DICE

Data Infrastructure Capacity for EOSC

Deliverable D4.1

Planning for the integration with other services & platforms

Responsible Partner:	Forschungszentrum Jülich
Status-Version:	Draft v1.0
Date:	30/09/2021
Distribution level (CO, PU):	Public

Project Number:	GA 101017207
Project Title:	DICE: Data infrastructure capacity for EOSC

Title of Deliverable:	Planning for the integration with other services & platforms
Due Date of Delivery to the EC	30.09.2021
Actual Date of Delivery to the EC	02.10.2021

Work package responsible for the Deliverable:	WP4 - Integration with other services & platforms
Editor(s):	Mallmann, D. – FZJ Ariyo, C. – CSC Azab, A. – SIGMA
Contributor(s):	----
Reviewer(s):	Carpenè, M. – CINECA Tonello, N. – BSC
Recommended/mandatory readers:	WP5

Abstract:	This deliverable consists of the technical report from task 4.1 “Compute and Analysis” on the integration of B2-services and object storage services, and from task 4.4 “Sensitive Data” on the design of the sensitive data workflow.
Keyword List:	Compute, Analysis, Sensitive Data
Disclaimer	This document reflects only the author’s views and neither Agency nor the Commission are responsible for any use that may be made of the information contained therein

The DICE project has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

Document Description

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	03.08.2021	First draft version	FZJ
v0.2	18.09.2021	Inputs received from Task leaders	CSC, SIGMA
V0.3	24.09.2021	Comments and suggestions received by internal reviewers	BSC, CINECA
V1.0	30.09.2021	Final version ready for submission	PMT

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

Table of Contents

List of Figures

The DICE project has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

Terms and abbreviations

AAI	Authentication and Authorisation Infrastructure
API	Application Programming Interface
BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputacion
CSC	CSC – Tieteen Tietotekniikan Keskus Oy
DoA	Description of Action
EC	European Commission
EOSC	European Open Science Cloud
EU	European Union
FZJ	Forschungszentrum Juelich GmbH
GA	Grant Agreement to the project
HPC	High Performance Computing
IdP	Identity Provider
KPI	Key Performance Indicator
PID	Persistent Identifier
TSD	Services for sensitive data
SDA	Sensitive Data Archive (SDA
UiO	University of Oslo
WP	Work Package

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

Executive Summary

The deliverable 4.1 is the first deliverable of work package 4 "Integration with other services & platforms". It consists of contributions from task 4.1 "Compute and Analysis" and task 4.4 "Sensitive Data"; more specifically including:

- The technical report on the integration of B2-services and object storage service describes the planning for the integration of data services with computing platforms to enable analysis, data replication, and data publication of data from HPC and cloud computing environments.
- The report on the design of the sensitive data workflow is a description of the design of a complex workflow for sensitive data across different existing EOSC tools and services.

The deliverable gives a detailed overview of the planned integration activities for compute services and sensitive data. Addressees are user communities involved in the integration plan with WP5 activities (deliverable D5.1) and users of B2-Services and sensitive data.

The next deliverable of WP4 is due in six months. It will report on the pilot use cases for the integration of data services with computing platforms in task 4.1 and a sensitive data risk analysis from task 4.4. In addition, task 4.2 "Discovery and Referencing" will report on the integration of the integrity check for PIDs and task 4.3 "Long Term Preservation" on the long-term preservation policies for B2SHARE and B2SAFE.

1 Introduction

Main objectives of the work package (WP4) are to:

- Integrate data services with European computing platforms and High Performance Computing (HPC) infrastructures, like FENIX[1] and EuroHPC[2], to provide users with necessary research tools from computation to data sharing and publishing.
- Define integration scenarios together with thematic communities collaborating with the project (WP5): biomedicine, radio astronomy and environmental sciences.
- Implement the integration of the services from DICE with the targeted platforms and infrastructures.
- Update the service integration based on the feedback of the users.

In particular, the integration activities will improve:

- Compute and Analysis,
- Discovery and Referencing,
- Long Term Preservation,
- Sensitive Data management.

1.1 About this deliverable

The deliverable 4.1 being the first one of work package 4 “Integration with other services & platforms” consists of contributions from task 4.1 “Compute and Analysis” and task 4.4 “Sensitive Data”.

Task 4.1 “Compute and Analysis” provides the planning for the integration of data services with computing platforms to enable analysis, data replication, and data publication from object storage (more and more used as a permanent or semi-permanent storage service even with HPC environments) and cloud computing environments.

Task 4.4. “Sensitive Data” offers the report on the design of the sensitive data workflow, which is a description of the design of a complex workflow for sensitive data across different existing EOSC tools and services.

1.2 Document structure

This document is structured into the following sections:

- **Section 1** presents the introduction and objectives of the deliverable and the different tasks that are involved.
- **Section 2** presents task 4.1 and the necessary technical planning report for the integration of the data services.
- **Section 3** presents task 4.4 and the necessary technical description of the design of the sensitive data workflow.
- **Section 4** provides Conclusions.

[1] https://fenix-ri.eu/

[2] https://eurohpc-ju.europa.eu/

2 Compute and Analysis

2.1 Introduction

The aim of this task is to integrate data services with computing platforms to enable analysis, data replication, and data publication for high performance computing (HPC) and cloud computing environments. This will help the users to get their research data from computation to publishing much easier and thereby help in getting wider access to research data.

This activity is planned by using B2DROP service as a tool to maintain “recipes” for analysis, B2SAFE as a tool to safely store results of simulations/analyses, and B2SHARE as a tool to share and publish datasets.

This task will contain following smaller entities:

- B2DROP
  - Ensure that small data (batch queue scripts etc. similar small objects) can be read from B2DROP to HPC environment, and that small data can be written back to B2DROP.
- B2SAFE
  - Autoingest feature, AAI, iRods policy usage, object storage usage.
- B2SHARE
  - Share and publish object that has been auto-ingested by B2SAFE, from HPC and computing environments.
- FENIX AAI
  - Possibility to provide the FENIX infrastructure communities opportunity to use the tools with an integrated AAI.

2.2 Integration of B2DROP service into computing environments

B2DROP[3] provides researchers a common service for synchronizing and exchanging volatile research data within a small group and with fine-grained access control mechanisms. It is a user-friendly and trustworthy storage environment, which allows users to synchronize their active data across different devices and to easily share this data with peers. B2DROP is based on NextCloud[4] software.

The service is intended for the long-tail and still volatile data objects which can change and are still subject to active research, e.g. drafts of papers. Therefore, B2DROP offers versioning of all ingested files but does not attach persistent identifiers to them. NOTE: B2DROP is not intended to be used with large datasets or large number of files.

Beside of these characteristics and functionalities, B2DROP offers an intuitive user-interface via the web. In addition to web access, users can mount B2DROP as a drive on their desktop machines via WebDAV or use a desktop client which also allows offline synchronization.

B2DROP service can be accessed in one of the following ways:

- NextCloud occ command line tool from computing environments[5]

[3] https://www.eudat.eu/services/b2drop

[4] https://nextcloud.com/

[5] https://docs.nextcloud.com/server/latest/admin_manual/configuration_server/occ_command.html

The DICE project has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

- WebDAV tools or other similar tools like DaviX[6]

In both cases, authentication is done by using B2ACCESS[7] services. So, end user must have account on B2ACCESS or use federated identities supported by B2ACCESS.

It was noted that FENIX infrastructure did not act yet as an Identity Provider (IdP) to B2ACCESS service. Discussion on possibilities to integrate FENIX AAI with B2ACCESS has been ongoing to increase the possibility to include the FENIX infrastructures among those connected with DICE data management services.

2.3 Integration of B2SAFE service into computing environments

B2SAFE[8] is EUDAT's service for secure long-term preservation of research data. Data in B2SAFE is kept safe by replicating them to one or several other EUDAT sites, i.e. creating redundant copies of data and maintaining those by different administrative units.

Additionally to the replication workflow, the B2SAFE technology offers the framework to implement community-specific data policies. B2SAFE can store and replicate large amounts of data. It is meant to be used by repositories to preserve and backup their data collections. Moreover, B2SAFE can replicate reference datasets to various compute sites which are usually co-located with the B2SAFE endpoints.

B2SAFE service, at the core, exploits the iRODS[9] rule engine to perform a set of actions to implement specific behaviour defined in data management policies. The actions are defined by a set of iRODS rules which can either be executed on regular basis or be triggered by actions like data ingest. The rules interact with external software components which deliver functionalities such as PID registration. Several Python scripts facilitate the interaction.

Assuming that data locates in object storage service providing S3 API, we have couple of possible options how to get the data ingested to B2SAFE. One possibility is to use i-commands. However, federated authentication with i-commands would need additional development work. We are planning to evaluate alternative solution by enabling B2SAFE to ingest data directly from object storage service.

To be able to do this, the following activities are planned:

- Define what would be needed to get B2SAFE connected to B2ACCESS,
  - Authentication using linux pam module and B2ACCESS,
  - group data to ensure that home directory can be created.
- Evaluate B2SAFE (iRODS) autoingest feature with object storage system and with most common file systems used in HPC environments.
- Evaluate pros and cons of keeping the data under user space.
  - By default data will be copied to B2SAFE service
  - In some cases, e.g. when the dataset is large, it may be worth considering to not to copy the data to B2SAFE. This use case is, however, challenging, because it is difficult to guarantee that the data is not modified or deleted by user.

[6] https://github.com/cern-fts/davix

[7] https://eudat.eu/services/b2access

[8] https://www.eudat.eu/services/b2safe

[9] https://irods.org/

- Both options are a possibility depending on the size of data involved but having the data under B2SAFE can guarantee the consistency of the data, so it is considered the best approach.

When B2SAFE service has identified the data, it can create a Persistent Identifier for it. After that, if access rights are correct, the data can be e.g. replicated to other site or shared with the B2SHARE service.

2.3.1 Important features and capabilities for the end users

Assuming that B2SAFE would be the "nexus" that takes care of data transfers between EUDAT services, and computing environments and storages, we need to ensure that we are able to fulfil customer expectations.

The possible test users or communities are at first those part of the DICE project under WP5:

- Integrated Carbon Observation Systems (ICOS), Sweden (https://www.icos-cp.eu/)
- CompBioMed, United Kingdom (https://www.compbiomed.eu/)
- LOFAR (Low Frequency Array), Netherlands (https://www.astron.nl/telescopes/lofar/)

What we are aiming to test and evaluate the following parameters:

- Ease of use,
- Reliability,
- Availability,
- Serviceability,
- Performance,

considering the following test environments:

- User has data in HPC file-system (e.g. Puhti's Lustre-storage)
- User stores data from HPC to object storage (e.g. Allas object storage)
- User transfers data from HPC file-system (e.g. Puhti's Lustre) to B2SAFE
- User transfers data from object storage (e.g. Allas object storage) to B2SAFE

2.3.2 Connection of B2SAFE to B2ACCESS

B2ACCESS[10] is an easy-to-use and secure Authentication and Authorization platform developed by EUDAT. B2ACCESS is versatile and can be integrated with any service. When B2ACCESS is integrated with a given service, the user may log in by using different methods of authentication:

- Home organisation identity provider
- Social media account (e.g. Google account)
- EUDAT ID

EUDAT IDs are created by the B2ACCESS upon registration. Therefore, B2ACCESS is an Identity Providers for the users that do not have neither a Google account nor a Home Organization Identity Provider. In these cases, B2ACCESS also offers the tool for the managements of the EUDAT IDs.

Integration of B2SAFE to B2ACCESS will guarantee better and easier use of B2SAFE.

[10] https://www.eudat.eu/services/b2access

As mentioned before, it is also under evaluation the possibility to integrate FENIX AAI with the B2ACCESS service to give the FENIX infrastructure possibility to be used in conjunction with the DICE data management services.

2.3.3 B2SAFE (iRODS) autoingest features

B2SAFE deployments are nearly always unique. Every organization has a different set of use cases, different existing infrastructure (users, storage, networking, compute), and different plans regarding maintenance and life cycle management for both their data and their processes.

Two patterns have been identified that involve the upload of new or updated data - and then its immediate synchronization into the main dataflow of an organization's data management infrastructure like B2SAFE.

These patterns are:

- "Landing Zone", whereby users are placing data that needs to be ingested into B2SAFE and then the original is moved aside.
- "File-system Scanner" in which an organization's existing filesystem is to be maintained as the source of truth and updates are to be mirrored into B2SAFE. This second scenario occurs when an organization with existing data decides to deploy B2SAFE. This is the cold start problem. They need to register what they have so they can find it later. It also allows an organization with its own history and best practices to continue using those practices without changing their scientists' and other employees work habits or tool chains.

Regardless of how the Automated Ingest Framework is configured, once new data hits the B2SAFE policy engine, events are triggered, and action can be taken.

2.4 Integration of B2SHARE service into computing environments

B2SHARE[11] is EUDAT's service for storing, sharing and publishing so called "long tail" data, i.e. small and medium scale research data in various formats, which is usually not covered by institutional data preservation policies.

B2SHARE is implemented based on CERN, the European Organization for Nuclear Research Invenio technology[12].

The most common way to access B2SHARE is via its web-based user interface, using an ordinary web browser. However, B2SHARE also supports two application programming interfaces (APIs) over HTTP. The first API supports the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), used by other repositories to collect basic metadata describing the datasets (and used by EUDAT's own searching service, B2FIND). The second API is a general-purpose API build in accordance with the representational state transfer (REST) principles. This API is used for e.g. batch uploading or integration with external web sites or research community portals.

Persistent Identifiers (PIDs) are also supported in B2SHARE, based on the European Persistent Identifier Consortium technology (ePIC). All data in B2SHARE can be traced and linked via PIDs.

[11] https://www.eudat.eu/services/b2share

[12] https://invenio-software.org/

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

Development will be done so that B2SHARE can share the data known by B2SAFE using ePIC PIDs and referring to the location with the PIDs.

In principle, there are two alternatives for which system stores the actual data:

- Data stored in B2SAFE or other storage services
    - In this case, B2SHARE maintain a link to data and metadata of the data using PID.
    - B2SHARE is not responsible of maintaining the data and cannot ensure integrity of it.
- Data copied to B2SHARE
    - Standard way to make use of the service.

We have also considered features that would most likely simplify the use of metadata description files and transfer of metadata description files between B2 services. This is, however, beyond the scope of this deliverable.

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

3 Sensitive Data

3.1 Introduction

The aim of this work is to provide interoperability between existing EOSC tools and services to enable complex workflows involving personal data.

While the high level of maturity of EOSC services for non-sensitive data already allows the discovery, reuse and sharing of data across domain and geographical borders, the variety of the privacy legislations at regional level has prevented the design and implementation of solutions suitable for all sensitive and personal data. The sensitive data task aims at exploring some blueprints suitable for certain data and certain research workflows which will involve all the steps of the sensitive research data life cycle, from analysing personal data to sharing and archiving sensitive and non-sensitive results. The workflow will include the use of a B2SHARE instance (secure-B2SHARE) adapted to host anonymised metadata of sensitive/personal data stored in Sensitive Data Archive (SDA). Sensitive data will be accessible only through APIs provided by dedicated hosts. The interoperability will be achieved by using JSON web token-based APIs to connect the involved back ends, namely B2SHARE, Galaxy Portal[13] and the Sensitive Data Archive (SDA).

The analysis workflow procedure will go as follows:

1. Galaxy portal will retrieve encrypted data from SDA by using APIs.
2. Data analysis will be initiated by the user through Galaxy on the backend cloud HPC.
3. After processing, the outcome of the analysis can be stored back to SDA as Galaxy history elements.
4. Metadata from new objects will be published by secure-B2SHARE service for secure sharing.
5. The metadata from secure-B2SHARE will be harvested to B2FIND[14] service, from where it is discoverable or from where it can be re-harvested to other search engines.

By default, data transfers between services will have to be accepted by both parties before any access can be given. This will require common agreement about required informatic security level.

The task provides a solution that might cover some very actual use cases, for example related to genomic sequencing in the research against the COVID-19 pandemic. The idea is to investigate blueprints that enhance the interoperability of the different components in such a way to be agnostic to the underlying technology and the adopted analysis portal.

3.2 Task Activities

3.2.1 Risk analysis

This activity will produce a detailed risk analysis study on the legal assessment for the encryption solution with the collaboration of internal regulatory resources of the partners involved, with the aim to investigate possible barriers, workarounds or impacts especially in non-certified data-sharing settings. The potential risks are highly dependent on the actual implementation.

[13] https://galaxyproject.org/use/laniakea/

[14] https://eudat.eu/services/b2find

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

3.2.2 Data Analysis

SDA is a sensitive data archiving service for secure storage of sensitive datasets with multi-stage encryption. The service is used as the main platform for storing biomedical data in the European Genome Phenome Archive (EGA) federation. Data ingestion and export is supported through a JSON Web Token (JWT) based API. B2SHARE offers an HTTP REST API, that can be used by external applications or workflows. The API can also be used for metadata harvesting, although an OAI-PMH API endpoint is also provided for this purpose. Galaxy will leverage on both these APIs (B2SHARE and SDA) to import both metadata and real data into the workload system in order to easy the analysis for the end users. The Data analysis is implemented through the Laniakea on-demand cloud infrastructure[15].

3.2.3 Data Sharing and Discovery

Data publishing with secure B2SHARE has been discussed in detail in the EOSC-hub project WP 6.6. In the current activity previous plans should be checked and implemented. We are evaluating the possibility to enable the Galaxy workflow manager with the capability to export into B2SHARE secure service the output of the typical analysis workflow.

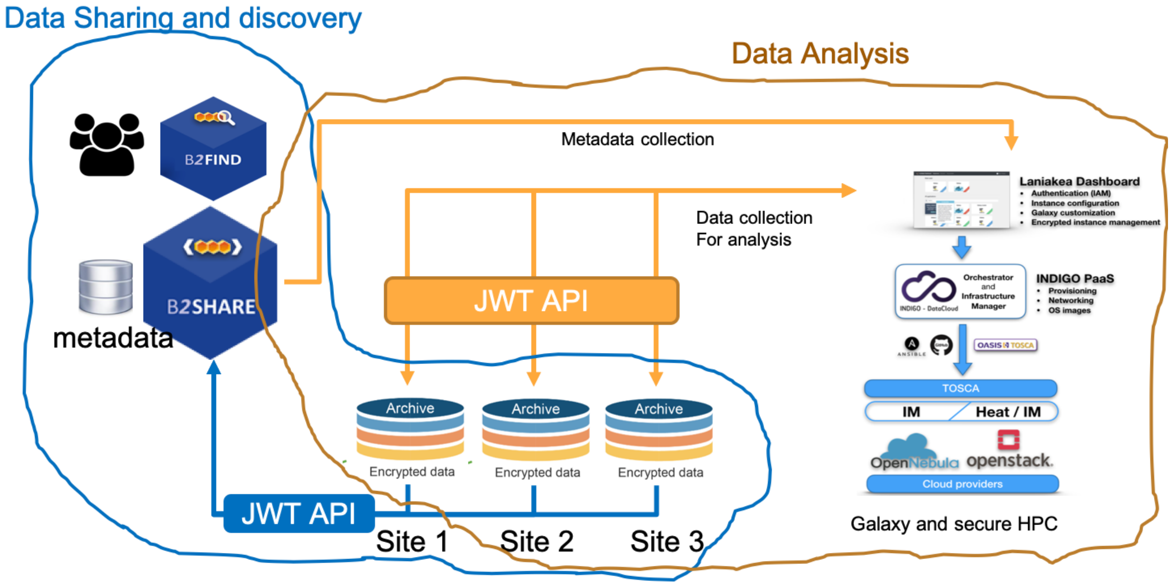

Figure 1. Secure data management and analysis model

[15] https://doi.org/10.1093/gigascience/giaa033

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

4 Conclusions

Integration of B2-services part of the DICE offering with computing environments and object storage is feasible with some development work as described in the respective sections.

The integration of B2DROP is easily attained as all the necessary parts and components are readily available.

Integration of B2SAFE has the highest amount of work to be carried out as there is the need to do an extensible development to integrate all the parts and components.

In conclusion, Task 4.1 “Compute and Analysis” is on the right path and the possible impact is vast; the positive outcome should be exploited to provide outstanding services for our users. As mentioned in the previous section, there are still some open questions that will need to be addressed and development works to be started as soon as possible.

The sensitive data task is working on an integration solution of three services: B2SHARE, SDA, and Laniakea Galaxy. The deployment targets two sensitive data e-Infrastructures: TSD (UiO), and CSC sensitive data service. The target is to enable researchers to securely store encrypted sensitive datasets, share the datasets with metadata description, and perform secure flexible data analysis using a user-friendly interface.