTS

Resource Providers: INFN

Resources:

- Deployment of a dedicated IAM and WaTTS instance:

  dodas-iam.cloud.cnaf.infn.it

  dodas-tts.cloud.cnaf.infn.it

Use case requirements:

- Dedicated instances of security services (IAM and WaTTS)

3.1.7 DARIAH use case

The DARIAH (Digital Research Infrastructure for the Arts and Humanities) Thematic Service (TS) aims to enhance and improve the usage of the cloud-based services and technologies in the domain of the digital arts and humanities research. It will enable end-users coming from the digital arts and humanities domains to seamlessly store, describe (metadata) and share their datasets, discover, browse and reuse datasets shared by the others and to perform analysis on various data volumes.

The DARIAH TS is providing a set of services and in particular, among them, the “Invenio-based repository as a service” enables researchers and scholars to easily create, deploy and configure their own Invenio-based repository and host it on cloud infrastructures.

The service is built around a set of EOSC-hub services:

- The FutureGateway that provides a user-friendly web interface for requesting the deployment of the repository: the authenticated user can customize the deployment request using a simple form; through the web interface it is also possible to monitor the status of the deployment and get the endpoint to access the deployed system.
- The PaaS Orchestrator receives the deployment request submitted by the users through the FutureGateway and coordinates the provisioning and configuration of the needed cloud resources on the “best” cloud provider. The latter is selected taking into account information such as the SLAs signed with the users, the monitoring data about the health of the provider services.
- The Infrastructure Manager (IM) is steered by the Orchestrator to interact with the cloud sites (through the APIs provided by the different Cloud Management Frameworks) in order to provision the virtual resources (servers, block devices, etc.) needed by the deployment. The contextualization of the virtual machines is managed by IM as well exploiting ansible to automate the installation and configuration of the software components.
- The INDIGO IAM provides the authentication/authorization infrastructure: OIDC tokens issued by IAM are used to access and interact with the PaaS services and also with the cloud providers.

Service: FutureGateway, PaaS Orchestrator, Infrastructure Manager (IM), Indigo-IAM

Use case requirements:

- Automated deployment of Invenio-based repository on Cloud environment

3.2 Integration by Thematic area

The following sections describe the integration activities as they were addressed in each specific Thematic area.

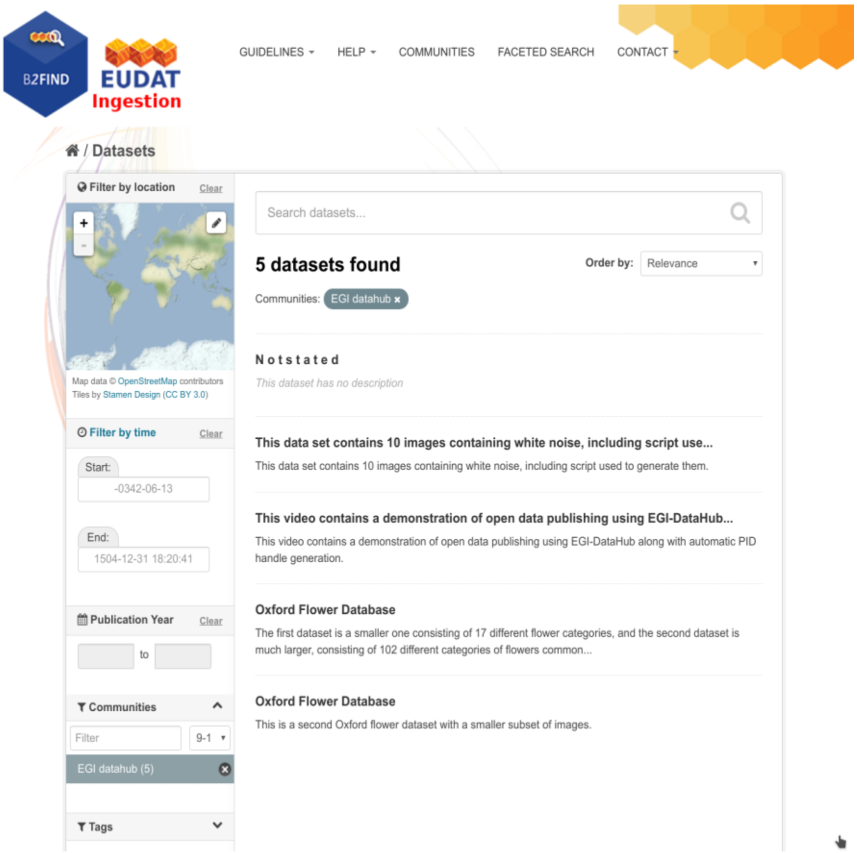
3.2.1 Data Discovery and Access

This section presents the overview of new or improved features achieved by extending or integrating existing services, and their relevance for thematic and specialized services.

3.2.1.1 Discoverability of EGI DataHub datasets via B2FIND

EGI DataHub and B2FIND services have been integrated by means of the OAI-PMH endpoint exposed by EGI DataHub (http://datahub.egi.eu/oai_pmh?verb=ListRecords&metadataPrefix=oai_dc), which exposes metadata of all published open data sets in EGI DataHub.

From a user perspective, this works in the following way. In order to publish a dataset, users have to create a share from their selected directory in EGI DataHub. The share by default is not public but can be accessed using a public URL endpoint in the EGI DataHub. Once the share is created, users have the option to publish it as an open data set. This step requires that the user selects a handle registration service and provides relevant metadata in Dublin Core format. For the EOSC-hub users, EGI DataHub has been integrated with B2HANDLE, thus registration of PID handles is automated. During publishing of the dataset using EGI DataHub, they will see the name of a PID minting service in the dropdown menu and EGI-DataHub will request generation of the PID for this dataset and from now on it will be included in the OAI-PMH endpoint listings including the provided DC metadata.



Start: -0342-06-13

End: 1504-12-31 18:20:41

5 datasets found

Communities: EGI datahub

Notstated

This dataset has no description

This data set contains 10 images containing white noise, including script use...

This data set contains 10 images containing white noise, including script used to generate them.

This video contains a demonstration of open data publishing using EGI-DataHub...

This video contains a demonstration of open data publishing using EGI-DataHub along with automatic PID handle generation.

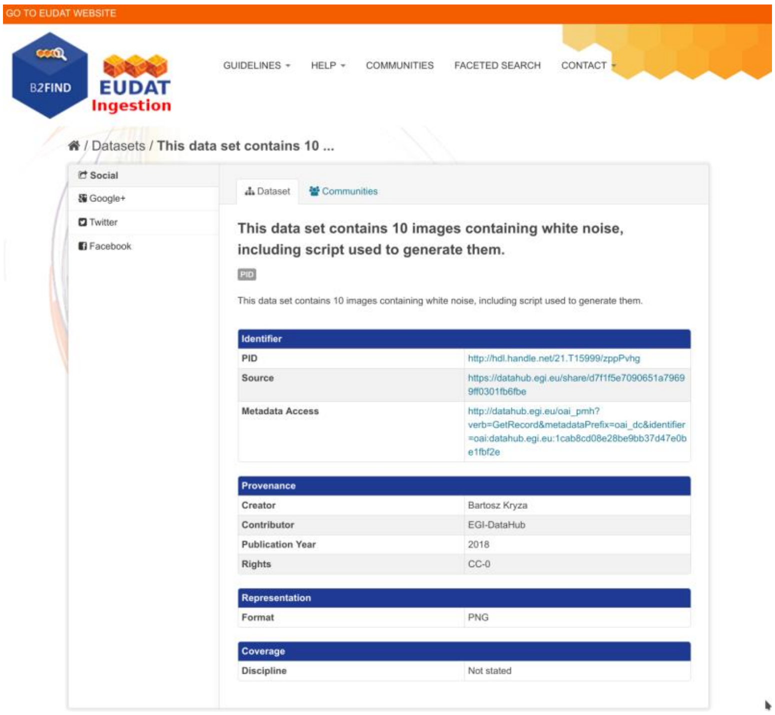
Oxford Flower Database

The first dataset is a smaller one consisting of 17 different flower categories, and the second dataset is much larger, consisting of 102 different categories of flowers common...

Oxford Flower Database

This is a second Oxford flower dataset with a smaller subset of images.

Figure 8. Test open data sets harvested from EGI DataHub by B2FIND



/ Datasets / This data set contains 10 ...

This data set contains 10 images containing white noise, including script used to generate them.

This data set contains 10 images containing white noise, including script used to generate them.

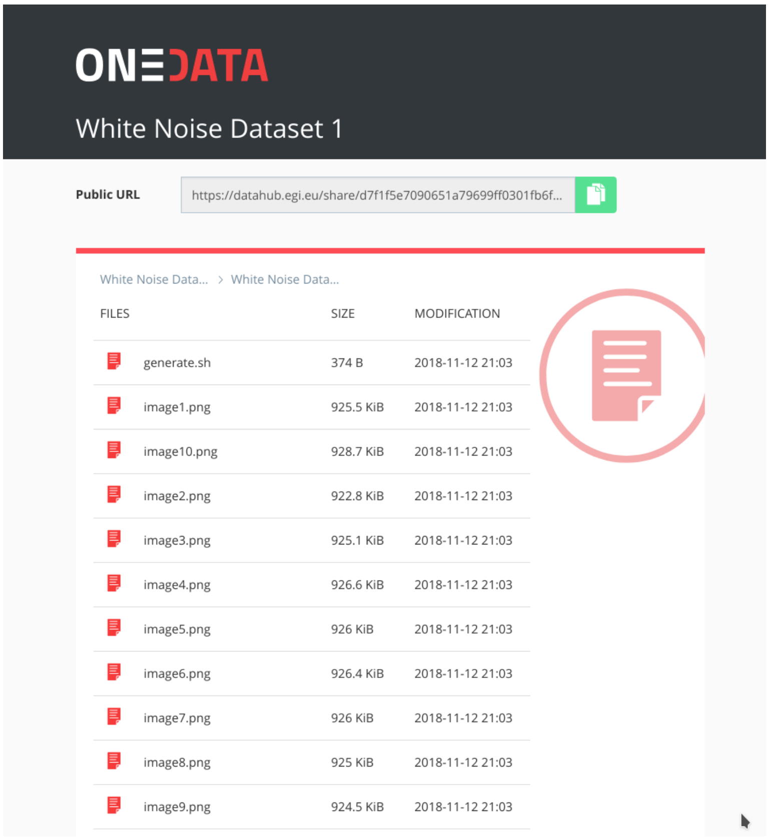
Identifier	
PID	http://hdl.handle.net/21.T15999/zppPvhg
Source	https://datahub.egi.eu/share/d7f1f5e70906519a7969 9ff0301fb6fbe
Metadata Access	http://datahub.egi.eu/oai_pmh?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:datahub.egi.eu:1cab8cd08e28be9bb37d47e0be1fbf2e

Provenance	
Creator	Bartosz Kryza
Contributor	EGI-DataHub
Publication Year	2018
Rights	CC-0

Representation	
Format	PNG

Coverage	
Discipline	Not stated

Figure 9. Metadata of a selected dataset harvested by B2FIND



White Noise Dataset 1

Public URL https://datahub.egi.eu/share/d7f1f5e7090651a79699ff0301fb6f...

White Noise Data... > White Noise Data...

FILES	SIZE	MODIFICATION
generate.sh	374 B	2018-11-12 21:03
image1.png	925.5 KiB	2018-11-12 21:03
image10.png	928.7 KiB	2018-11-12 21:03
image2.png	922.8 KiB	2018-11-12 21:03
image3.png	925.1 KiB	2018-11-12 21:03
image4.png	926.6 KiB	2018-11-12 21:03
image5.png	926 KiB	2018-11-12 21:03
image6.png	926.4 KiB	2018-11-12 21:03
image7.png	926 KiB	2018-11-12 21:03
image8.png	925 KiB	2018-11-12 21:03
image9.png	924.5 KiB	2018-11-12 21:03

Figure 10. Referenced data set can be accessed from web browser directly by following the handle

The remaining issues involve alignment of the metadata schemes between B2FIND and EGI DataHub, in particular ensuring that fields relevant for particulars communities will be provided when publishing the datasets via EGI DataHub.

Since the release of D6.4, focus has been placed on facilitating training and demonstration activities related to data discovery and open data publishing. For this a demo instance of EGI DataHub has been setup at https://datahub-demo.egi.eu, which allowed creating test open data sets, which were then discovered by a special instance of B2FIND without cluttering the official public B2FIND repository with test records. This feature was fully enabled by the end of March 2020, when EGI DataHub was upgraded to the latest currently published Onedata version - 19.02.1. Meanwhile, B2FIND incrementally harvests the EGI DataHub metadata test records from their official OAI-PMH endpoint http://datahub.egi.eu/oai_pmh and the records have been integrated into the active B2FIND portal.

3.2.1.2 Staging data stored in EGI DataHub by B2STAGE for processing

During the former reporting period the integration between B2STAGE and EGI DataHub has been investigated, reaching the conclusion that the task was not feasible due to the lack of integration between B2STAGE and B2ACCESS. As a result, the activity has been diverted towards the integration between EGI DataHub and B2SAFE. To reach the goal EGI DataHub started to implement a WebDAV client to expose B2SAFE resources through the davrods interface. Furthermore, transfer tests have been performed.

In the scope of this activity, WebDAV driver has been added to Onedata which is the underlying platform for EGI DataHub, which allows to both import existing data from B2SAFE as well as storing data from DataHub in B2SAFE, using a WebDAV endpoint. Transfer tests were undertaken using a test WebDAV endpoint provided by CINECA and a test instance of Onedata. In order to provide a real production service a token refresh mechanism is required in Onedata which will allow for the permanent registration of WebDAV endpoints (by default access tokens generated using B2ACCESS expire after a few hours).

Since the integration between B2STAGE and B2ACCESS has been completed with the release of D6.4, as described in detail in section 3.2.4. in the last project period focus was placed on exploiting this integration and proceeding with the task to integrate EGI DataHub and B2STAGE, as originally planned.

This integration activity allows to move data, from any configured endpoint, not only from B2SAFE. Detailed instructions can be found within the EGI wiki page 'Jointly exploit EGI and EUDAT services' [19] and in a step-by-step demonstration[20] shows in two scenarios how data can be staged from and into EGI datahub Onedata storage.

[19] https://wiki.egi.eu/wiki/Jointly_exploit_EGI_and_EUDAT_services

[20] https://datahub.egi.eu/share/8c81c2dac4a2b3683b1727e49b5657f6

3.2.1.3 *Integration between B2STAGE and B2ACCESS*

B2STAGE HTTP-APIs fully implemented the OAuth2 workflow required to support the B2ACCESS authentication (described in the Service Integration Documentation[21] from B2ACCESS) by exposing two different endpoints. The first (/auth/askauth) is intended to let B2STAGE manage the whole OAuth2 workflow, and it requires the user to operate through a web browser. The second one (/auth/b2safeproxy) alternatively allows users to skip the authorization part on B2ACCESS by directly providing an access token, so that this endpoint can also be requested from command line interfaces and/or included in automated scripts.

When the user calls the /auth/askauth endpoint from a web browser the request is automatically redirected to the B2ACCESS website, where the user can log-in and authorize HTTP APIs to access the user profile. Then B2ACCESS redirects back to B2STAGE service by including two tokens: an access token (with a validity of a few hours) and a refresh token (that can be used to request for new access tokens). By using the access token, B2STAGE retrieves from B2ACCESS the user profile, in particular to obtain the email address used to map the request over B2SAFE users. The workflow described till now is not executed when the user calls the /auth/b2safeproxy endpoint, since in this case the B2ACCESS access token is provided by the user as input. In both cases B2STAGE creates, and provides to the user, a new JWT token linked to both B2ACCES tokens. That JWT token can be used to make further requests on restricted endpoints, and it allows B2STAGE to transparently use B2ACCESS tokens. The access token is provided to B2SAFE to authenticate the user by adopting the PAM protocol. In case of authentication errors, the access token is intended to be expired and B2STAGE uses the refresh token to ask B2ACCESS for a new access token.

HTTP APIs are connected to B2SAFE by means of the python-irodsclient[22](PRC) but this library did not support the PAM protocol. The lack of this functionality delayed the completion of this activity and postponed other integration activities based on B2ACCESS common credentials. To be able to proceed with this task, it was decided to directly contribute to the development of the python irods client and extend the required functionalities by providing a merge request with our implementation of the PAM protocol. The merge request has been accepted by the irods team and PAM is now officially supported by PRC.

3.2.1.4 *Retrieve processed data from B2STAGE and share into B2SHARE Sharing processed data by B2SHARE*

Work on this task was stalled for a while due to personnel resource issues with task's B2SHARE related implementation and due to some misunderstandings with what is being pursued with the task. Misunderstandings have been solved and there is now a better understanding of what will be done in course of this task.

A feature will be implemented to enable users to fetch files stored in an external storage system such as B2SAFE, within B2SHARE UI by utilizing B2STAGE for the actual transfer of bytes from B2SAFE to the user's computer.

[21] https://eudat.eu/services/userdoc/b2access-service-integration

[22] https://github.com/irods/python-irodsclient

In preparation for the common authentication layer between B2SHARE and B2STAGE, support for OAuth2 workflows and support for PAM protocol have been implemented to B2STAGE. Chapter 2.2.4 describes these activities more in detail.

Previously identified fixes to B2STAGE HTTP-API required to support B2SHARE integration have been implemented. Support for HTTP HEAD method was implemented, and it is now possible to provide JWT token in the URL instead of Authentication-header. This was required due to a technical constraint (HTTP GET request made by B2SHARE cannot provide an Authentication-header).

For B2SHARE, detailed implementation plans have been sketched and workflow defined in them has been fully tested. In essence, B2SHARE will require support for exchanging B2ACCESS user info tokens for B2STAGE JWT tokens and support for delivering this JWT token back to B2STAGE when the user requests to download a file stored in B2SAFE. Implementation of these features should be done with good usability in mind. Implementation of these features will continue during 2020.

3.2.1.5 EUDAT dataset discoverability in OpenAIRE Community Dashboard

OpenAIRE and EUDAT-B2FIND enhanced the compliance of their guidelines for data providers[23]. This is particularly evident in the use of common standards (such as OAI-PMH) and the compatibility of the metadata schemas used.

Furthermore, a cooperation between EOSC and OpenAIRE Advance[24] that comprises the provisioning of the enriched metadata indexed in B2FIND to OpenAIRE was started. This will lead both to a further spreading and to an improved curation of meta data indexed in B2FIND by the services of OpenAIRE. In order to implement this, B2FIND planned in the last reporting period to offer its processed metadata in a format compatible with OpenAIRE via OAI-PMH.

Meanwhile an OAI-PMH API was built for datasets stored in B2FIND's metadata database. This is implemented based on the ckan extension for OAI-PMH exposure from CKAN[25]. The installation works technically and has been released to OpenAIRE for harvesting. However, some content issues and policies still need to be clarified. This requires communication activities concerning standardization agreements, which is (and probably will always be) work in progress.

In this context initiatives aiming to consolidate and unify metadata standards and schemes on the level of EOSC wide generic and FAIR metadata management should also be highlighted. Of particular relevance is the initiative of B2FIND and B2SHARE to establish a common metadata schema for EUDAT services and the cooperation with EOSCpilot 6.2 'Data Interoperability'[26] and the RDA's Data Discovery Paradigms IG[27].

[23] compare guidelines of OpenAIRE (https://guidelines.openaire.eu/en/latest/data/index.html) and EUDAT-B2FIND (http://b2find.eudat.eu/guidelines/introduction.html)

[24] https://docs.google.com/document/d/1zXcDrrS2Ud8XL2lDFJcj2b1CQjMzmHKp7USBBJkKvVc

[25] https://github.com/openresearchdata/ckanext-oaipmh

[26] https://eoscpilot.eu/content/d66-2nd-report-data-interoperability

[27] https://www.rd-alliance.org/groups/data-discovery-paradigms-ig

Since the release of D6.4, the CKAN-OAI-extension for B2FIND has been configured and exposes the B2FIND metadata in both DataCite and B2FIND’s own b2f metadata format. OpenAIRE is now testing the harvesting workflow and works on mapping the metadata to the OpenAIRE schema.

3.2.1.6 Future Integration Plans

This section presents the overview of planned features to be achieved by extending or integrating existing services in the future and enhancing their relevance for thematic and specialized services.

- Normal software and service maintenance activities, e.g., upgrade and consolidate the metadata schema of B2FIND.
- Extend the uptake of metadata and indexing more data resources registered in EGI-datahub and B2SAFE
- Set up an OAI endpoint on top of B2FIND from where OpenAIRE and other indexers can harvest metadata in a proper format
- Further development of B2DROP to enhance the allowed size of files and user storage space to support applications with big data volumes.
- Improve two-way integration with B2NOTE and B2FIND.
- Investigate the integration between B2STAGE and EGI DataHub and complete the data transfer tests between B2SAFE and DataHub.

3.2.2 Federated Compute

The following sections describe the integration activities as they were addressed in each specific Task area.

3.2.2.1 Elastic Kubernetes cluster support

Kubernetes provides an ideal platform to run container-based workloads be it an application composed of several microservices or a High Throughput Computing application composed of loosely coupled jobs. Indeed, Kubernetes supports autoscaling capabilities by means of the Horizontal Pod autoscaler that is in charge of determining the right amount of pods inside a Kubernetes cluster depending on the varying workload. However, this does not affect the number of nodes of the Kubernetes cluster. To this aim, Kubernetes offers the Cluster Autoscaler which is responsible to scale the cluster nodes when the pods cannot be scheduled on nodes because there are no free resources available. However, this component is only functional for Amazon Web Services, Microsoft Azure and Google Cloud Platform.

It was the aim of this task to include auto-scaling support for Kubernetes clusters deployed by the users on any IaaS Cloud site. To this goal, first a set of Ansible roles for the automatic installation and configuration of a Kubernetes cluster have been created. It enables the provisioning of a fully functional Kubernetes cluster and the addition/removal of new nodes on runtime. It includes the configuration of the different network plugins (Flannel, Calico, Weave, etc.), the deployment of the dashboard and the installation of the Helm tools:

https://github.com/grycap/ansible-role-kubernetes

Then a plugin for the CLUES elasticity system has been developed to enable the elastic management of the Kubernetes cluster adding/removing nodes based on the workload.

https://github.com/grycap/clues/blob/master/cluesplugins/kubernetes.py

The EC3 template has been created to enable launching the elastic Kubernetes cluster using the EC3 tool automatically.

https://github.com/grycap/ec3/blob/master/templates/kubernetes.radl

Since the release of deliverable D6.4, the migration to Helm v3 and the installation of the cert-manager to generate certificates using valid CA entities as Let's encrypt have been addressed. The role has been updated to the 1.20 version that has introduced some major changes such as the removal of the support to the insecure port in the localhost.

3.2.2.2 EGI Workload Manager

The EGI Workload Manager service integrated HTC resources orchestrated via the central Task Queue service. From the user perspective, this is a single entry point they need in order to access all the computing resources that they are allowed to use. This service replaced without any lacking functionality the gLite WMS service that played the role of the HTC resources orchestrator and was decommissioned in the beginning of the project.

The EGI Workload Manager also integrated resources of the EGI Federated Cloud to be accessible in the same way as HTC resources from the user perspective. Some HPC centers were demonstrated to be accessible in the same way via specially configured queues as defined in the service configuration.

The EGI Workload Manager was integrated with the EGI Check-In AAI system. This allowed user access to the service portal using the OAuth2 tokens obtained via the Check-In SSO authentication procedure. The client API was enhanced by tools for obtaining user X509 certificate proxies once the OAuth2 tokens were obtained. This allowed using the current X509 based client/server protocol with the Check-In authentication. The work is in progress to provide a client API based on the HTTP protocol which will be fully based on the OAuth2 tokens security mechanism.

3.2.3 Processing and Orchestration

This section summarizes the integration improvements carried out in the processing and orchestration layer in the final stage of the project, after the latest activity reported in Deliverable D6.4 "Second report on the maintenance and integration of common services". Therefore, an update of the integration activities achieved after the aforementioned deliverable is described. The description is performed for each major software component involved in the processing and orchestration layer.

3.2.3.1 INDIGO Orchestrator evolution

The INDIGO Orchestrator is being extended in the framework of the EOSC-hub project both for ensuring better performances and enhanced reliability and for supporting the new requirements coming from the thematic services.

The INDIGO Orchestrator has advanced its integration with EGI Check-In in order to directly accept requests from users using this AAI framework. A Kubernetes plugin has been developed in order to submit a TOSCA template to deploy a Helm chart on top of a Kubernetes cluster, integrated with IAM. Also, now networks can be selected when tenants provide multiple public or private networks. The visual dashboard of the INDIGO Orchestrator registered in the EOSC marketplace has been improved with a new landing page aimed at users that come from external communities to better understand the dashboard.

As a side note, there has been a recent effort[28] to integrate in the upstream version of the EGI-Federation's information system, the cloud-info-provider (CIP), the extensions that enable the dynamic gathering of information required for the operation of the aforementioned set of cloud-oriented technologies supported by the Orchestrator, including Mesos, Amazon EC2 and Onedata. Likewise, these extensions provide the required scheduling information to leverage GPU resources both from OpenStack and Mesos frameworks. As a result of this integration effort, the CIP tool will be able to feed both AppDB (through AMS) and CMDB endpoints, and thus, be a key enabler of the Orchestrator integration into the EGI Federation.

3.2.4 Data and Metadata Management

Data and metadata management services allow users to store data sets e.g., in a repository, also associating persistent identifiers to one or multiple copies. Data can be published, and specific metadata associated as well, then those metadata can be harvested and indexed to make the data findable. This section presents results achieved during the last year as a continuation of work done and reported in D6.4 (Section 6.1).

3.2.4.1 Maintenance, interfaces and integration options of the services

B2SAFE

B2SAFE is a long-term preservation and policy-based data service. It allows community repositories to implement data management policies on their research data that is distributed across multiple administrative domains. For detailed description of the service see D6.1. and https://www.eudat.eu/b2safe .

Recent improvements include:

- Integrating B2SAFE-core (and related components) in the brand new gitlab.eudat.eu DevOps environment, hosted at KIT. This has been a particularly good step in the direction of consolidating B2SAFE-core components versions, also creating a brand-new environment where to build/test new features and fixes, also managing software dependencies in an easier way so as to improve both the integration of multiple services and the software release process.
- Completed the integration of B2SAFE-core code to gitlab.eudat.eu web portal managed at KIT, now working to the setup of a new gitlab-runner environment to complete the CI/CD project configuration.

[28] https://github.com/EGI-Foundation/cloud-info-provider/pull/214

- Completed the porting of B2SAFE-core to Python3, introducing pid microservices as well as completing code refactoring. New pre-release version B2SAFE-4.3.0 has been released and new fixes have been merged. All these efforts go in the direction to align and integrate different EUDAT components, relying on different versions of languages and software libraries.
- Docker images as well as rpm packages will be published in the new JFrog repository[29] hosted at Sara, so as to integrate and consolidate software components, as well as to create and improve a brand new "build and integration" system, where developers will schedule pipelines, also to satisfy dependency requirements between multiple software services.
- Started merging the B2SAFE-B2SHARE iBridge integration component within the B2SAFE-core master version in the new gitlab.eudat.eu repository. This will improve the integration between B2SAFE and B2SHARE, the component will be included in the final release of B2SAFE in the coming weeks.
- Finally, local configurations have been created locally in CINECA in order to build/test specific B2SAFE-core versions, exploiting a new container-based Kubernetes environment. This will be fundamental also to test specific configurations and solutions on sites. This will definitively prove the advantages of adopting DevOps as a methodology for releasing and deploying new versions of B2SAFE-core components.
- New collaboration has started with Compbiomed community in order to implement data replication workflows, a new document including use case has been produced which defines the next step to federate EUDAT sites in order to satisfy requirements expressed by community users.

B2SHARE

B2SHARE is a data storage and sharing service for research communities and individual researchers. It allows discovery and publication of research datasets by providing detailed descriptions in the form of standardized metadata. For a detailed description of the service see D6.1. and the EUDAT website.[30]

- The integration of B2SHARE with B2SAFE is moving ahead and will be completed soon, the new B2SAFE-B2SHARE iBridge software component is going to be merged in the B2SAFE-core master branch and to be tested in the new gitlab.eudat.eu repository.
- Recent improvements also included moving ahead with B2NOTE integration using the brand new B2NOTE 3 APIs. B2NOTE's new advanced features will allow better integration of the B2NOTE software component within different web portals and services, also exploiting the most recent software libraries and languages.
- Updated back-end layer using Invenio v3: since version 3 of the open source Invenio repository framework, used as the back end for B2SHARE, has been released. Invenio "has been completely rewritten from scratch with a radically improved architecture and technical implementation."

[29] https://artie.ia.surfsara.nl/ui/

[30] https://eudat.eu/services/userdoc/b2share

- Updated B2SHARE WebUI to React v16: React.js is used to render the B2SHARE web interface and make it interactive. Version 16 has been released some time ago and the latest version needs to be implemented in the current B2SHARE software stack. This will improve the integration between B2SHARE and software plugins as B2NOTE (also based on React.js), as well as the integration with multiple software libraries for web applications.

B2NOTE

B2NOTE is a data annotation service integrated with data repositories/data publication services. It allows the service users to add extra information without modifying the underlying data record.

For B2NOTE lot of work has been done in the last year including:

- B2NOTE version 3 has been released, it has been redesigned completely using JavaScript and better organized, also including APIs description and documentation. The software developer can now integrate B2NOTE using iFrame or directly invoking a JS client. For the integration steps of the B2NOTE within a specific software service or web portal application, nothing changes with a new version released. However recent upgrades of B2NOTE service seriously improve the capability of such components to be integrated in web portal pages or web interfaces in a simple and maintainable way.
- Tests have been done for B2NOTE 3 version. Search and discovery functionalities have been improved and work was also undertaken on the integration for the B2ACCESS production version. New use cases and communities have been investigated, including digital Humanity, Covid19 and Neuroscience.
- Simplified architecture for the B2NOTE full JS implementation, for the UI. Python3 Flask framework and REST APIs updated so as to improve again the integration with external clients and services.

New functionalities and developments:

- New functionalities, as the button for annotating datasets, have been added in B2NOTE, this will appear as soon as the software developer will integrate the new B2NOTE 3 version within the generic web portal (e.g., B2SHARE).
- New developments about the UI as user profiles are enabled to select public/private information display for each user, including annotation and provenance. Users can see now what has been published for each project and this will be visible as soon as the B2NOTE will be integrated on the specific web page.
- Now with the new B2NOTE version if the user wants to publish information, he/she needs just a proper id to get related information.
- Moreover, the idea of annotating resources is also being explored. In the form of integrating with new service providers, the goal is to have such functionality for any resource or service within the federation. To this extent technical and security issues of course need to be addressed.
- Integration of the B2NOTE component has been discussed both with OpenAIRE and other communities (Zenodo). Integrating with OpenAIRE AAI has been achieved. B2NOTE has been integrated also with the OpenAIRE community dashboard. Finally, the integration with B2SHARE has been finalised.

B2HANDLE

B2HANDLE is a service for the provisioning of Persistent Identifiers (Handles), including provision of namespaces (prefixes), hosting of Handle servers and additional server components for search and programmatic access. For detailed description of the service see D6.1. and https://eudat.eu/services/userdoc/b2handle.

- In the last year, the integration with the DataHub basically has been completed, remarkably interesting work and many demos have been done.
- Discussion is still ongoing about how metadata and PID records may be improved (Low priority at the moment)
- Planning to upgrade the B2HANDLE server instance now ongoing, consequently testing the integration with the Handle.net service.
- Integration with B2SHARE is actually in progress, tests will be completed soon.
- Made pre-release available supporting Python 3, this in order to be aligned with the rest of the EUDAT services and to help the integration of the new versions of software components.
- On the server-side new setup has been done in production and lot of bug fixes have been made, also for the purpose of facilitating the integration of multiple components and services with B2HANDLE.
- Added integration for the B2HANDLE code stuff in the Eudat Gitlab: gitlab.eudat.eu. This will improve the build/test integration process and will help manage software dependencies between EUDAT components. The B2HANDLE will be tested together with B2SAFE-core and the related code definitively moved to gitlab.eudat.eu like for B2SAFE. Artifacts will be published on the new repository based on JFrog for further releases.

Future integration plans

This section presents the overview of new or improved features achieved by extending or integrating existing services, and their relevance for thematic and specialized services.

- Normal software and service maintenance activities; complete to merge B2SAFE-B2SHARE iBridge component into the B2SAFE-core master branch and release final new B2SAFE-core version (expected in the coming weeks).
- Further development of B2SHARE, also enabling and supporting new communities to start integrating and using B2SHARE publishing their own records on the B2SHARE service and portal.
- Integrate with further, non-EUDAT services that are part of EOSC-hub service catalogue.
- Remove unused components and consolidate already existing software services with an overlook to the future sustainability plans.
- Finalize the integration between B2SHARE and B2SAFE, completing the integration of the iBridge software component in the B2SAFE-core master branch (almost done).
- Continue the collaboration with the CompBioMed community in order to provide integration solutions and tools in order to implement the use case produced this year, supporting communities for moving data (data transfer) and to perform data and metadata management.

- Complete the integration of the Data Policy Manager in the new DevOps environment based on gitlab.
- Extend the gitlab.eudat.eu web portal with new configurations, also improving the gitlab-runner VMs facility that will be hosted at KIT. This will again facilitate the whole build and integration process for all the EUDAT software components and middleware stuff.
- Extend the data policies to support further services and communities that will integrate our software services.

3.2.5 Sensitive Data Services

SECURE B2SHARE

The secure B2SHARE concept has been developed during the activities of Task 6.6 and has been used to expose sensitive data services to users via well-known EUDAT services. At its core, the secure B2SHARE concept is a concept which describes a way for publishing datasets that contain sensitive data. The concept defines a layered architecture which requires the integration of several existing services, effectively providing a compound service. This layered architecture is described in the following sections.

Secure B2SHARE has three distinct components: B2SHARE, Secure Data Submission -service (SDS) and Authorization service. Dataset owner uploads files in SDS, creates datasets and describes metadata for the datasets in B2SHARE and manages authorization to datasets in Authorization service. Datasets created in B2SHARE always only refer to files previously uploaded through SDS. Files themselves are stored in Secure Storage.
Researchers can find datasets with search functionality provided by B2SHARE, or through metadata discovery services such as B2FIND.

When a researcher discovers an interesting dataset an access request must be made. Data owner (or representative) reviews the access request and either rejects or accepts it. If authorization is granted, Secure B2SHARE notifies Secure Storage that a specific person has been granted access to a specific dataset.
Besides general guidelines conforming with information security best practices, Secure B2SHARE does not specify how access to sensitive data should be implemented, as this very much depends on the sensitive data infrastructure Secure B2SHARE is implemented on.

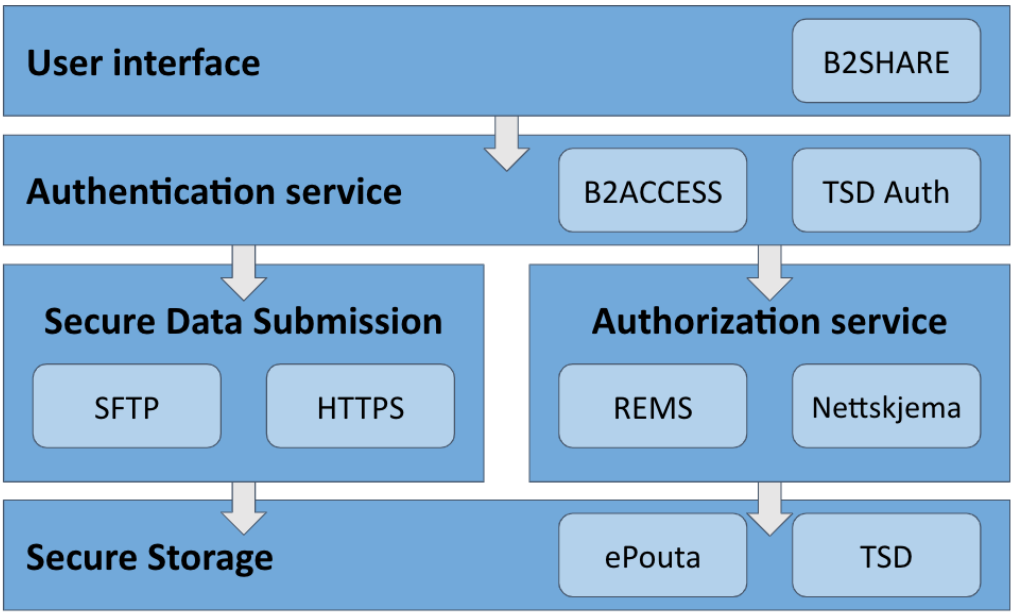


Figure 11: Secure B2SHARE layered architecture diagram.

Since Secure B2SHARE sends authorization decisions to Secure Storage, it must be possible for Secure Storage to link data access requests to specific authorization decision, i.e., person who makes data access request to Secure Storage, must be identifiable to be the same person that requests access to the dataset at Secure B2SHARE. This can be achieved by identity federation or by using common authentication service for both Secure B2SHARE and Secure Storage.

Implementation on TSD Infrastructure

1. Data owner registers a dataset in B2SHARE: Creates a B2SHARE record with a UUID. The Authorization service gets hold of the dataset-UUID association.
2. Data requester requests access to the dataset: Submit a data access request to the authorization service through B2SHARE web interface. The authorization service then notifies the data owner about the request.
3. Data owner grants/denies access to the dataset: Through the authorization service.

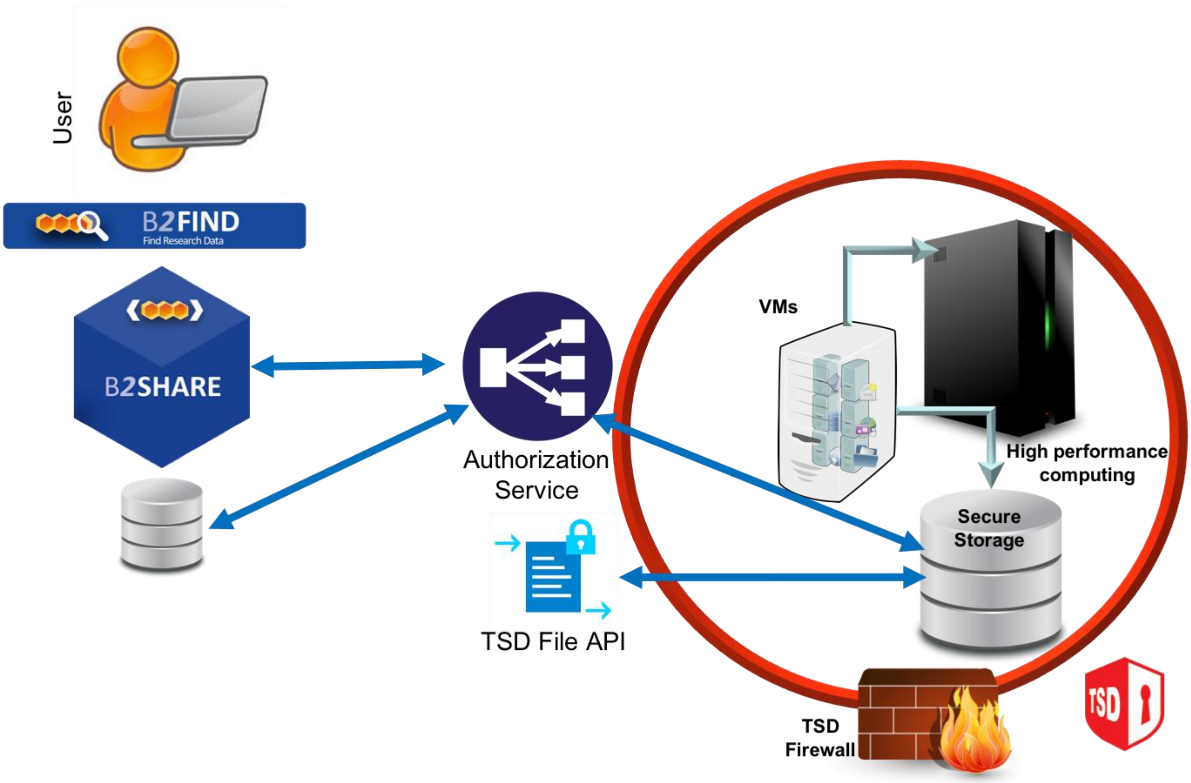
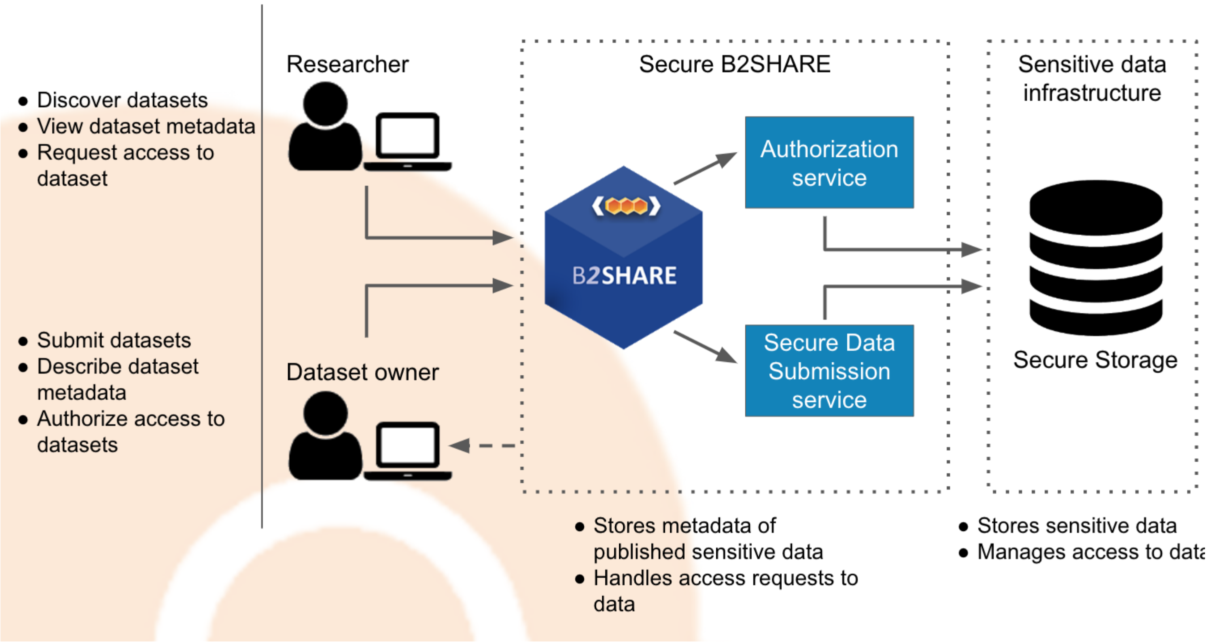


Figure 12. Secure-B2SHARE integration in TSD

Implementation on CSC Secure Storage

In CSC specific implementation of Secure B2SHARE, REMS Resource Entitlement Management System (REMS) implements the authorization service–component. Datasets submitted into CSC specific Secure B2SHARE have to be encrypted. Before storage, data is decrypted and re-encrypted using local keys.

Data processing happens via remote desktop connection to ePouta tenant, as policy of the platform requires that data is not transferred outside the secure environment.



- Discover datasets
- View dataset metadata
- Request access to dataset
- Submit datasets
- Describe dataset metadata
- Authorize access to datasets
Researcher
Dataset owner
Secure B2SHARE
B2SHARE
Authorization service
Secure Data Submission service
Sensitive data infrastructure
Secure Storage
- Stores metadata of published sensitive data
- Handles access requests to data
- Stores sensitive data
- Manages access to data

Figure 13. Secure-B2SHARE integration in CSC secure storage

Authentication

B2ACCESS -service provides authentication (and authorization) of users. When a user-logins to B2ACCESS, B2ACCESS provides the email address of the user and an anonymous permanent aid of the user to the service requesting the authentication. B2ACCESS use for user authentication for TSD, and CSC is under investigation. Currently, local authentication/login mechanisms are applied.

Sensitive metadata

Sensitive metadata can be stored in a file, stored in CSC-SS and referenced in a record as one would do with any other type of file. Format for the file has not been decided, but it should be a format that is commonly used for metadata records.

Discovery of datasets

When a sensitive dataset is published, only it is metadata is publicly available. Metadata is stored in B2SHARE service. Files, i.e., the sensitive data itself is stored in CSC Secure storage. B2SHARE supports harvesting of metadata via OAI-PMH protocol, which allows metadata catalogue services, for example B2FIND, to include sensitive datasets stored in Secure B2SHARE in B2FIND search results. B2SHARE also provides search functionality out-of-the-box.

Persistent identifiers used by B2SHARE

Secure B2SHARE supports obtaining Handle PIDs for publicly available metadata record of the dataset, not to the files of the dataset. Using Handle PIDs as identifiers for the actual sensitive dataset files is under investigation.

4 Summary

The initial goals outlined for the EOSC-hub common services have been met and, in many cases, extended upon based on the evolving nature of the service and integration requirements which come from the research communities.

Throughout the EOSC-hub project WP6 has focused on providing a catalogue of core services which can be used by research community projects and end users alike. These services have been maintained and enhanced during the project period to meet the evolving needs of these users and communities, ensuring that at each step of the project EOSC-hub is addressing the research use-cases which drive this endeavour. In addition to these core services, numerous use-case driven integration activities have been undertaken to provide added value which goes beyond a simple service offering. These integration activities bring value in several major ways. Firstly, by directly addressing the needs of the research communities, the EOSC-hub project reduces the load on the communities themselves, allowing researchers to concentrate on research and not infrastructure. Additionally, once these integration use-cases are addressed they may be used by many different communities. A final benefit arises since as services are added to the EOSC-hub they become part of an environment where they can be integrated with other services to bring added value. The whole is greater than the sum of the parts. Service providers gain from being part of EOSC-hub and communities and end users alike benefit from off the shelf services and the integration of these services which provides high level compound functionality.

The activities undertaken by WP6 contributes to the EOSC-hub Key Exploitable Results 4 (Internal Services in the Hub Portfolio), 5 (External Services in the EOSC Service Portfolio) and 8 (Interoperability and integration guidelines). Numerous services have already been integrated into the updated EOSC-portal and marketplace that will be maintained by the EOSC-hub follow-up projects. Additionally, the experience and expertise gained from the integration activities provides valuable input for the Interoperability and integration guidelines, highlighting best practices etc, which will be provided to follow up projects.