DICE

Data Infrastructure Capacity for EOSC

Deliverable D4.3

Final integration with other services & platforms

Responsible Partner:	Forschungszentrum Jülich
Status-Version:	Final - v1.2
Date:	30.06.2023
Distribution level (CO, PU):	Public

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

Title of Deliverable:	Final integration with other services & platforms
Due Date of Delivery to the EC	30.06.2023
Actual Date of Delivery to the EC	30.06.2023

Work package responsible for the Deliverable:	WP4 - Integration with other services & platforms
Editor(s):	Daniel Mallmann (FZJ)
Contributor(s):	T4.1 Compute and Analysis: Chris Ariyo, CSC T4.2 Discovery and Referencing: Tibor Kálmán, GWDG T4.3 Long Term Preservation: Wilko Steinhoff, DANS T4.4 Sensitive Data: Abdulrahman Azab, SIGMA
Reviewer(s):	Tonello, N. – BSC Testi, D. - CINECA
Recommended/mandatory readers:	WP5 Integration with community platforms WP2 Dissemination and outreach

Abstract:	This deliverable includes the final report about the integration of data services with computing platforms, the integration of PID Graph resources in B2FIND, the implementation of the LTP policy for B2SHARE in one CTS certified archive, the report on enabling sensitive data workflow by adapting standard interoperability frameworks to connect the endpoints.
Keyword List:	Compute, Analysis, Identifier, Long-term Preservation, 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	30.03.2023	Template for the tasks leaders	FZJ
v0.8	07.06.2023	Integration of the four documents with task contributions	FZJ, CSC, GWDG, DANS, SIGMA2
V1.0	20.06.2023	Internal review comments	BSC, CINECA
V1.1	22.06.2023	General Update addressing most of internal review comments	FZJ
V1.2	29.06.2023	Update of tasks according to internal reviewer comments	FZJ, CSC, GWDG, DANS, SIGMA2

Table of Contents

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

List of Figures

List of Tables

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

ASTRON	Astron
BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputacion
CESNET	CESNET, z. s. p. o.
CINECA	Cineca
CSC	CSC – Tieteen Tietotekniikan Keskus Oy
CyI	The Cyprus Institute
Datacite	DataCite
DKRZ	Deutsches Klimarechenzentrum GmbH
DoA	Description of Action
EC	European Commission
EOSC	European Open Science Cloud
ETHZ	Eidgenössische Technische Hochschule Zürich
EU	European Union
EUDAT ltd	EUDAT ltd
FZJ	Forschungszentrum Juelich Gmbh
GA	Grant Agreement to the project
GRNET	National Infrastructures for research and technology
GWDG	Gesellschaft für Wissenschaftliche Datenverarbeitung mbh Göttingen
HZB	Helmholtz-Zentrum Berlin für Materialien und Energie
INFN	Istituto Nazionale di Fisica Nucleare
IT4I	Vysoka Skola Banska - Technicka Univerzita Ostrava
IVOA	International Virtual Observatory Alliance
KIT	Karlsruhe Institut für Technologie
KNAW-DANS	Koninklijke Nederlandse Akademie van Wetenschappen
KPI	Key Performance Indicator
MPG	Max Planck Gesellschaft zur Foerderung der Wissenschaften e.V.
PID	Persistent Identifier
SIGMA	SIGMA2
SNIC	Uppsala Universitet
SURF	SURFsara BV
TRUST	Trust-IT services
TSD	The research platform for working with sensitive data at the University of Oslo (Norwegian name: Tjenester for Sensitive Data)
UCL	University College London
ULUND	University of Lund
VA	Virtual Access
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 DICE work package 4 fosters the integration of data services offered via DICE with European platforms and infrastructures. The deliverable 4.3 “Final integration with other services & platforms“ of the DICE projects compiles the contributions of the four tasks of WP4:

- Task 4.1 “Compute and Analysis” final report about the integration of data services with computing platforms,
- Task 4.2 “Discovery and Referencing” reports on the integration of PID Graph resources in B2FIND,
- Task 4.3 “Long Term Preservation of Data” describes implementation of the LTP policy,
- Task 4.4 “Sensitive Data” reports on enabling sensitive data workflow by adapting standard interoperability frameworks to connect the endpoints.

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

1 Introduction

D4.3 “Final integration with other services & platforms “is the final deliverable WP4 “Integration with other services & platforms” of the DICE project.

It comprises contributions from all four tasks, which perform independently of each other.

Task 4.1 “Compute and Analysis” describes in chapter 2 the work realized for the integration of EUDAT data services with computing platforms to enable analysis, data replication, and data publication in a generic way. Two use cases are provided, one that illustrates how the ICOS community exploits the integration results of task 4.1, and a second, that shows how B2SHARE is used in a research community portal for using or publishing data on B2SHARE.

Task 4.2 “Discovery and Referencing” reports about the integration of PID Graph resources, namely PIDs for scientific instruments, people IDs and repository IDs, into B2FIND, and the productization of B2INST. Use cases of two communities describe the implementation of instrument information in B2FIND: Helmholtz-Zentrum Berlin für Materialien und Energie (HZB) and the International Virtual Observatory Alliance (IVOA).

Task 4.3 “Long Term Preservation of Data” authored the policy templates for Long-term Preservation services and implemented them for B2SHARE and B2FIND. In addition, T4.3 developed and implemented an interface for the archiving of data records from B2SHARE in a digital preservation service.

Task 4.4 “Sensitive Data” enhanced the Secure B2SHARE service and deployed it in two Sensitive Data Infrastructures at CSC and the University of Oslo. T4.4 included B2FIND to improve discoverability of sensitive data stored in a Secure B2SHARE instance and designed the integration of processing services outside of the sensitive data infrastructure.

To measure the success of WP4, we defined the KPI #9 “DICE services integrated with other platforms”. The KPI was expected to be “6 services/platforms” at the end of the project. The number originated from the expectation of two services for each task, with task 4.3 being policy work and not contributing to this KPI. However, task 4.3 developed an extension of the B2SHARE service and integrated it with a data preservation service in addition to the two policy templates, being not an integration with other services/platforms but are very valuable results for B2SHARE and B2SAFE. All in all, WP4 reached 12 integrations for the whole duration of the project:

Task 4.1 “Compute and Analysis”

1. B2DROP with HPC-Systems and Jupyter-Hub at JSC
2. B2SAFE integration with HPC-Systems at CSC
3. Data transfer from computing platforms to B2SHARE for publishing

Task 4.2 “Discovery and Referencing”

4. B2INST in B2FIND
5. ORCID in B2FIND
6. Repositories of re3data, Fairsharing and OpenDOAR in B2FIND
7. B2FIND in OpenAIRE Explore
8. Integrity check for PID infrastructures in B2HANDLE (described in D4.2)
9. Integrity check for PID metadata in B2HANDLE (described in D4.2)

Task 4.3 “Long Term Preservation of Data”

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

10. Data Preservation Service for B2SHARE

- *(Long-term Preservation Policy Template for B2SHARE)*
- *(Long-term Preservation Policy Template for B2SAFE)*

Task 4.4 “Sensitive Data”

11. Secure B2SHARE in CSC Sensitive Data Infrastructure
12. Secure B2SHARE in TSD research platform at the University of Oslo

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

2 Final report about the integration of data services with computing platforms (T4.1)

2.1 Introduction

The aim of task T4.1 "Compute and Analysis" is to integrate EUDAT data services with computing platforms to enable analysis, data replication, and data publication for high performance computing and cloud computing platforms.

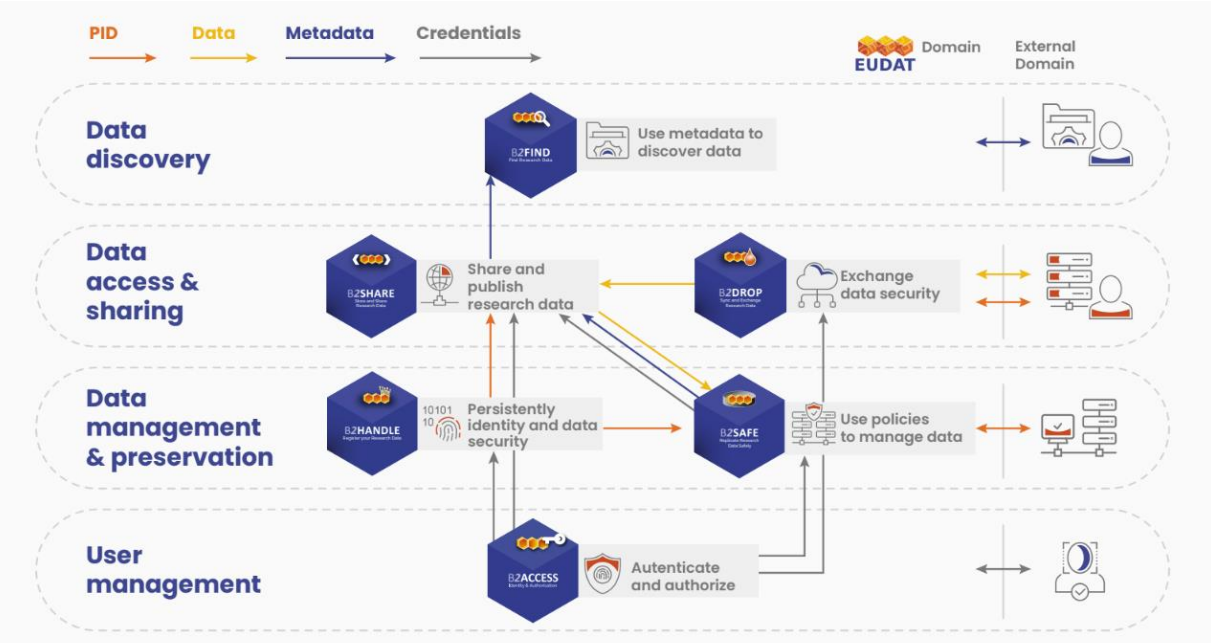

PID
Data
Metadata
Credentials
EUDAT Domain
External Domain
Data discovery
B2FIND
Use metadata to discover data
Data access & sharing
B2SHARE
Share and publish research data
B2DROP
Exchange data security
Data management & preservation
B2HANDLE
10101
Persistently identity and data security
B2SAFE
Use policies to manage data
User management
B2ACCESS
Autenticate and authorize

Figure 1. EUDAT CDI data services

This is done by using B2DROP service to store, share and hold small datasets for analysis, B2SAFE service to register large datasets, and B2SHARE service to publish datasets.

In this deliverable, we describe and show the preparations, steps and commands necessary for integration of EUDAT data services and computing platforms:

- Section 2.2: B2DROP
  Ensure that small data (batch queue scripts etc. similar small data objects) can be read from B2DROP to computing environment and that small data can be written back to B2DROP

- Section 2.3: B2SAFE
  Data transfer between B2SAFE and computing platforms and also data transfer between computing platforms and object storage systems

- Section 2.4: B2SHARE
  Transfer data from computing platforms to B2SHARE for publishing.

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

2.2 B2DROP use case

B2DROP[1] is a low-barrier, user-friendly and trustworthy storage environment which allows users to synchronize their active data across different desktops and to easily share this data with peers with the following features:

- Access via Web GUI, desktop clients and WebDAV
- Multiple versions of files are kept
- Enabled apps: Contacts, Calendar, Tasks, Circles (social communities)
- Sharing within B2DROP, across different instances (via OCM-API) and via links
- Publishing of datasets to B2SHARE

B2DROP offers two ways to make files accessible to HPC environments. The first way is to use the web browser to create a "Share link", which can then be accessed with any https client tool such as wget, curl etc. The second way is to use the WebDAV protocol, which allows for more advanced use cases.

2.2.1 Using shared links

Shared links are primarily intended for usage via the web browser, but for simple cases, they offer a quick and convenient way to download data files from almost anywhere.

To create a new shared link, the user needs to click the share icon next to a file/folder in B2DROP. B2DROP will then create a new endpoint that the user can visit with the web browser. Shared links do not require additional authentication and can be used from anywhere where an internet connection to B2DROP is possible, including login nodes of HPC clusters.

To download the file from HPC using a command-line tool, the user needs the shared link with an extra "/download" appended to the URL, for example:

```
wget -content-disposition https://b2drop.eudat.eu/
                          s/7mAX6FqfmMz7afF/download
```

This will download the file and write it to the local disk under its original name.

Shared links offer additional possibilities such as automatic expiry and can be removed by the owner at any time. However, there is no easy-to-use way to upload files from a command-line environment.

2.2.2 Using Web Distributed Authoring and Versioning (WebDAV)

WebDAV is a much more capable protocol[2], providing extensions to HTTP allowing for advanced file system operations like upload, listing and creating directories etc. The WebDAV endpoints offered by B2DROP require authentication via "app secrets" (consisting of the B2DROP user ID and a password).

The first step in using WebDAV is therefore creating an "app secret" via the user settings in B2DROP at this link: https://b2drop.eudat.eu/settings/user/security

Using the newly created username/password, the user can now use any WebDAV client or any other HTTP client tool such as curl to access their data via the WebDAV protocol. This is possible

[1] https://eudat.eu/catalogue/b2drop

[2] https://github.com/fstanis/awesome-webdav

from all HPC nodes with internet access, typically, these will be the login nodes or dedicated data transfer nodes.

The B2DROP WebDAV interface[3] accesses the same files/directories that the web browser sees at the URL https://b2drop.eudat.eu/apps/files/ so all files owned by and shared with the user are accessible. Since WebDAV allows easy uploading, use cases that involve writing results back to B2DROP are then possible.

2.2.3 Challenges in using WebDAV for automounting on HPC & computing systems

Currently, it's not possible to use B2DROP on HPC systems with WebDAV for mounting B2DROP as a filesystem, because an entry in /etc/fstab is required to use the automatic mount of WebDAV without root access using DAVfs[4]. HPC centers do not have this entry written in the /etc/fstab of their supercomputers, because it would specify one mount point for the whole system for all users. At FZJ, several solutions have been discussed. The most promising solution was not updated since May 2020[5] and another probable contender, Davix[6], was examined as well. FZJ added the mounting with DAVfs in Jupyter-JSC[7] and users can now use directly access their B2DROP files on the Cloud resources. Since users start their JupyterLab in a container, every resource can use the same mount path and they don't share a whole filesystem but only have access to their files.

2.3 B2SAFE use case

B2SAFE[8] is a robust and highly available service which allows community and departmental repositories to implement data management policies on their research data across multiple administrative domains in a trustworthy manner. It offers an abstraction layer of large scale, heterogeneous data storages, guards against data loss in long-term archiving, allows to optimize access for users (e.g. from different regions), and brings data closer to facilities for compute-intensive analysis. This service is based on Integrated Rule-Oriented Data System (iRODS)[9].

Features of B2SAFE:

- Support for data management policies (e.g. registration of PIDs, cross-site replication, data integrity checks)
- Support for policies customized to community and organizational needs
- Support for less frequently used archival data, but can also support active data
- Support for large scale storage resources (e.g up to PB-scale)
- A single namespace across heterogeneous storages
- Supports integration with different kind of storage systems (e.g. Tape based HSM, POSIX filesystems, Object storage)
- Access via GridFTP, Webdav, iRODS commands
- Service offered by a network of EUDAT service providers

B2SAFE service, at the core, exploits the iRODS rule engine to perform a set of actions to implement specific behavior defined in data management policies. The actions are defined by a set of iRODS rules, which can either be executed on a regular basis or be triggered by actions

[3] https://b2drop.eudat.eu/remote.php/webdav
[4] https://sabre.io/dav/clients/davfs/
[5] https://github.com/miquels/webdavfs
[6] https://davix.web.cern.ch/davix/docs/devel/
[7] https://jupyter-jsc.fz-juelich.de
[8] https://eudat.eu/catalogue/b2safe
[9] https://irods.org/

like data ingestion. The rules interact with external software components, which deliver functionalities such as PID registration.

The B2SAFE module also offers rules for integrity checks across zones, recovering failed transfers and updating the information on data location in the PID system in case of changing the iRODS path to the data. Furthermore, the ruleset contains experimental features like community metadata handling and messaging.

B2SAFE offers safe data replication across different data centers. Communities, repositories, and data projects can use B2SAFE to distribute valuable data across the EUDAT network to keep it safe and to bring it closer to compute infrastructures.

B2SAFE offers a few ways to make files accessible to High Performance Computing (HPC) and cloud computing environments.

1. HTTP-API[10] protocol
2. Web Distributed Authoring and Versioning (WebDAV)[11] protocol
3. GridFTP[12] File Transfer Tool

2.3.1 Data transfer between B2SAFE and HPC

This use case concentrates on using the HTTP-API available in B2SAFE to transfer data between B2SAFE and the HPC systems Mahti[13] and Puhti[14] and is also applicable to Lumi[15] and other HPC systems.

The use of the HTTP Rest API protocol is very important as this makes integration and data transfer to and from B2SAFE possible in most if not all Unix based computing platforms. To use this protocol, you need to provide the following:

- Authorization header needs to be provided on each request.
- Authorization is basic authentication username/password combination to the backend.

The username and password must be provided from the B2SAFE system that will be used. The detailed usage and configuration of the use case is summarized in Appendix 1 of this deliverable.

For the usage of the HTTP-API interface we focus on the usage of the command line tool and library (cURL)[16] for transferring data with URLs. The commands for listing and creating collections and uploading or deleting files are added to Appendix 1.

[10] https://gitlab.com/noumar/http-api/-/blob/master/DESCRIPTION.md
[11] http://www.webdav.org
[12] https://fasterdata.es.net/data-transfer-tools/gridftp/
[13] https://research.csc.fi/-/mahti
[14] https://research.csc.fi/-/puhti
[15] https://www.lumi-supercomputer.eu/
[16] https://curl.se/

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

2.3.2 Data transfer between Object Storage and HPC

Accessing CSC – IT Center for science object storage (Allas[17]) in the CSC computing environments will be used for this use case. To transfer data between Allas and the HPC systems Mathi[18] and Puhti we have the following possible tools[19]:

- A-commands
- Swift commands
- S3cmd commands
- Rclone command

We are going to concentrate on using the Rclone[20] command line file transfer tool on HPC to transfer data to and from the object storage system (Allas) in this use case, because Rclone provides a very powerful and versatile way to use Allas and other object storage services. It is able to use both the S3 and Swift protocols (and many others), but in the case of Allas, the Swift protocol is preferred. It is also the default option on the CSC servers. The usage is applicable to other object storage systems connected to HPC systems.

The rclone configuration and commands for creating, listing, and downloading objects are added in Appendix 1 of this deliverable.

The access to Allas on Puhti or Mahti is supported by a dedicated module and configuration command (allas_conf. The allas-conf command prompts for your CSC password (the same that you use to login to CSC servers). It lists your Allas projects and asks you to define a project (if not already defined as an argument). allas-conf generates an rclone configuration file for the Allas service and authenticates the connection to the selected project.

You can only be connected to one Allas project at a time in one session. The project you are using in Allas does not need to match the project you are using in Puhti or Mahti, and you can switch to another project by running allas-conf again.

2.3.3 Better integration of B2SAFE to B2ACCESS

The aim here was to better improve the usability of B2SAFE by better integrating it with B2ACCESS[21]. The possibility to do this will rely completely on the availability of OpenID Connect[22] in the coming iRODS version. Due to this we could not test this at all.

2.3.4 Integrated Carbon Observation System (ICOS) community use case for data transfer to B2SAFE

The aim of this use case is to get ICOS[23] data transferred into EUDAT B2SAFE in CSC and from there data can be replicated to Jülich Computing Center (JSC). The data can then be transferred to an HPC system in CSC or JSC for computation and analysis. The result from the computation or analysis is then uploaded back into B2SAFE.

[17] https://research.csc.fi/-/allas
[18] https://research.csc.fi/-/mahti
[19] https://docs.csc.fi/support/faq/how-to-move-data-between-puhti-and-allas/
[20] https://rclone.org
[21] https://eudat.eu/catalogue/b2access
[22] https://openid.net/connect/
[23] https://www.icos-cp.eu/

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

To create a federated B2SAFE between CSC and JSC, there are certain general setup steps needed. The following workflow applies in order to join the B2SAFE federation.

- Deployment of an iRODS/B2SAFE instance[24]
- Agreements on iRODS federations with other EUDAT centers and community centers
- Entry in the Resource Coordination Tool (RCT) registry making the new B2SAFE node known to EUDAT

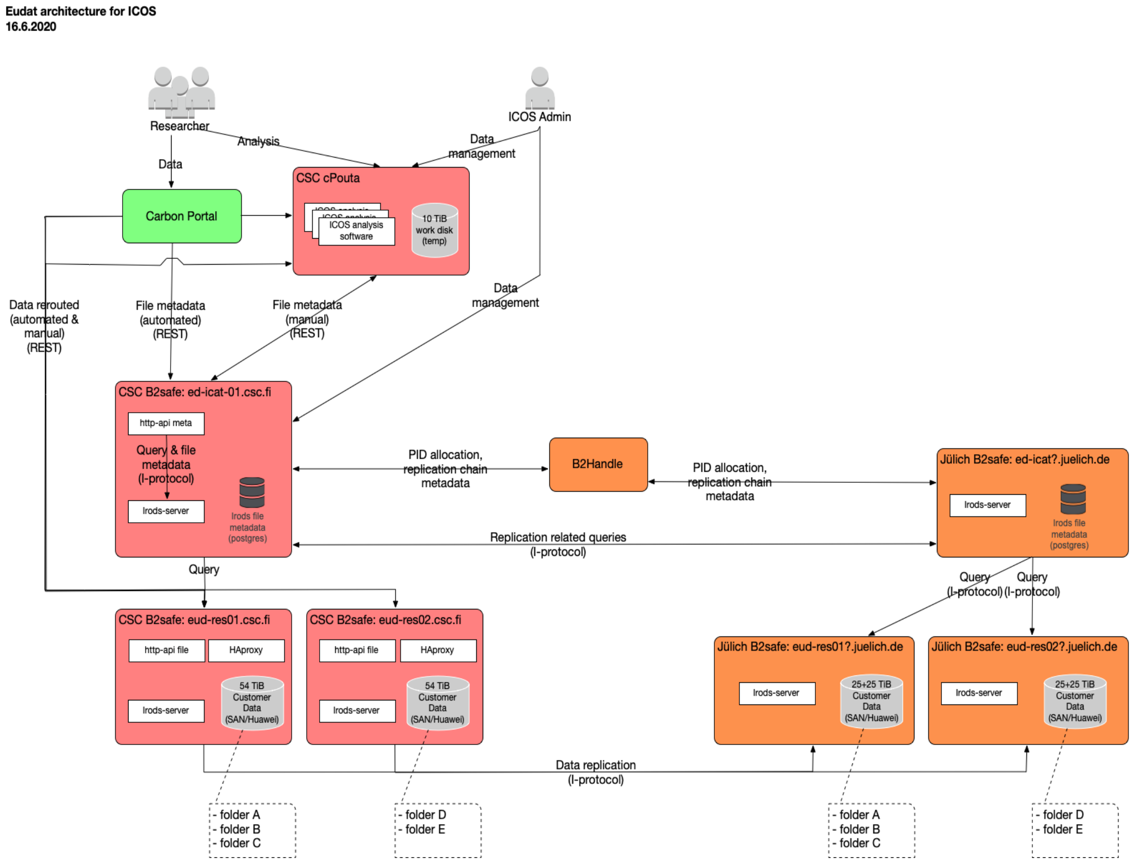


Figure 2. ICOS server architecture and data flow

Figure 2 shows the interactions between ICOS and the two B2SAFE cases. The automatic replication of the data is not yet in production phase.

Research data is transferred from ICOS to CSC and back using the HTTP REST API.

2.4 B2SHARE use case

B2SHARE[25] is a user-friendly, reliable, and trustworthy way for researchers, scientific communities and citizen scientists to store, publish and share research data in a FAIR way. B2SHARE is a solution that facilitates research data storage, guarantees long-term persistence of data and allows data, results or ideas to be shared worldwide. B2SHARE supports community domains with metadata extensions, access rules and publishing workflows. EUDAT offers communities and organizations customized instances and/or access to repositories supporting large datasets.

[24] https://documentation.eudat.eu/b2safe/foradministrators/
[25] https://eudat.eu/catalogue/b2share

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

B2SHARE has the following features:

- Support of metadata descriptions via the EUDAT metadata schema
- Registers DOIs for datasets and Handle PIDs for data objects
- Supports versioning
- Harvested by B2FIND[26] and OpenAIRE explorer[27]
- Direct upload from B2DROP
- Accessible via a Web GUI and an HTTP REST API to support automatic publishing workflows
- Supports community domains
- Allows communities to define metadata extensions, access rules and publishing workflows

In addition to its web-based GUI, B2SHARE offers an HTTP REST API[28]. The B2HARE HTTP REST API can be used for interacting with B2SHARE via external services or applications, for example for integrating with other websites (research community portals) or for uploading or downloading large data sets that are not easily handled via a web browser or computing platforms. This API can also be used for metadata harvesting.

Certain HTTP REST API requests to the B2SHARE service require authentication, for example to create or modify draft records. Each such request to the server must provide an access token parameter that identifies the user. The access token is an opaque string, which can be created in the user profile when logged in to the B2SHARE web user interface[29]. B2SHARE's access tokens follow the OAuth 2.0 standard.

2.4.1 HTTP REST API token generation from the B2SHARE web user interface

Enter a name in the text field below 'Create new token:' which easily identifies the purpose for this key. By clicking on New Token, a new access token is generated which is only shown at this time. Store it somewhere in order to use later, like in a file. It is assumed that the generated token is stored in a file called token.txt.

Notes

1. Please note that this is the only time the access token is visible, so copy it to a safe place.
2. It is not possible to programmatically register new or administer existing tokens. This can only be done through the B2SHARE web interface.
3. The generated token is unique for all instances and therefore can only be used for the instance you created it on. A token of the production instance of B2SHARE can therefore not be used on the training instance and vice-versa or in other B2SHARE instances!

2.4.2 A publication workflow

When your dataset is ready for publication, it can be uploaded to the B2SHARE service by creating a draft record and adding files and metadata. This page will guide you through the creation process of a new draft record, preparing and finally publishing it as a record. It covers:

[26] https://eudat.eu/b2find
[27] https://explore.openaire.eu/
[28] https://documentation.eudat.eu/b2share/httpapi/
[29] https://b2share.eudat.eu/

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

- The creation of a new draft record,
- The addition and removal of files and metadata, and
- Committing the draft record to publish it

The HTTP API does not impose a specific workflow for creating a record. The following example workflow only defines the most basic steps:

1. Identify a target community for your data by following the HTTP API List all communities guide
2. Using the community's identifier, retrieve the community's JSON Schema of the record's metadata. The submitted metadata will have to conform to this schema. Use the Get community schema guide to achieve this
3. Create a draft record: follow the Create draft record guide to create a draft record with initial metadata in it
4. Upload files into the draft record
5. Set the complete metadata and publish the record

In Figure 3 the general deposit workflow of B2SHARE is shown. All blue boxes require a request interaction with the B2SHARE service.

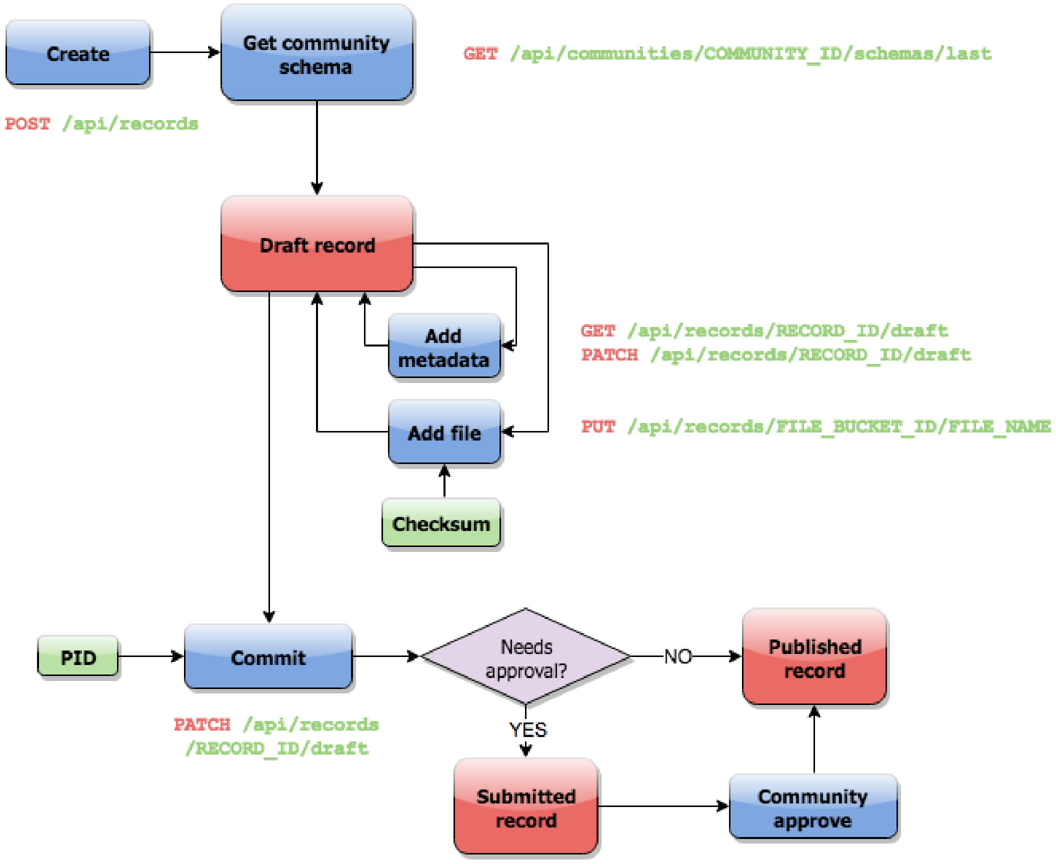

Create
Get community schema
GET /api/communities/COMMUNITY_ID/schemas/last
POST /api/records
Draft record
Add metadata
GET /api/records/RECORD_ID/draft
PATCH /api/records/RECORD_ID/draft
Add file
PUT /api/records/FILE_BUCKET_ID/FILE_NAME
Checksum
PID
Commit
Needs approval?
NO
Published record
PATCH /api/records /RECORD_ID/draft
YES
Submitted record
Community approve

Figure 3. General deposit workflow for B2SHARE

The red boxes indicate an object state, where in this workflow only draft, submitted and published records exist. Files and metadata can be added multiple times. Persistent identifiers (PIDs) and checksum are automatically added by B2SHARE (green boxes). Once a draft record is

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

committed, depending on the community's requirements, the record is either in submitted state and needs further approval or is immediately published.

In Appendix 1 of this deliverable, we added the detailed configuration and commands for publishing data on B2SHARE using cURL tool.

2.5 Challenges in using of JWT (JSON Web Token) Authentication Tokens Security

JSON Web Token (JWT)[30] is an industry standard that has been widely used by many companies. The security of this standard depends very much on the right implementation.

- JWT tokens are not inherently insecure, i.e., they are not insecure by design.
- As with any software, vulnerabilities in *implementations* of JWT are found from time to time, which are usually quickly resolved by the vendor of that software
- JWT implementations rely on underlying software such as cryptographic libraries or JSON parsers, which may have vulnerabilities themselves, which are corrected quickly
- Vulnerabilities related to JWT in various software products[31]
    - As of today, there is only one **(CVE-2022-25898**) with a workaround

All in all, we do NOT see any particular danger in using JWT. On the contrary, its widespread usage and high availability of well-tested implementations make it an excellent choice for tasks such as authentication, authorization and delegation.

2.6 Challenges

We faced some challenges in the process of implementation and connection of some services to computing and HPC systems:

- Automatic WebDAV mounting of B2DROP requires root access, this limits the user experience when using HPC systems. There are workarounds that could be used.
- Connection of B2SAFE to B2ACCESS using OpenID Connect (OIDC) is not available in current core technology (iRODS) but expected in a later date.

[30] https://jwt.io/introduction
[31] https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=jwt

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 Integration of PID Graph resources in B2FIND (T4.2)

3.1 Introduction

The persistent identification of research output is crucial to enable FAIR data management practices, in particular when it comes to metadata. The idea behind the concept of persistent identification is not only to have a persistent identifier (PID) for datasets, but also to use PIDs to link information about people, institutions, repositories, software, instruments, etc. that are part of the research output. Several services for registering persistent identifiers already exist and some of them expose their PID information as graphs. Within the framework of the DICE project, several persistent identifiers have been included in the B2FIND ingestion software (as well as into the discovery portal) in a way that it paves the path for future work of integrating further graph resources. Benefits and challenges for this inclusion are described below in the first section. In terms of interoperability, maintenance and adoption of software are a basic issue, our efforts for that are described in the second section. Here the difficulty is the different approaches of B2FIND as (1) mainly an interdisciplinary discovery portal and (2) as well as metadata aggregator and provider to OpenAIRE (and therefore to EOSC). How we meet the needs of both approaches is described in that section as well. Finally, a conclusion will summarise the main insights and potential further enhancements after the project ends.

3.2 PID information

As mentioned above, there are several ways to include information about research output using persistent identifiers. Within the DICE project we focused on the integration of PIDs for (1) scientific instruments, (2) people IDs and (3) repository IDs. Instrument PIDs were selected, because (1) there is a strong need to identify instruments in a persistent way in many communities and (2) within the Research Data Alliance there has been a dedicated working group[32] dealing with the development of minimal metadata for PIDs for instruments. To develop and establish a registry as a productional service and to include the outcome within an interdisciplinary search portal has been a great effort. Thus, prerequisites and the operational workflow (implemented in three already existing services) are described in more detail here, in particular as the result validates the conceptual framework.

With reference to enhanced discoverability of people being involved in research outputs, Open Researcher and Contributor IDs (ORCID) have been integrated into B2FIND. For better provenance information, several Repository IDs (re3data, Fairsharing and OpenDOAR) have been integrated in B2FIND ́s metadata export.

3.2.1 Instruments

Scientific instruments are used by very divergent communities in very different ways. While information about instruments is already implemented within the internal workflows of several communities, it is not that easy to exchange metadata about instruments using a generic metadata schema. EUDAT core[33] (with the purpose of transferring metadata information across different EUDAT CDI services) has already included an optional metadata element <Instrument> with allowed attributes <instrumentIdentifier> and <instrumentIdentifierType>[34], which is used

[32] https://www.rd-alliance.org/groups/persistent-identification-instruments-wg
[33] https://schema.eudat.eu/eudatcore/
[34] https://schema.eudat.eu/eudatcore_metadataelements/#instrument-o

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

e.g. for the mapping of Blue-Cloud[35]. As there has been some effort to include persistent identifiers for scientific instruments, we will present a workflow on how that information is made available in B2FIND for (1) a community that uses the EUDAT service B2INST for registering PIDs for instruments and (2) a community that is using DOI for persistent identification of instruments. But before doing that we will describe the work done for developing a new service to register scientific instruments in a persistent way.

3.2.1.1 B2INST service - from prototype to production

B2INST is an emerging EUDAT service for registering scientific instruments and a collaborative solution driven by the scientific community, aiming to provide a global and unique identification system for instruments used in scientific research. Instruments play a crucial role in various fields, such as environmental science, life sciences, and medical domains, encompassing sensors, DNA sequencers, microscopes, and more. B2INST serves as a public service, allowing research communities and individual researchers to describe, register, and reference their instruments.

Before the DICE project started, there had been a proof-of-concept implementation of B2INST [1] and this prototype was taken up by DICE. It was built on the B2SHARE technology offered by EUDAT but is slightly modified to serve the purposes of B2INST, as the B2SHARE technology was designed for repositories and B2INST has a focus on flexible registration of metadata that supports different kinds of communities and their corresponding different metadata schemas. We summarize the development activities and new features we implemented in DICE at Table 1.

Table 1. DICE Developments in B2INST Service

Status Prototype ("before DICE")	DICE Enhancements
Metadata Schema ("root schema") is similar to RDA's PIDINST schema, but not compatible	Instrument metadata is compatible with RDA's endorsed PIDINST schema
Community extensions ("community schema") cannot be supported (software bugs)	Full support for community extensions ("community schema")
Validation of instrument metadata ("records") against the "root" schema only	Validation against both the community schema (if any) and the root schema assigned
Minting DOIs with very limited metadata	Integrating the functionality to create Handle PIDs with rich metadata

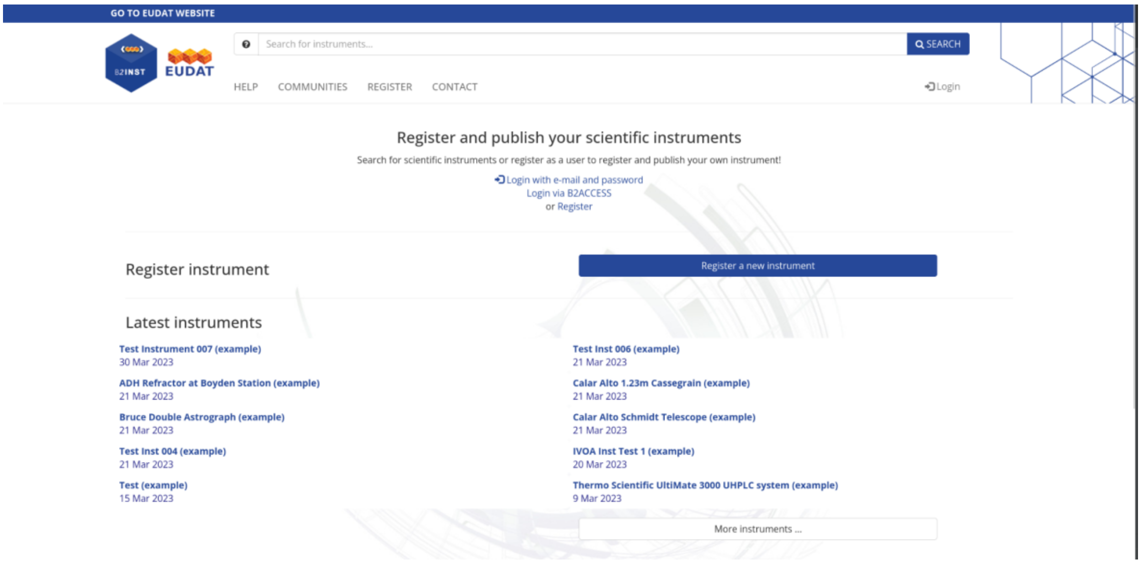
[35] Blue-Cloud is a Horizon2020 Project for Marine Research that aims to federate already existing infrastructure to create a trusted virtual space where scientists can access the ocean data, tools, services and research outputs they need to perform research in a more efficient way. B2FIND has integrated metadata records from the Blue-Cloud Data Discovery & Access service (https://blue-cloud.org/services/blue-cloud-data-discovery-and-access-service) including the specific names of the Argo profiling float and the measurement instrument (Argo is an international array of about 4,000 profiling floats that measure temperature and salinity throughout the global oceans, down to 2,000 meter). Those names are searchable within the facet <Instrument> for all Blue-Cloud records: https://b2find.eudat.eu/organization/bluecloud.

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

Travis Pipeline with prototypical actions (such building of docker image, running tests and creating test reports).	GitLab Continuous Integration pipeline with several actions in an automated fashion (building Docker image, deploying the image, checking whether the correct images are up and running, testing whether the images are configured correctly).
User logins via B2ACCESS only	Improving federated AAI support to integrate further communities by adding OIDC based single-sign-on, as well as authentication based on local username/password.
No documentation available	Writing EUDAT documentation for B2INST for users, administrators, and developers. Integrating this into https://doc.eudat.eu/
Metadata schema and metadata fields are not documented	Writing documentation for the EUDAT Instruments schema and integrate this into https://schema.eudat.eu/
Basic usability (prototype)	Enhanced user experience (several UI improvements, like prefilled instrument identifier schema version field, etc.)

B2INST provides a REST API, as well as a graphical, web-based user interface, which is designed to be self-explanatory. Viewing the instrument information or using the search functionality are publicly available without registration or authentication. A search field is available for all users at the top of the B2INST home page (see Figure 4). Text and keywords can be entered in the search field. The text can be part of a title, keyword, abstract or any other metadata.

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



Register and publish your scientific instruments

Search for scientific instruments or register as a user to register and publish your own instrument!

Login with e-mail and password
Login via B2ACCESS
or Register

Register instrument

Register a new instrument

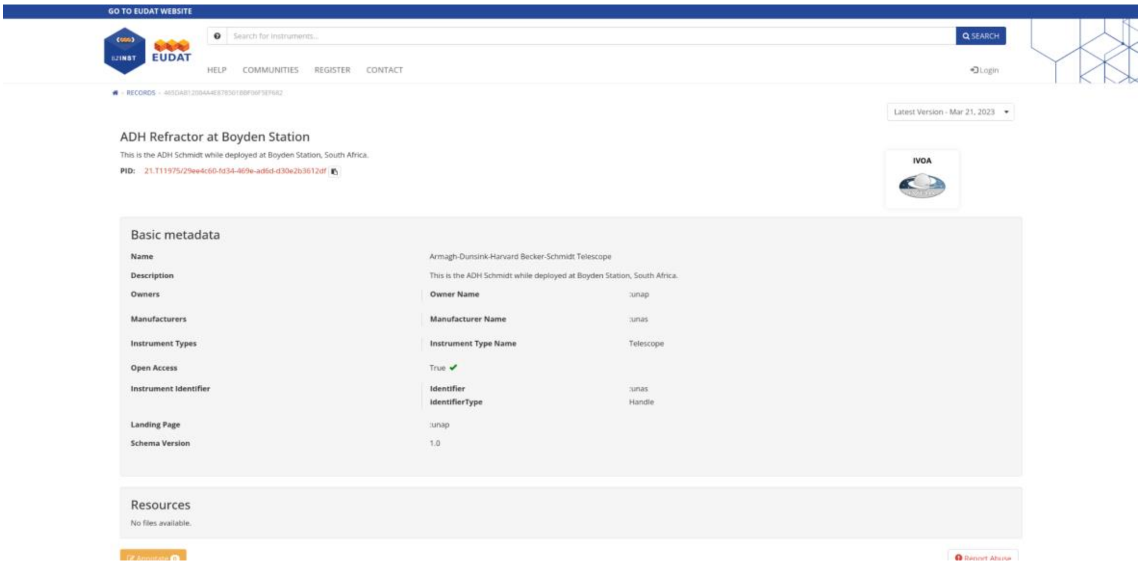
Latest instruments

Test Instrument 007 (example) 30 Mar 2023	Test Inst 006 (example) 21 Mar 2023
ADH Refractor at Boyden Station (example) 21 Mar 2023	Calar Alto 1.23m Cassegrain (example) 21 Mar 2023
Bruce Double Astrograph (example) 21 Mar 2023	Calar Alto Schmidt Telescope (example) 21 Mar 2023
Test Inst 004 (example) 21 Mar 2023	IVOA Inst Test 1 (example) 20 Mar 2023
Test (example) 15 Mar 2023	Thermo Scientific UltiMate 3000 UHPLC system (example) 9 Mar 2023

More instruments ...

Figure 4. Web-based user interface of B2INST

Once a record has been found using the search functionality or directly on the homepage of the B2INST service, by clicking on the title a so-called landing page is shown (s. Figure 5). This page displays the data of the instrument record, like the files and metadata. Each record has one or more files attached together with metadata structured according to the metadata schema of the corresponding community under which the record was originally published. For each file, the file size, checksum and PID are shown.



ADH Refractor at Boyden Station

This is the ADH Schmidt while deployed at Boyden Station, South Africa.

PID: 21.T11975/29ee4c60-fd34-469e-ad6d-d30e2b3612df

Basic metadata

Name	Armagh-Dunsink-Harvard Becker-Schmidt Telescope	
Description	This is the ADH Schmidt while deployed at Boyden Station, South Africa.	
Owners	Owner Name	:unap
Manufacturers	Manufacturer Name	:unas
Instrument Types	Instrument Type Name	Telescope
Open Access	True ✓	
Instrument Identifier	Identifier IdentifierType	:unas Handle
Landing Page	:unap	
Schema Version	1.0	

Resources

No files available.

Figure 5. Landing page of a registered instrument in B2INST

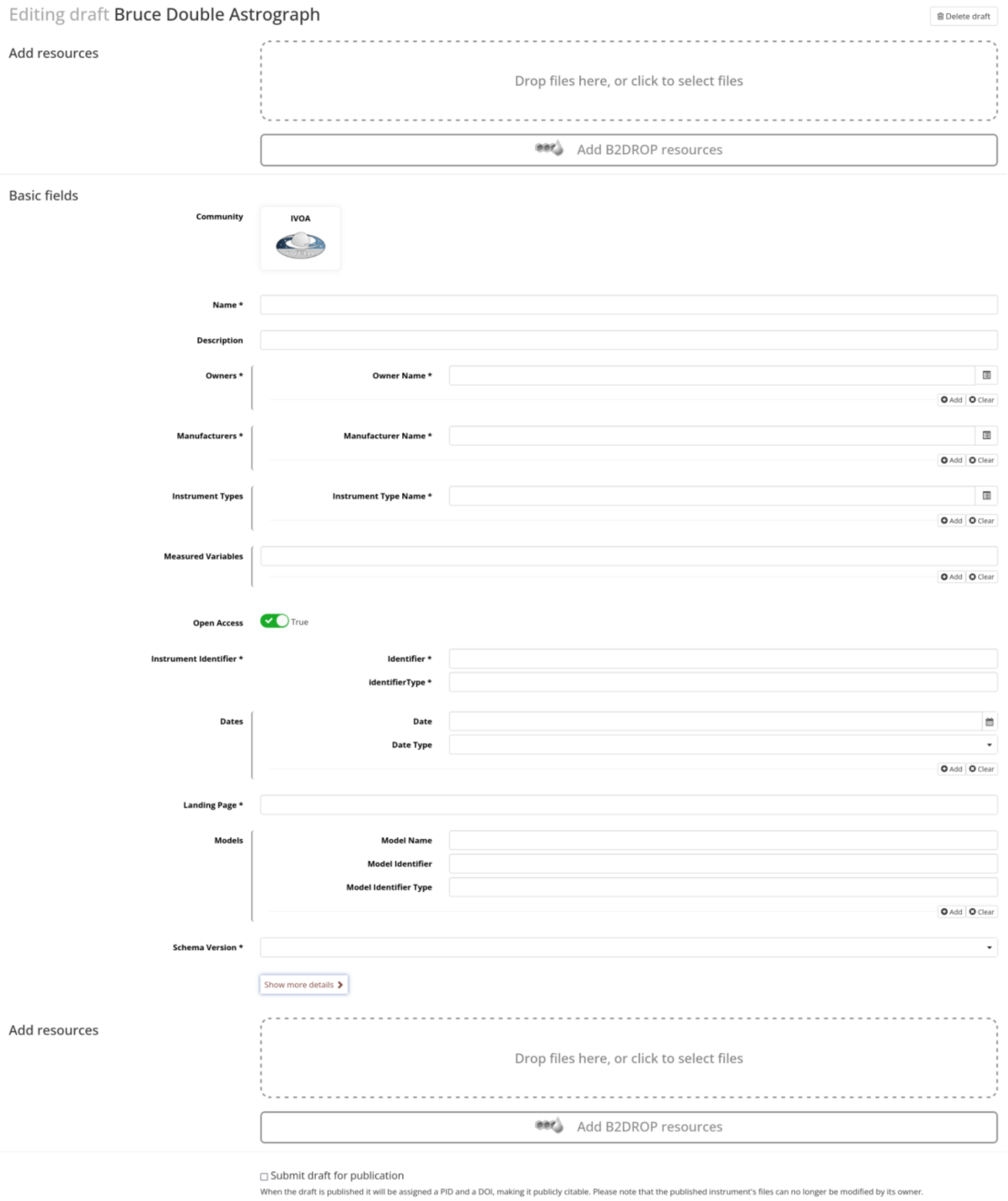
3.2.1.2 Metadata Schema development in B2INST

For B2INST, we developed and implemented the EUDAT Instruments schema. The basis for this schema is provided by RDA's PIDINST Working Group [2].

During the examination of various community use cases for instruments, it has been observed that each community has distinct metadata requirements. The metadata schema developed by

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

the PIDINST working group and being endorsed by the Research Data Alliance [3], while comprehensive, only encompasses a minimum set of very common metadata elements for instrument descriptions. To address this diversity in metadata needs, the B2SHARE technology offers the possibility of extending metadata through community-specific extensions over its joint root schema. Using the root schema as a foundation, communities have the flexibility to enhance metadata capabilities by incorporating extensions tailored to their specific needs. This enables the inclusion of more comprehensive instrument descriptions and facilitates compliance with the unique requirements of different communities. By leveraging community extensions, the B2SHARE platform ensures a more robust support system for a broader range of instrument metadata requirements. Figure 6 shows an example for supported metadata fields when a new instrument is being registered.



Editing draft Bruce Double Astrograph

Add resources

Basic fields

Community: IVOA

Name *

Description

Owners * — Owner Name *

Manufacturers * — Manufacturer Name *

Instrument Types — Instrument Type Name *

Measured Variables

Open Access: True

Instrument Identifier * — Identifier *, identifierType *

Dates — Date, Date Type

Landing Page *

Models — Model Name, Model Identifier, Model Identifier Type

Schema Version *

Add resources

□ Submit draft for publication

When the draft is published it will be assigned a PID and a DOI, making it publicly citable. Please note that the published instrument's files can no longer be modified by its owner.

Figure 6. Example for supported metadata fields when a new instrument is registered

Currently, the instrument schema is being fully documented[36]. Additionally, we have created a type definition for the RDA's PIDINST schema in a Type Registry[37]. Data types characterise data at any granularity. They are registered and used by humans and machines. Users often encounter unknown data types within an id, type, and value triple. Data type registries provide structure for processing such data. Users can query registries for type definitions, relationships, properties, and pointers to relevant services or software. The Data Type Registry addresses parsing, understanding, and reusing data. It serves as a repository for managing these types, promoting data sharing in an interconnected world with complex problems and high interoperability among sources. Moreover, efficiently processing vast scientific data requires automated parsing without human intervention. Defining and associating types with data is crucial for optimised interactions. Standardised, unique, and discoverable types are necessary. That was our motivation to create the type definition for RDA's PIDINST.

We have been integrating the latest developments, how Type Definitions should be registered in Data Type Registries. This is however an ongoing development, and we will continue the discussion with the community in the FAIRCORE4EOSC project.

3.2.1.3 Metadata Schema enhancement in DataCite to support Instruments

Following the RDA Persistent Identification of Instruments Working Group discussions, DataCite has extended the resource types to include instruments. This includes:

- Addition of Instrument[38] to the resourceTypeGeneral[39] controlled list values.
  This value may be used in 10.a resourceTypeGeneral[40] and other places where resourceTypeGeneral is used (12.f resourceTypeGeneral[41], 20.a relatedItemType[42]).
  Example:
  <resourceType resourceTypeGeneral="Instrument">Reflectometer</resourceType>
- Addition of new relationType[43] pair: IsCollectedBy[44] and Collects[45]

Detailed information about the change is documented in DataCite Schema 4.5 documentation[46]. Currently, DataCite is working on extending the APIs and User Interfaces to support the new schema 4.5 to enable DataCite members to create DOIs for instruments and the community to search and view metadata including connections to other PIDs such as ORCIDs (People), RORs (Organisations) and DOIs (Datasets and Publications).

[36] The metadata elements are described on the following link: https://schema.eudat.eu/eudatinstruments_metadataelements/

[37] The schema can be found on the Data Type Registry via the following link: https://dtr-profiles.pidconsortium.net/#objects/21.11145/3c003669dfb6b895ff9b.

[38] https://datacite-metadata-schema.readthedocs.io/en/4.5/appendices/appendix_1/resourceTypeGeneral.html#instrument

[39] https://datacite-metadata-schema.readthedocs.io/en/4.5/appendices/appendix_1/resourceTypeGeneral.html

[40] https://datacite-metadata-schema.readthedocs.io/en/4.5/properties/mandatory/property_resourcetype.html#a

[41] https://datacite-metadata-schema.readthedocs.io/en/4.5/properties/recommended_optional/property_relatedidentifier.html#f

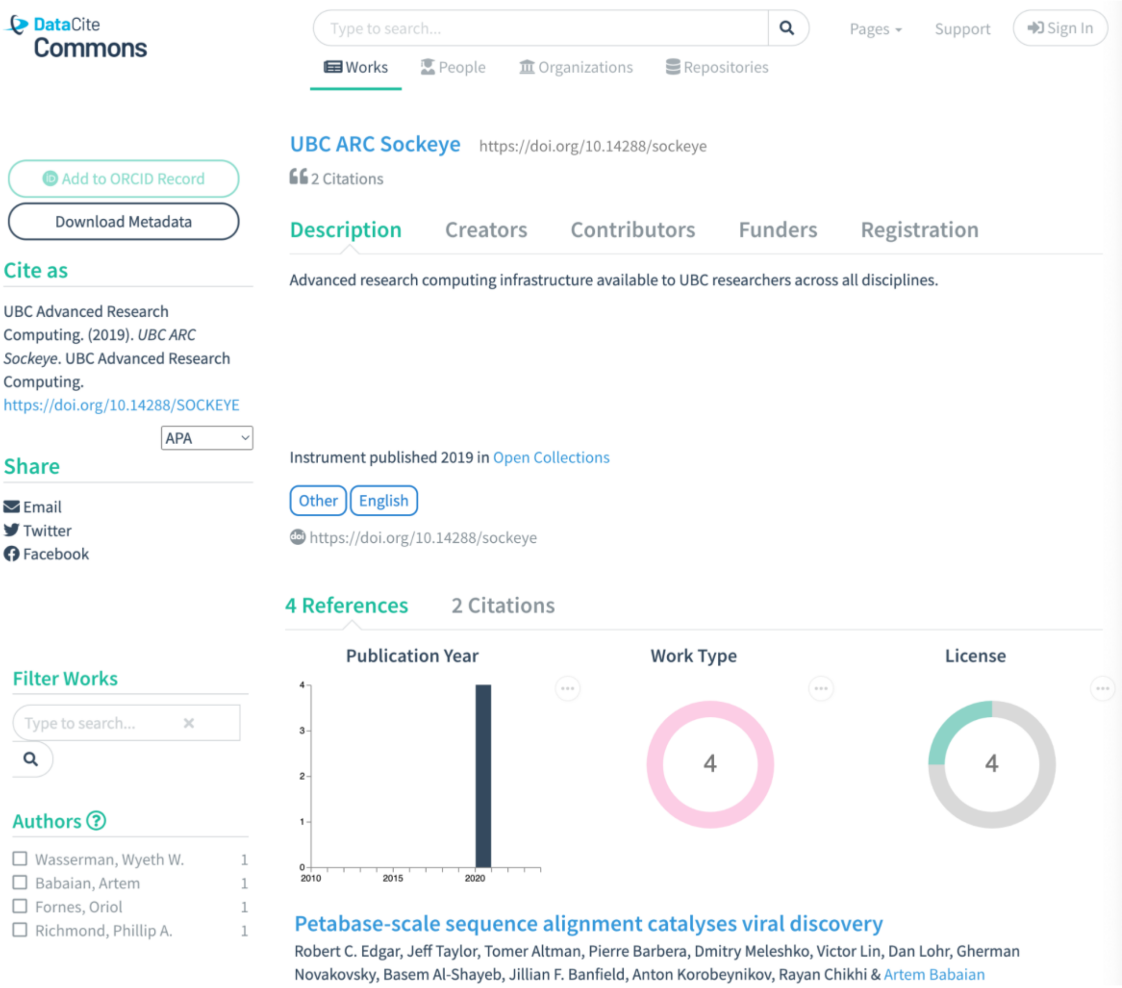
[42] https://datacite-metadata-schema.readthedocs.io/en/4.5/properties/recommended_optional/property_relateditem.html#a

[43] https://datacite-metadata-schema.readthedocs.io/en/4.5/appendices/appendix_1/relationType.html

[44] https://datacite-metadata-schema.readthedocs.io/en/4.5/appendices/appendix_1/relationType.html#iscollectedby

[45] https://datacite-metadata-schema.readthedocs.io/en/4.5/appendices/appendix_1/relationType.html#collects

[46] https://datacite-metadata-schema.readthedocs.io/en/4.5/introduction/version_update.html#support-for-instruments



UBC ARC Sockeye https://doi.org/10.14288/sockeye

2 Citations

Cite as

UBC Advanced Research Computing. (2019). *UBC ARC Sockeye*. UBC Advanced Research Computing. https://doi.org/10.14288/SOCKEYE

Advanced research computing infrastructure available to UBC researchers across all disciplines.

Instrument published 2019 in Open Collections

https://doi.org/10.14288/sockeye

Authors	
□ Wasserman, Wyeth W.	1
□ Babaian, Artem	1
□ Fornes, Oriol	1
□ Richmond, Phillip A.	1

Petabase-scale sequence alignment catalyses viral discovery

Robert C. Edgar, Jeff Taylor, Tomer Altman, Pierre Barbera, Dmitry Meleshko, Victor Lin, Dan Lohr, Gherman Novakovsky, Basem Al-Shayeb, Jillian F. Banfield, Anton Korobeynikov, Rayan Chikhi & Artem Babaian

Figure 7. Instrument view via DataCite commons[47]

3.2.1.4 Metadata Schema Enhancement in B2FIND

As a prerequisite for presenting instrument information on the B2FIND CKAN GUI, a concept was finalised to include the definition of appropriate properties within the EUDAT Core metadata schema as well as within EUDAT Extended metadata schema, in order to allow interoperability across the EUDAT services B2FIND and B2SHARE. For the web interface, metadata is displayed in a structured manner (see Table 2):

Table 2. Display of Metadata on B2FIND web GUI

Sections	Metadata Elements	Comment
Identifier	DOI / PID / Source RelatedIdentifier MetadataAccess	Internal "ranking" for several identifier: DOIs preferred, followed by other PIDs or any URL

[47] https://commons.datacite.org/doi.org/10.14288/sockeye

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

Provenance	Creator Contributor **Instrument** Publisher PublicationYear Contact Rights FundingReference OpenAccess	**with <instrumentIdentifier> and <instrumentIdentifierType>**
Representation	ResourceType Format Language Size Version Discipline SpatialCoverage TemporalCoverage	

Based on this schema implementation, the integration of instrument information from two dedicated communities was tested to showcase how this information from divergent research areas (Astronomy and High Energy Physics) can be made available within a generic search portal. Finally, the enhanced ingestion software has been deployed on the productive B2FIND. In the following, we describe these two use-cases.

3.2.1.5 Use-Cases to implement PID information for Instruments

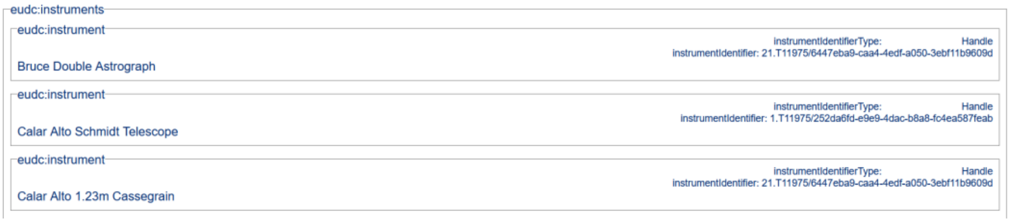
Two communities have been chosen to highlight the implementation of instrument information in B2FIND: Helmholtz-Zentrum Berlin für Materialien und Energie (HZB) and the International Virtual Observatory Alliance (IVOA). Both communities rely on the usage of scientific instruments and are interested in both persistent identification and in enabling discoverability of them.

1. **IVOA - persistent identification of instruments using B2INST**
   The Virtual Observatory (VO) is a network of astronomical data centres offering access to hundreds of millions of datasets and hundreds of billions of object records, held together by a common set of standards and a metadata Registry. The International Virtual Observatory Alliance (IVOA) is an organisation that debates and agrees on the technical standards that are needed to make the VO possible. B2FIND is harvesting metadata from the metadata Registry which is using OAI-PMH as transfer protocol and EUDAT Core as metadata schema. The reason for that is that instruments are already used within the VO and thus, should be visible in B2FIND. But as not all VO records do have a DOI, using Datacite as an exchange schema was not an option (because Datacite allows only 1 value for <IdentifierType> = 'DOI'). Thus, the VO metadata Registry (hosted by the German Astrophysical Virtual Observatory) implemented the exposure of their metadata with EUDAT Core in order to allow a faceted search for instruments[48].

[48] This may be seen here: https://b2find.eudat.eu/organization/ivoa shows all IVOA records, clicking on the facet Instrument allows users to search within this facet for a specific instrument.

Workflow and Example

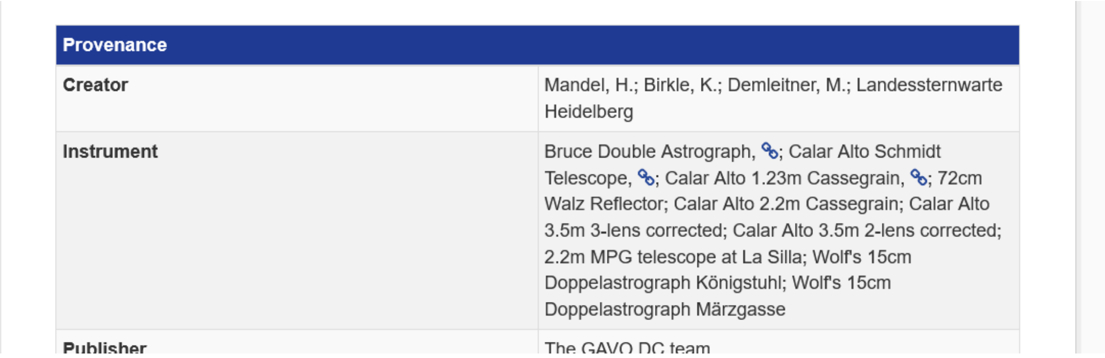
Even though there is a continuous discussion within the IVOA to include persistent identifiers for instruments, up to now there is no common decision on how to use them. However, as EUDAT is offering a new service specifically for the registration of PIDs for Instruments, B2INST was used as a test case for adding instrumentIdentifier within VO records. Therefore, the instrument "Bruce Double Astrograph" was registered in B2INST with 'name' and 'owner', thus receiving a PID for it. This PID was then included in the exposure of the VO metadata Registry with EUDAT Core, including instrumentIdentifier and instrumentIdentifierType, as shown on Figure 8.



```
eudc:instruments
  eudc:instrument
                                                        instrumentIdentifierType:                Handle
                     instrumentIdentifier: 21.T11975/6447eba9-caa4-4edf-a050-3ebf11b9609d
  Bruce Double Astrograph
  eudc:instrument
                                                        instrumentIdentifierType:                Handle
                      instrumentIdentifier: 1.T11975/252da6fd-e9e9-4dac-b8a8-fc4ea587feab
  Calar Alto Schmidt Telescope
  eudc:instrument
                                                        instrumentIdentifierType:                Handle
                     instrumentIdentifier: 21.T11975/6447eba9-caa4-4edf-a050-3ebf11b9609d
  Calar Alto 1.23m Cassegrain
```

Figure 8. IVOA OAI-PMH output using EUDAT Core

This information is harvested by B2FIND, mapped accordingly and uploaded to the search portal where the record is shown like it is depicted at Figure 9:

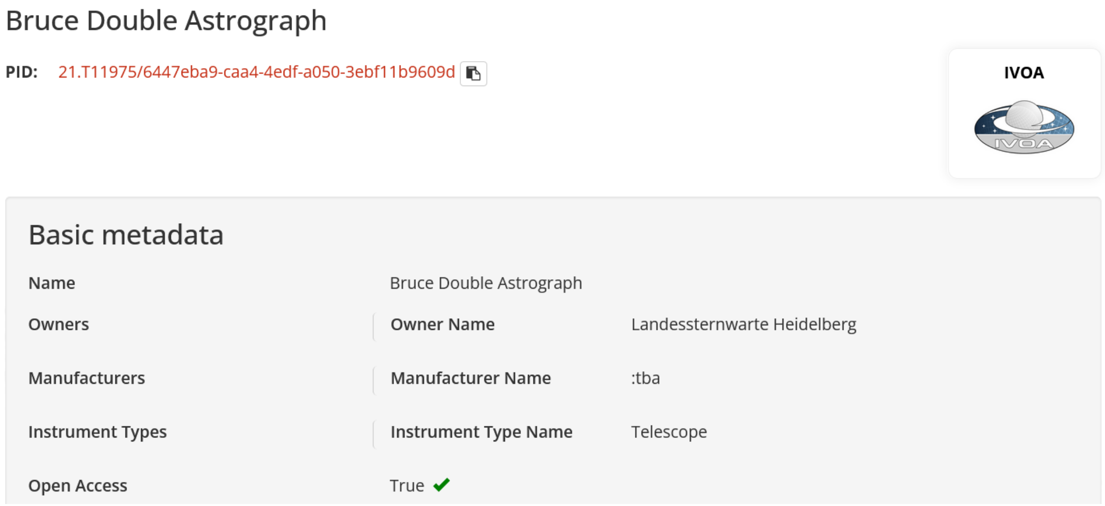


Provenance	
Creator	Mandel, H.; Birkle, K.; Demleitner, M.; Landessternwarte Heidelberg
Instrument	Bruce Double Astrograph, 🔗; Calar Alto Schmidt Telescope, 🔗; Calar Alto 1.23m Cassegrain, 🔗; 72cm Walz Reflector; Calar Alto 2.2m Cassegrain; Calar Alto 3.5m 3-lens corrected; Calar Alto 3.5m 2-lens corrected; 2.2m MPG telescope at La Silla; Wolf's 15cm Doppelastrograph Königstuhl; Wolf's 15cm Doppelastrograph Märzgasse
Publisher	The GAVO DC team

Figure 9. Single record page in B2FIND for IVOA

The icons behind the instrument names are clickable and link back to the landing page of B2INST.

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



Bruce Double Astrograph

PID: 21.T11975/6447eba9-caa4-4edf-a050-3ebf11b9609d

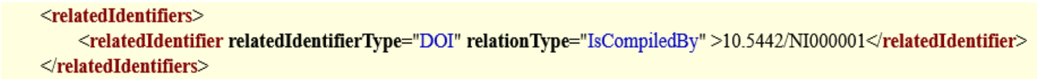
Basic metadata

Name	Bruce Double Astrograph	
Owners	Owner Name	Landessternwarte Heidelberg
Manufacturers	Manufacturer Name	:tba
Instrument Types	Instrument Type Name	Telescope
Open Access	True ✓	

Figure 10. Single record page in B2INST for 'Bruce Double Astrograph'

2. HZB - persistent identification of instruments using DOIs

The German Helmholtz-Zentrum Berlin für Materialien und Energie focuses on research for Energy and matter, in particular by a third-generation synchrotron radiation source. HZB uses OAI-PMH as a metadata transfer protocol and offers two subsets for harvesting, 'hzb_pub' and 'hzb_inv', both exposing metadata with Datacite Metadata Schema. While the first endpoint offers data publications, the latter refers to so-called 'Investigations', which are the proposals for specific experiments on the synchrotron. These investigations include information about the used instruments while the value for <creator> is ':unav'. HZB registers DOIs for persistent identification of their instruments, using the Datacite metadata element <relatedIdentifier> with the attribute relatedIdentifierType="DOI" and relationType="IsCompiledBy".



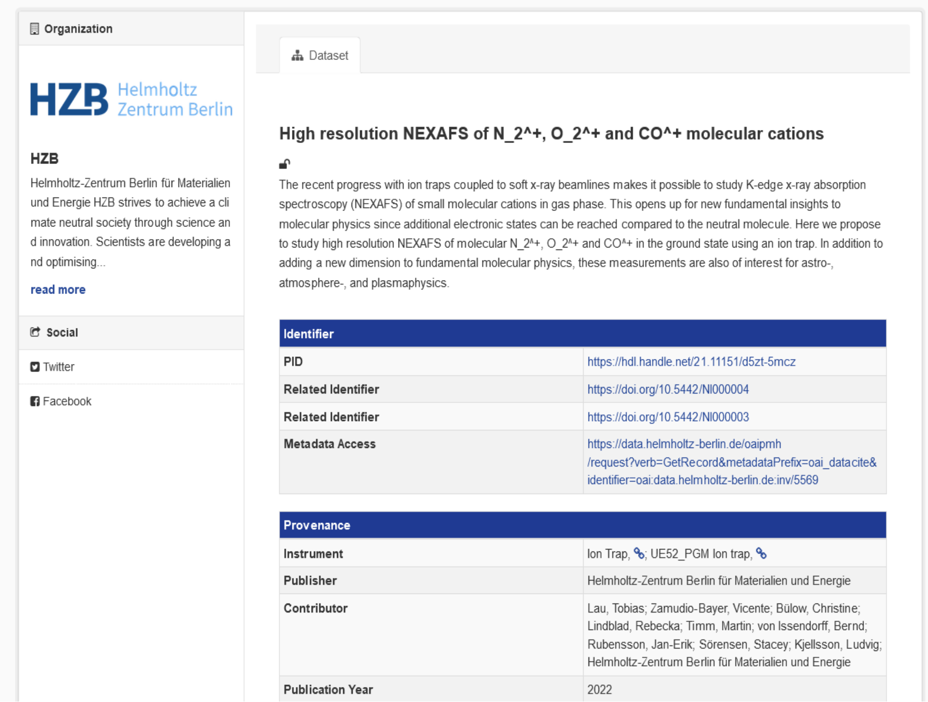
```
<relatedIdentifiers>
    <relatedIdentifier relatedIdentifierType="DOI" relationType="IsCompiledBy" >10.5442/NI000001</relatedIdentifier>
</relatedIdentifiers>
```

Figure 11. HZB OAI-PMH output for 'investigations' using Datacite schema

Workflow and Example

For the integration of HZB as a data provider in B2FIND, we used our generic datacite reader[49] that is sufficient for the subset 'hzb_pub'. Even though there is not yet a dedicated metadata element "Instrument" in Datacite metadata schema, B2FIND ingestion software is flexible enough to include additional methods for e.g. each harvesting endpoint. For retrieving instrument information from the subset 'hzb_inv', we implemented some update methods within the mapfile including formatting methods to represent the instrument DOIs as a clickable icon that is displayed behind the instrument name and refers to the landing page of HZB.

[49] The whole B2FIND ingestion software stack is published as open source code in Github: https://github.com/EUDAT-B2FIND. 'Reader' are the basic mapping scripts that support different metadata schemas, including generic schemas such as Datacite and DublinCore or thematic ones such as e.g. DDI2.5 for Social Sciences and ISO 19115/19139 for georeferenced data. An overview may be seen here: https://github.com/EUDAT-B2FIND/md-ingestion/tree/master/mdingestion/reader.



Organization

HZB

Helmholtz-Zentrum Berlin für Materialien und Energie HZB strives to achieve a cli mate neutral society through science an d innovation. Scientists are developing a nd optimising...

read more

Social

Twitter

Facebook

Dataset

High resolution NEXAFS of N_2^+, O_2^+ and CO^+ molecular cations

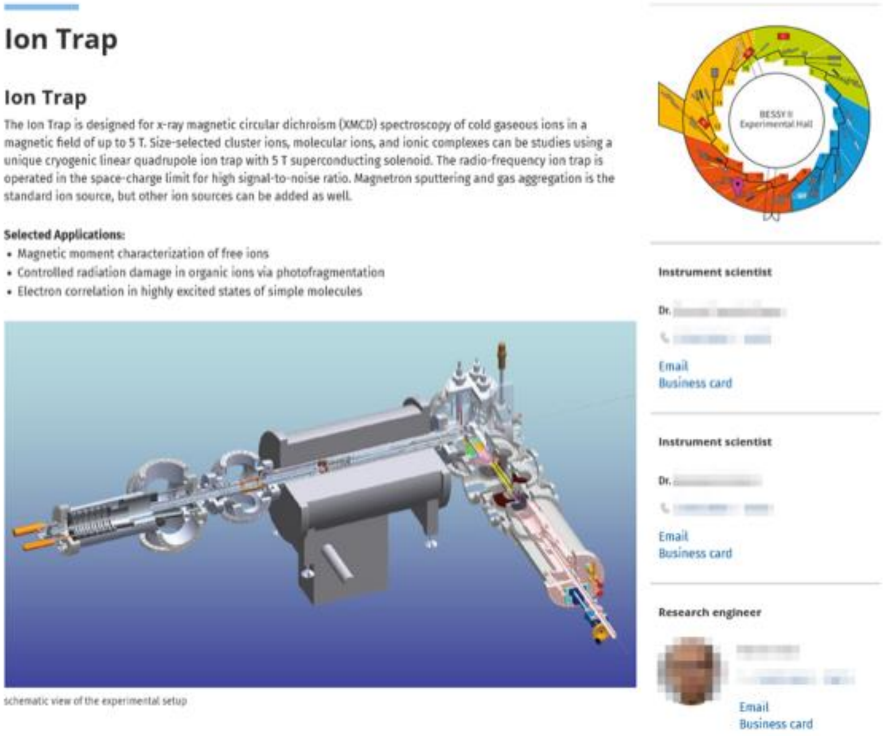
The recent progress with ion traps coupled to soft x-ray beamlines makes it possible to study K-edge x-ray absorption spectroscopy (NEXAFS) of small molecular cations in gas phase. This opens up for new fundamental insights to molecular physics since additional electronic states can be reached compared to the neutral molecule. Here we propose to study high resolution NEXAFS of molecular N_2^+, O_2^+ and CO^+ in the ground state using an ion trap. In addition to adding a new dimension to fundamental molecular physics, these measurements are also of interest for astro-, atmosphere-, and plasmaphysics.

Identifier	
PID	https://hdl.handle.net/21.11151/d5zt-5mcz
Related Identifier	https://doi.org/10.5442/NI000004
Related Identifier	https://doi.org/10.5442/NI000003
Metadata Access	https://data.helmholtz-berlin.de/oaipmh/request?verb=GetRecord&metadataPrefix=oai_datacite&identifier=oai:data.helmholtz-berlin.de:inv/5569

Provenance	
Instrument	Ion Trap, 🔗; UE52_PGM Ion trap, 🔗
Publisher	Helmholtz-Zentrum Berlin für Materialien und Energie
Contributor	Lau, Tobias; Zamudio-Bayer, Vicente; Bülow, Christine; Lindblad, Rebecka; Timm, Martin; von Issendorff, Bernd; Rubensson, Jan-Erik; Sörensen, Stacey; Kjellsson, Ludvig; Helmholtz-Zentrum Berlin für Materialien und Energie
Publication Year	2022

Figure 12. Single result page for HZB record in B2FIND

Figure 12 shows an excerpt of a HZB record in B2FIND. While the section Identifier displays all <identifier> for this record, the section Provenance includes information about the used instruments for the experiment, here the Ion Trap and the specific beamline with variable polarisation, UE52_PGM. Both identifiers lead to a HZB landing page with further information.



Ion Trap

Ion Trap

The Ion Trap is designed for x-ray magnetic circular dichroism (XMCD) spectroscopy of cold gaseous ions in a magnetic field of up to 5 T. Size-selected cluster ions, molecular ions, and ionic complexes can be studies using a unique cryogenic linear quadrupole ion trap with 5 T superconducting solenoid. The radio-frequency ion trap is operated in the space-charge limit for high signal-to-noise ratio. Magnetron sputtering and gas aggregation is the standard ion source, but other ion sources can be added as well.

Selected Applications:

- Magnetic moment characterization of free ions
- Controlled radiation damage in organic ions via photofragmentation
- Electron correlation in highly excited states of simple molecules

schematic view of the experimental setup

Instrument scientist

Dr. [redacted]

[redacted]

Email
Business card

Instrument scientist

Dr. [redacted]

[redacted]

Email
Business card

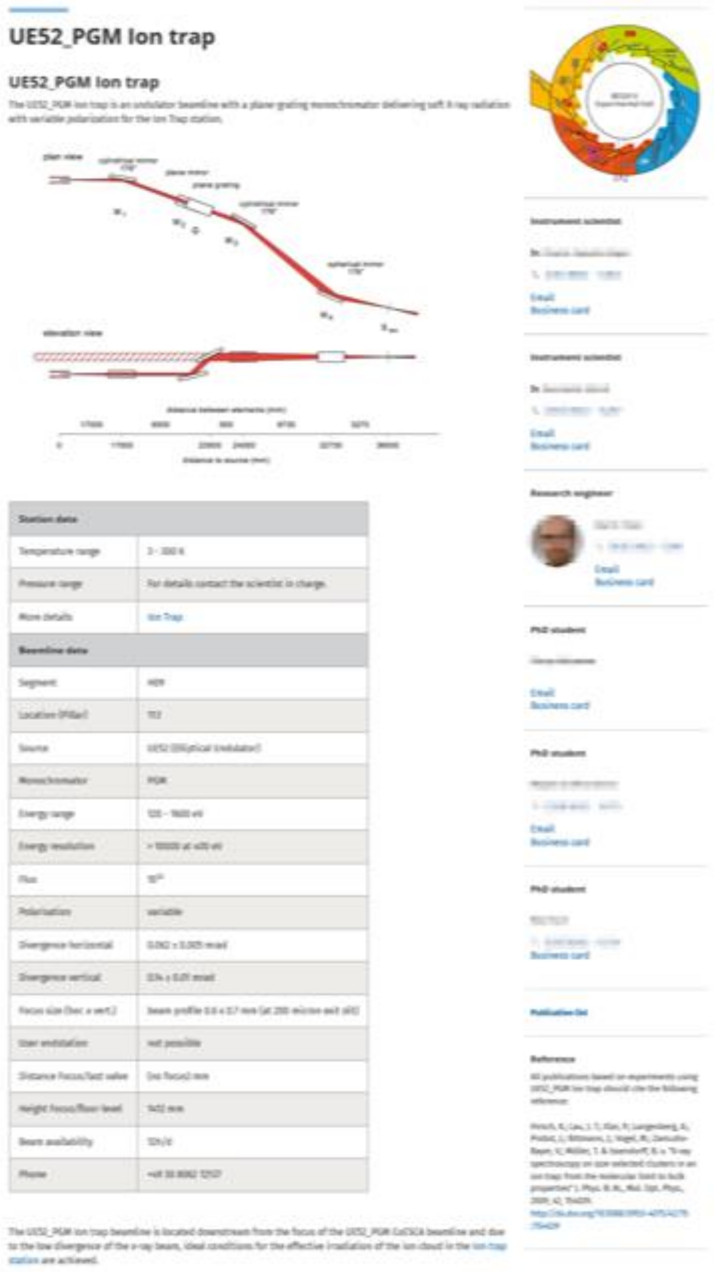
Research engineer

[redacted]

[redacted]

Email
Business card

Figure 13. Landing page for the link to Ion Trap at HZB



UE52_PGM Ion trap

UE52_PGM Ion trap

The UE52_PGM Ion trap is an undulator beamline with a plane grating monochromator delivering soft X-ray radiation with variable polarization for the Ion Trap station.

Station data	
Temperature range	3 - 300 K
Pressure range	For details contact the scientist in charge.
More details	Ion Trap
Beamline data	
Segment	H09
Location (Pillar)	11.2
Source	UE52 (Elliptical Undulator)
Monochromator	PGM
Energy range	120 - 1600 eV
Energy resolution	> 10000 at 400 eV
Flux	10^{12}
Polarisation	variable
Divergence horizontal	0.062 ± 0.005 mrad
Divergence vertical	0.14 ± 0.01 mrad
Focus size (hor. x vert.)	beam profile 0.6 x 0.7 mm (at 200 micron exit slit)
User endstation	not possible
Distance Focus/last valve	(no focus) mm
Height Focus/floor level	1402 mm
Beam availability	12h/d
Phone	+49 30 8062 12137

The UE52_PGM Ion trap beamline is located downstream from the focus of the UE52_PGM CoESCA beamline and due to the low divergence of the x-ray beam, ideal conditions for the effective irradiation of the ion cloud in the ion trap station are achieved.

Reference

All publications based on experiments using UE52_PGM Ion trap should cite the following reference:

Hirsch, K.; Lau, J. T.; Klar, Ph.; Langenberg, A.; Probst, J.; Rittmann, J.; Vogel, M.; Zamudio-Bayer, V.; Möller, T. & Issendorff, B. v. "X-ray spectroscopy on size-selected clusters in an ion trap: from the molecular limit to bulk properties" J. Phys. B At. Mol. Opt. Phys. 2009, 42, 154029. http://dx.doi.org/10.1088/0953-4075/42/15/154029

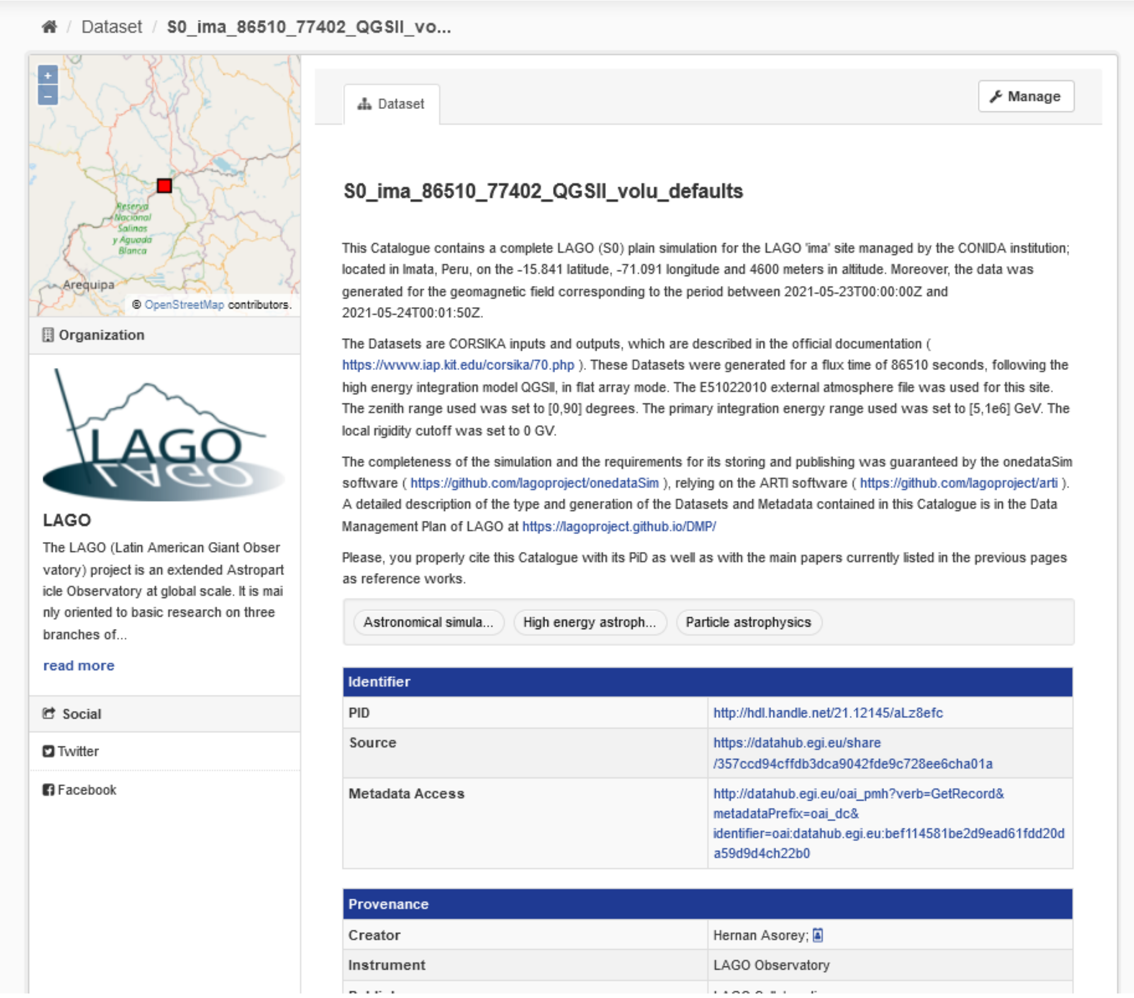
Figure 14. Landing page for the link to UE52_PGM Ion Trap at HZB

3.2.2 People

Another type of persistent identifiers refers to the unambiguous identification of people referenced within a dataset metadata record. Here, the Open Researcher and Contributor ID (ORCID) enables persistent identification of e.g. creators of research data and allows to find other data or publications created by a specific person. The Datacite metadata schema already allows to transfer this information using the element Creator with <nameIdentifier> and <nameIdentifierType = 'ORCID'>. As several data providers from which B2FIND is harvesting metadata already include this information, within B2FIND ingestion software the already mentioned 'reader' for datacite was enhanced in order to allow a mapping of ORCIDs as additional identifiers for <Creator>. ORCIDs may also be transferred with DublinCore, however

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

this is not clearly defined within the DublinCore metadata schema; several elements may be used to include identifiers for people, being contradictory to each other. Thus, it was not an option to change the B2FIND 'reader' for DublinCore as this one defines a generic mapping. Nonetheless, if ORCIDs are offered by a repository using DublinCore as a metadata exchange schema, it is possible to retrieve this information by using an update method within the repository specific mapping. This happened with records from the Latin American Giant Observatory (Figure 15).


Dataset / S0_ima_86510_77402_QGSII_vo...
Reserva Nacional Salinas y Aguada Blanca
Arequipa
© OpenStreetMap contributors.
Organization
Dataset
Manage
Social
Twitter
Facebook

S0_ima_86510_77402_QGSII_volu_defaults

This Catalogue contains a complete LAGO (S0) plain simulation for the LAGO 'ima' site managed by the CONIDA institution; located in Imata, Peru, on the -15.841 latitude, -71.091 longitude and 4600 meters in altitude. Moreover, the data was generated for the geomagnetic field corresponding to the period between 2021-05-23T00:00:00Z and 2021-05-24T00:01:50Z.

The Datasets are CORSIKA inputs and outputs, which are described in the official documentation (https://www.iap.kit.edu/corsika/70.php). These Datasets were generated for a flux time of 86510 seconds, following the high energy integration model QGSII, in flat array mode. The E51022010 external atmosphere file was used for this site. The zenith range used was set to [0,90] degrees. The primary integration energy range used was set to [5,1e6] GeV. The local rigidity cutoff was set to 0 GV.

The completeness of the simulation and the requirements for its storing and publishing was guaranteed by the onedataSim software (https://github.com/lagoproject/onedataSim), relying on the ARTI software (https://github.com/lagoproject/arti). A detailed description of the type and generation of the Datasets and Metadata contained in this Catalogue is in the Data Management Plan of LAGO at https://lagoproject.github.io/DMP/

Please, you properly cite this Catalogue with its PiD as well as with the main papers currently listed in the previous pages as reference works.

Astronomical simula... High energy astroph... Particle astrophysics

Identifier	
PID	http://hdl.handle.net/21.12145/aLz8efc
Source	https://datahub.egi.eu/share /357ccd94cffdb3dca9042fde9c728ee6cha01a
Metadata Access	http://datahub.egi.eu/oai_pmh?verb=GetRecord& metadataPrefix=oai_dc& identifier=oai:datahub.egi.eu:bef114581be2d9ead61fdd20da59d9d4ch22b0

Provenance	
Creator	Hernan Asorey; 🆔
Instrument	LAGO Observatory

LAGO

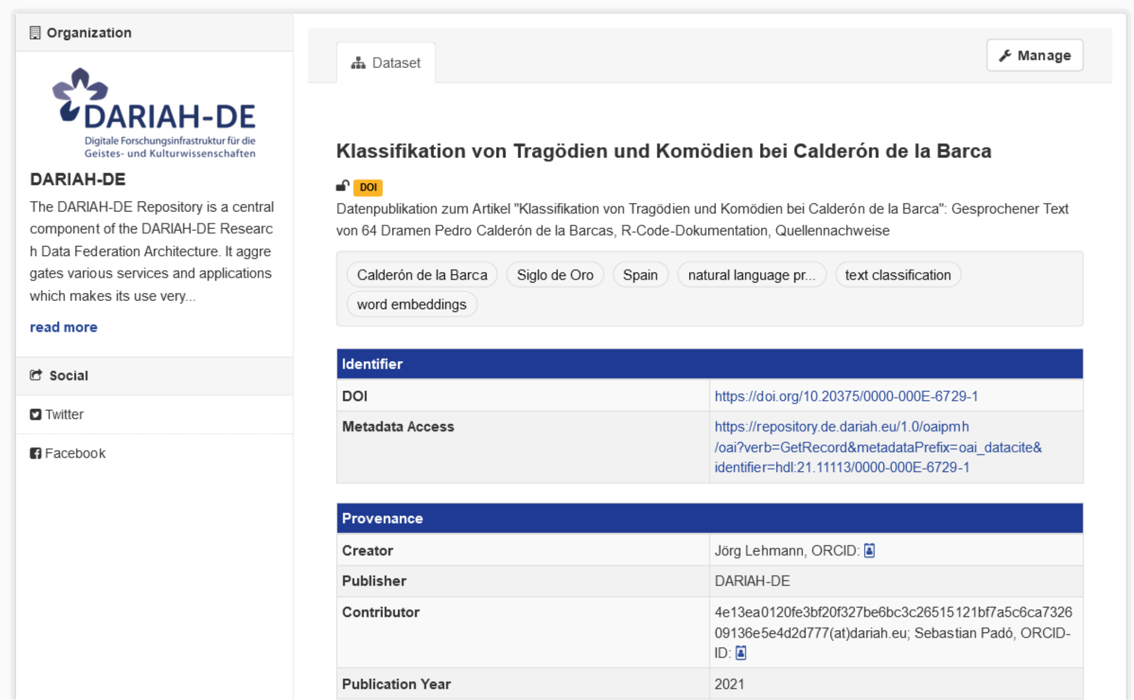
The LAGO (Latin American Giant Obser vatory) project is an extended Astropart icle Observatory at global scale. It is mai nly oriented to basic research on three branches of...

read more

Figure 15. Single records page in B2FIND for LAGO record

If B2FIND retrieves ORCIDs within the harvesting, they are mapped accordingly to EUDAT Core and displayed on the web GUI within the metadata elements <creator> or <contributor>. Currently, ORCID identifiers are shown after the value of <creator> or <contributor>. This value usually is a name after which a string called 'ORCID:' is shown, followed by a clickable icon. If multiple creators and ORCIDs exist, they are shown as a semicolon separated list. Clicking on the icon behind a certain creator or contributor leads to the corresponding landing page of ORCID. The following images show the integration of ORCIDs from different data repositories on the single result page in B2FIND, using the enhanced Datacite reader. Figure 16 shows an example of a record in 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.



Organization
DARIAH-DE
Digitale Forschungsinfrastruktur für die Geistes- und Kulturwissenschaften

DARIAH-DE

The DARIAH-DE Repository is a central component of the DARIAH-DE Researc h Data Federation Architecture. It aggre gates various services and applications which makes its use very...

read more
Social
Twitter
Facebook
Dataset
Manage

Klassifikation von Tragödien und Komödien bei Calderón de la Barca

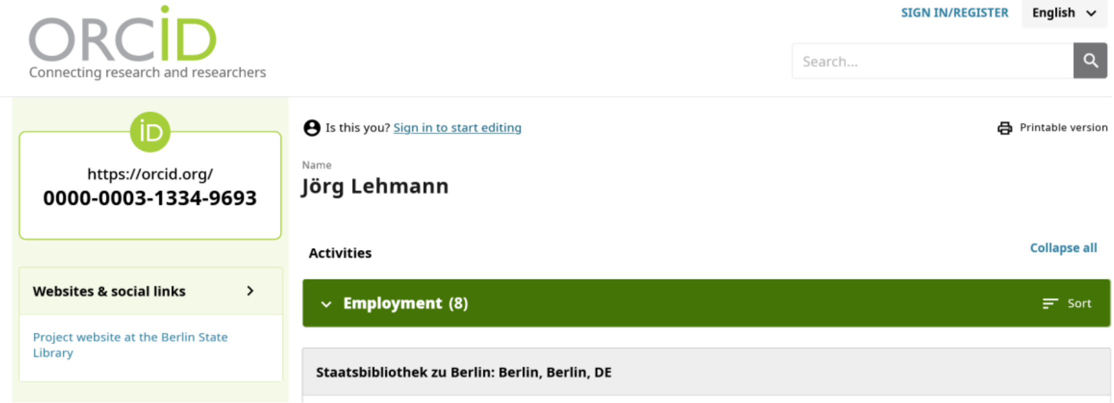
Datenpublikation zum Artikel "Klassifikation von Tragödien und Komödien bei Calderón de la Barca": Gesprochener Text von 64 Dramen Pedro Calderón de la Barcas, R-Code-Dokumentation, Quellennachweise

Identifier	
DOI	https://doi.org/10.20375/0000-000E-6729-1
Metadata Access	https://repository.de.dariah.eu/1.0/oaipmh/oai?verb=GetRecord&metadataPrefix=oai_datacite&identifier=hdl:21.11113/0000-000E-6729-1

Provenance	
Creator	Jörg Lehmann, ORCID:
Publisher	DARIAH-DE
Contributor	4e13ea0120fe3bf20f327be6bc3c26515121bf7a5c6ca732609136e5e4d2d777(at)dariah.eu; Sebastian Padó, ORCID-ID:
Publication Year	2021

Figure 16. Single records page in B2FIND for DARIAH-DE record

Clicking on the icon behind the creator leads to the corresponding ORCID page of this specific person (Figure 17):



https://orcid.org/
0000-0003-1334-9693

Websites & social links
Project website at the Berlin State Library
Is this you? Sign in to start editing
Printable version

Name
Jörg Lehmann

Staatsbibliothek zu Berlin: Berlin, Berlin, DE

Figure 17. Public ORCID landing page for https://orcid.org/0000-0003-1334-9693

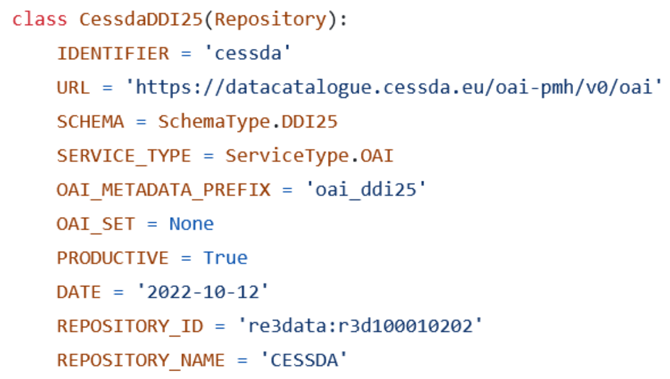
3.2.3 Repositories

As B2FIND is not merely an interdisciplinary discovery portal for research data but as well a metadata aggregator, its role as metadata curator deepens. In particular, for offering metadata to OpenAIRE explore we decided to expand our OAI-PMH extension for CKAN in order to expose additional persistent identifiers for harvested repositories to be able to track the provenance of the metadata harvested by B2FIND. According to the OpenAIRE Guidelines[50] these identifiers must come either from re3data, FAIRSharing or OpenDOAR. In order to increase the visibility of

[50] https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/latest/use_of_oai_pmh.html

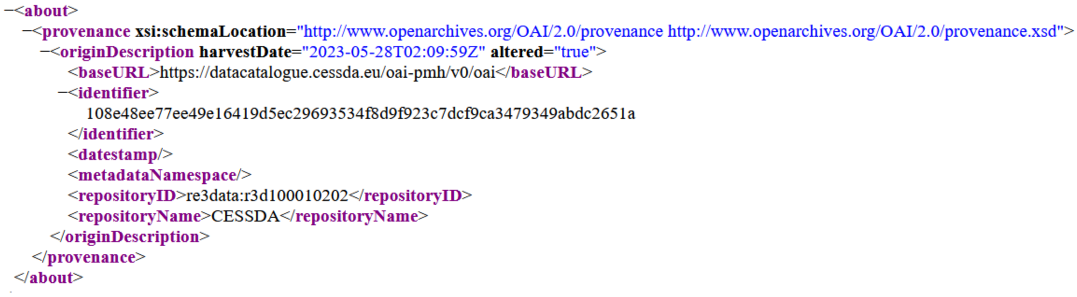
The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

metadata provenance, a new section <about> was included in the B2FIND OAI-PMH[51] output with two dedicated elements, namely 'repositoryID' and 'repositoryName'. It is important here that the used name and ID should correspond to the distinct harvesting endpoint B2FIND is using. To ensure this information is valid, a mapping was included for all repositories in B2FIND using a manually collected list. This list was reviewed, revised and then implemented within the repository mapfiles of B2FIND ingestion software. Thus, if a specific repository in B2FIND is registered in re3data[52], Fairsharing[53] or OpenDOAR[54], the name and ID are attached to the <about> section of each record in the OAI-PMH output harvested from the specified endpoint. (Figure 18) shows as an example how Repository ID and Name are included in the mapfile and exposed within the B2FIND OAI-PMH output.



```
class CessdaDDI25(Repository):
    IDENTIFIER = 'cessda'
    URL = 'https://datacatalogue.cessda.eu/oai-pmh/v0/oai'
    SCHEMA = SchemaType.DDI25
    SERVICE_TYPE = ServiceType.OAI
    OAI_METADATA_PREFIX = 'oai_ddi25'
    OAI_SET = None
    PRODUCTIVE = True
    DATE = '2022-10-12'
    REPOSITORY_ID = 're3data:r3d100010202'
    REPOSITORY_NAME = 'CESSDA'
```

Figure 18. Excerpt of B2FIND mapfile for CESSDA with re3data ID and Name



```
−<about>
  −<provenance xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/provenance http://www.openarchives.org/OAI/2.0/provenance.xsd">
    −<originDescription harvestDate="2023-05-28T02:09:59Z" altered="true">
       <baseURL>https://datacatalogue.cessda.eu/oai-pmh/v0/oai</baseURL>
      −<identifier>
         108e48ee77ee49e16419d5ec29693534f8d9f923c7dcf9ca3479349abdc2651a
       </identifier>
       <datestamp/>
       <metadataNamespace/>
       <repositoryID>re3data:r3d100010202</repositoryID>
       <repositoryName>CESSDA</repositoryName>
     </originDescription>
   </provenance>
 </about>
```

Figure 19. B2FIND OAI-PMH output of a CESSDA record with re3data ID and Name

Of course, this mapping applies only to those repositories that are registered. If no identifier exists, there will be no 'repositoryID' and 'repositoryName' within the OAI-PMH output. An example for this case is BBMRI, which is integrated into B2FIND via the central B2SHARE instance hosted at the Finnish national IT Center for Science (CSC): while BBMRI-ERIC is an European infrastructure for biomedical research (connecting biobanks) with its own directory to search and access data[55] there is also a BBMRI 'Community' in B2SHARE where e.g. supplementary data for papers is stored. While the BBMRI-ERIC directory is registered in re3data[56] the

[51] more precisely: in the OAI-PMH extension for CKAN that has been developed at the Finnish IT Center for Science (CSC) and modified by B2FIND staff, see the code in Github: https://github.com/EUDAT-B2FIND/ckanext-oaipmh-server.

[52] https://www.re3data.org/

[53] https://fairsharing.org/

[54] https://v2.sherpa.ac.uk/opendoar/

[55] https://directory.bbmri-eric.eu/#/

[56] https://www.re3data.org/repository/r3d100013028

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

harvesting endpoint for B2FIND is a specific OAI Set from B2SHARE that does (not yet) have a registered identifier. Consequently no ‘repositoryID’ and ‘repositoryName’ is offered within the B2FIND OAI-PMH output. Another example is IVOA, the International Virtual Observatory Alliance. Here the point is that the VO is a federated system, where multiple databases are queried (using pretty advanced tools) to create metadata records that are then exposed for harvesting via a central registry. Thus, there is not one single database or even repository, but many. Thus, it is not possible to attach one specific identifier.[57]

3.3 Interoperability Issues

The term interoperability here refers to (a) the maintenance of services as a basic prerequisite for any interoperability issues, (b) the development of new features to enhance the ways services or tools may interact with each other and (c) the deployment of these developments within already existing services or operations. This section describes the work done on the (1) provider side for the DataCite PID Graph, (2) the metadata exposure from B2FIND to OpenAIRE and their metadata ingestion respectively and (3) the OAI-PMH implementation in B2INST.

3.3.1 DataCite PID Graph

Currently, DataCite metadata and the PID graph can be accessed in multiple ways.

- REST API[58] - allows any user to retrieve, query and browse DataCite DOI metadata records
- OAI-PMH[59] - exposes metadata stored in the DataCite Metadata Store (MDS) using the Open Archives Initiative Protocol for Metadata Harvesting
- GraphQL API[60] - support queries of the DataCite API using the GraphQL query language[61]

Even though these APIs provide easy access to the DataCite metadata and the PID graph for individual records and subsets, accessing all content remains challenging. As part of this project, we have started to work on developing a service to download the metadata and PID graph as a data dump.

As part of DICE we worked on improving the underlying infrastructure that is needed to support the data dump service. DataCite worked on updating our production ElasticSearch cluster to version 7.10, moving to a new instance type (Graviton) which is cheaper and faster and using reserved instances to reduce cost.

As part of FAIRCORE4EOSC, we will continue this work to develop a harvester service to enable the community to download DataCite metadata and the PID graph. DataCite plans to release a beta version at the end of this year and the progress of this work can be followed via FAIRCORE4EOSC T3.5 project dashboard[62].

[57]Obviously it would be an option for the central metadata registry of the VO (hosted by GAVO) to register itself as a “Repository” in re3data. Only if this is desirable on a political level is highly questionable. Several data centres that are part of IVOA are already registered in re3data, e.g. the GAVO Data Centre or ESA Planetary Science Archive.

[58] https://support.datacite.org/reference

[59] https://oai.datacite.org/oai

[60] https://api.datacite.org/graphql

[61] https://graphql.org/

[62] https://github.com/orgs/datacite/projects/14

3.3.2 B2FIND and OpenAIRE

As B2FIND is an important (meta)data provider for OpenAIRE, to increase interoperability of the metadata transfer has been a key factor. During several harvesting iterations it became clear that nearly half of the metadata records offered by B2FIND were not correctly ingested in OpenAIRE Explore; in order to understand why and to solve that unsatisfactory status, some effort was needed. Both the structural challenges and our deployed solution are described.

The "OpenAIRE Guidelines for Data Archives"[63] state that *"If ResourceType is used, resourceTypeGeneral is mandatory"*. As a lot of B2FIND records contain a ResourceType, mapped as a simple key-value pair without further attributes like resourceTypeGeneral, the main challenge has been (1) to keep this information (if originally provided) throughout the mapping process to B2FIND's Eudat Core metadata schema and (2) to deliver this information to OpenAIRE. This is needed to group the records on OpenAIRE side into either "Publications", "Research Data", "Research software" or "Other research products” and make the metadata findable within these categories in OpenAIRE Explore. To export the resourceTypeGeneral, in a first step B2FIND had to keep this information, if delivered from the harvested repositories. To realise this within our non-hierarchical, key-value pair based Eudat Core metadata schema, we took the information from e.g. the attribute "resourceTypeGeneral'' from records B2FIND harvests in Datacite Schema. Then we added it as an additional value in the field "Resource Type" in B2FIND. To export this information via our OAI-PMH CKAN extension, we wrote a Python script that parses the <resourceType> values and maps the found values (e.g. ‘dataset’) to the respective controlled list values of the OpenAIRE schema applicable for B2FIND metadata. If a match is found, it is written into and exported by the "resourceTypeGeneral'' attribute, which again enables OpenAIRE to correctly map the harvested metadata records to their Research Type categories. This workaround was successfully implemented insofar as during the last harvesting iteration (May 2023) by OpenAIRE the amount of correctly classified records increased significantly:

Table 3. Comparison of B2FIND output classification in OpenAIRE over time

Date	Research Publications	Research data	Research software	Other research output
11/2022	95	59	4	1.035.267
05/2023	349	841.361	341	311.027

Another interoperability issue refers to DataCite <identifier>: within the DataCite metadata schema there is only one <identifierType> =’DOI’. That is reasonable as DataCite metadata schema was developed for registering DOIs; however, DOI is only one option for persistent identification. As B2FIND is harvesting many repositories that use other identifiers (like e.g. Handles), the option has been (1) to ‘misuse’ the DataCite metadata schema for B2FIND output and use <identifier> for representing the harvested identifiers even though they are not DOIs or (2) to put all identifiers that are not DOIs within the metadata element <alternateIdentifier>. We chose the latter option when developing our OAI-PMH extension for CKAN some years ago, which unfortunately collides with OpenAIRE, because the <identifier> element is mandatory there (even though several <identifierTypes> are supported). The workaround here is a script on the OpenAIRE side that parses all harvested metadata and allows mapping values from <alternateIdentifier> to <identifier> if there is no value for that in the first step.

[63] https://guidelines.openaire.eu/en/latest/data/field_resourcetype.html

Finally, the deployment of new features within the B2FIND software stack and the maintenance of B2FIND service operations demanded many resources. Firstly, because any basic change within the specific mapfiles requires a reingest of all metadata of a specific harvesting endpoint (that has been the case e.g. for the implementation of 'repositoryID' and 'repositoryName'); even though this process may be automated to a certain extent, it still requires effort. Secondly, due to an DKRZ internal shift from CentOS 8 to AlmaLinux we needed to setup and configure a new environment for the B2FIND service, including a new productive machine (separated into SolR and CKAN Postgres), a testbed and several virtual machines, including test machines for metadata ingestion and software development. Setting up a new productive machine again required a complete reingest of all repositories, due to the amount of records in B2FIND that alone took several weeks.[64]

3.3.3 B2INST and its Harvesting API

In order to enhance interoperability between B2INST and various other data management services, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) API has been used in B2INST. The metadata related to the instruments have been exposed to other services in a structured manner so that the other services can make requests to the OAI-PMH API to harvest all the metadata related to the instruments. One of the key advantages of the OAI-PMH is its simplicity. The protocol is based on the HTTP protocol and uses a small set of verbs and nouns to define the metadata harvesting operations. This simplicity makes it easy for developers to implement and for different systems to communicate with each other.

To use the OAI-PMH, B2INST needed to first expose its metadata using the protocol. We therefore created a metadata format that is compliant with the OAI-PMH and made it available via a web service. Once the metadata is exposed, other repositories can use the protocol to request the metadata and integrate it into their own systems.

We adapted the record serializers in B2INST to support the following metadata formats via the OAI-PMH API:

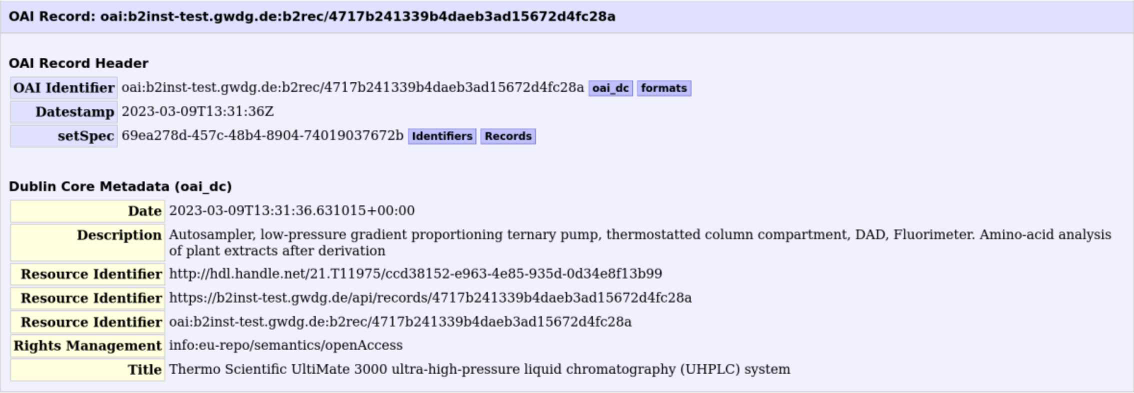
- Dublin Core (OAI-DC)
- EUDAT Core
- EUDAT Extended
- DataCite

Our record serializers were implemented in a way that they are able to expose the following metadata fields for the registered instruments (as shown on Figure 20):

- Title, Description and Date
- Resource Identifier (as Handle PID),
- Resource Identifier (as URL to the backend API),
- Resource Identifier (as OAI-specific internal format)
- Rights management

That way B2INST can support various services, which harvests metadata using different formats.

[64] The limiting factor here is the upload to Postgres due to the CKAN schema; the average upload time is 1,5 seconds / record. There is possibly room for improvement here (e.g in principle CKAN does allow bulk upload).



OAI Record: oai:b2inst-test.gwdg.de:b2rec/4717b241339b4daeb3ad15672d4fc28a

OAI Record Header

OAI Identifier	oai:b2inst-test.gwdg.de:b2rec/4717b241339b4daeb3ad15672d4fc28a
Datestamp	2023-03-09T13:31:36Z
setSpec	69ea278d-457c-48b4-8904-74019037672b

Dublin Core Metadata (oai_dc)

Date	2023-03-09T13:31:36.631015+00:00
Description	Autosampler, low-pressure gradient proportioning ternary pump, thermostatted column compartment, DAD, Fluorimeter. Amino-acid analysis of plant extracts after derivation
Resource Identifier	http://hdl.handle.net/21.T11975/ccd38152-e963-4e85-935d-0d34e8f13b99
Resource Identifier	https://b2inst-test.gwdg.de/api/records/4717b241339b4daeb3ad15672d4fc28a
Resource Identifier	oai:b2inst-test.gwdg.de:b2rec/4717b241339b4daeb3ad15672d4fc28a
Rights Management	info:eu-repo/semantics/openAccess
Title	Thermo Scientific UltiMate 3000 ultra-high-pressure liquid chromatography (UHPLC) system

Figure 20. Listing an OAI-PMH record (ListRecord) in B2INST

3.4 Conclusion

Within the framework of the DICE project, several persistent identifier types have been included in the B2FIND ingestion software, as well as into the generic discovery portal. In this section we described the integration of PIDs for (1) scientific instruments, (2) people IDs and (3) repository IDs into B2FIND. Here we summarise some of our conclusions:

- Several referencing systems (PIDs) for different purposes are in use already and there are some overlaps between these PID systems. It is however not always clear, (a) which system is "responsible" for a particular PID type and (b) what distinguishes the different purposes. Thus, the whole referencing system is still an evolving one and it heavily relies on the active usage of data providers. Good metadata management (e.g. curation) is also important.

- Interoperability becomes difficult when different levels of structured metadata are confronted: both EUDAT Core and the PIDINST schema were designed with a "low barrier approach" in mind (e.g. minimum set of metadata elements, flat structure, simple key-value pairs). This can help the ingestion of metadata even from "not advanced data providers". However, it is difficult to transfer this information into a hierarchically structured schema (like DataCite). In particular, if B2FIND harvests records in Datacite format (with well structured information), maps them onto a "flat" schema like EUDAT Core and maps it back to Datacite metadata schema for metadata exposure some information might get lost.

- interoperability is an ongoing process, it doesn't stop (as services constantly develop)

- The PID information now integrated in B2FIND helps to disambiguate metadata and to link information beyond the original search. It is thus a good first step towards connecting metadata information across repositories for federated search systems, like knowledge graphs, will hopefully be able to realise this in the future. However, the PID information will unfold its full potential only in such federated graph-based search portals, where e.g. an ORCID of a creator in a metadata record from one repository can link to other publications from the same author from a different metadata repository. This is not yet technically possible with the current B2FIND system and remains for future development work.

4 Implementation of the LTP policy (T4.3)

4.1 Introduction

Task 4.3 of the DICE project concerns the implementation of Long Term Preservation (LTP) policies for digital data. The description of work involves the formulation and implementation of LTP policies in relation to the EUDAT services.

The chapter consists of two main parts:

- Part I provides an update on the status of the LTP Template after it was published in the first joint deliverable D4.2[65] of the DICE project in April 2022 in B2SHARE.
- Part II describes the proof-of-concept of the Digital Preservation Service, based on the technical specifications, supplied in Appendix 2 to this deliverable.

4.2 Long-Term Preservation Policy Template

4.2.1 Introduction

Many research data repositories want to be clear about what they can promise their customers with respect to long-term preservation (LTP), or are officially required to be explicit about their LTP policies. Also, if a digital archive wants to be certified, an LTP policy needs to be in place. To promote and assist the formulation of such policies, the DICE project created a template and accompanying guidance that data services and repositories can use to formulate an LTP policy. The template follows the components of the well-known OAIS reference model (Open Archives Information System, ISO 14721, 2012) and hence guarantees that all basic subjects of LTP are covered. The template intends to be generic, so that it can be used by a wide range of repositories and policy-based data archives to formulate their own LTP policies. It can be used in situations where a repository itself provides an LTP service, but also when this service is outsourced to a dedicated archive. The template can be used as a “fill-in-the-blanks exercise”, or as a reference and example. Of course, data services and repositories can adapt the template to their own needs because requirements will depend on the local situation of each service and may even change over time. Moreover, to allow for different levels of aspiration, the template uses the four levels of curation defined by CoreTrustSeal (CTS) to define the requirements for each curation level. Furthermore, an introductory section on cost modelling for long-term preservation is included. It provides insight into different cost models and the variables influencing the costs.

The LTP Policy Template is designed to be connected to, or included in the “EUDAT Service Management Framework”, version 2.5 (2021-02-19), but can also be used independently.

The template document has been identified as a Key Exploitable Result (KER) by the DICE project. The template is published in the joint DICE project deliverable (D4.2) amongst other topics of the work package 4 tasks, that are also included. A separate stand-alone version of the LTP Policy Template was created afterwards, including the guidance part.

[65] https://doi.org/10.23728/b2share.b27ba476b37740d598a87092e1a527f0

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

4.2.2 Stand-alone Long-Term Preservation Policy Template

To improve the FAIRness and visibility of the document, we deposited the LTP Policy Template as a stand-alone document in Zendo[66] from where it is openly accessible: "Long-Term Preservation Policies for Research Data Repositories: a Template".

In this way, by offering it as a dedicated, single document, it should be easier to find the policy template by services that would like to use the LTP Policy Template as a basis for the formulation of their own LTP policy.

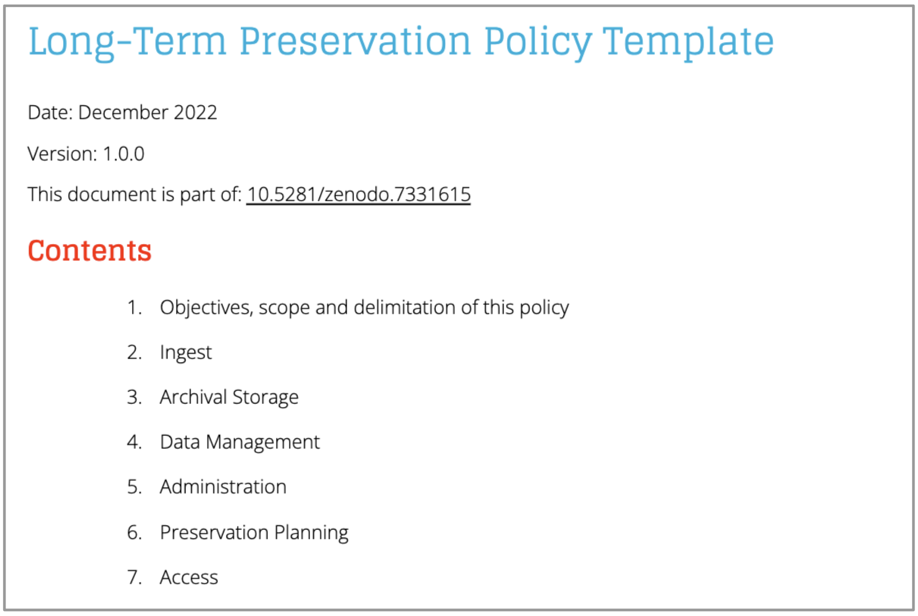
The actual stand-alone version of the LTP Policy Template has a few minor updates in comparison to the version that was part of the previous deliverable D4.2.

To increase the scope and outreach of the template, it has also been submitted to the EOSC Long-Term Data Preservation Task Force[67], as will be discussed in the next section.

The LTP Policy publication (available in Zenodo) consists of three files:

1. Background information: "Long-Term Preservation Policies for Research Data Repositories a Template.pdf"
2. LTP Policy Template form in PDF format: "Template for LTP Policies.pdf"
3. LTP Policy Template in Word format: "Template Template for LTP Policies.docx"

The first file is a single document that contains the complete report, including the LTP Policy Template, as it can be found in the joint deliverable D4.2. The other two files contain the LTP Policy Template itself, without any guidance or explanation, in two different file formats. The docx format document is included here for easy editing and adjusting the policy template. Figure 21 shows the Table Of Contents (TOC) of the LTP Policy Template.



Long-Term Preservation Policy Template

Date: December 2022

Version: 1.0.0

This document is part of: 10.5281/zenodo.7331615

Contents

Figure 21. Table of contents of the Long-Term Preservation Policy Template

[66] https://doi.org/10.5281/zenodo.7331614 (visited: 11th May 2023)
[67] https://www.eosc.eu/advisory-groups/long-term-data-preservation (visited: 11th May 2023)

The report (file #1) consists of four parts:

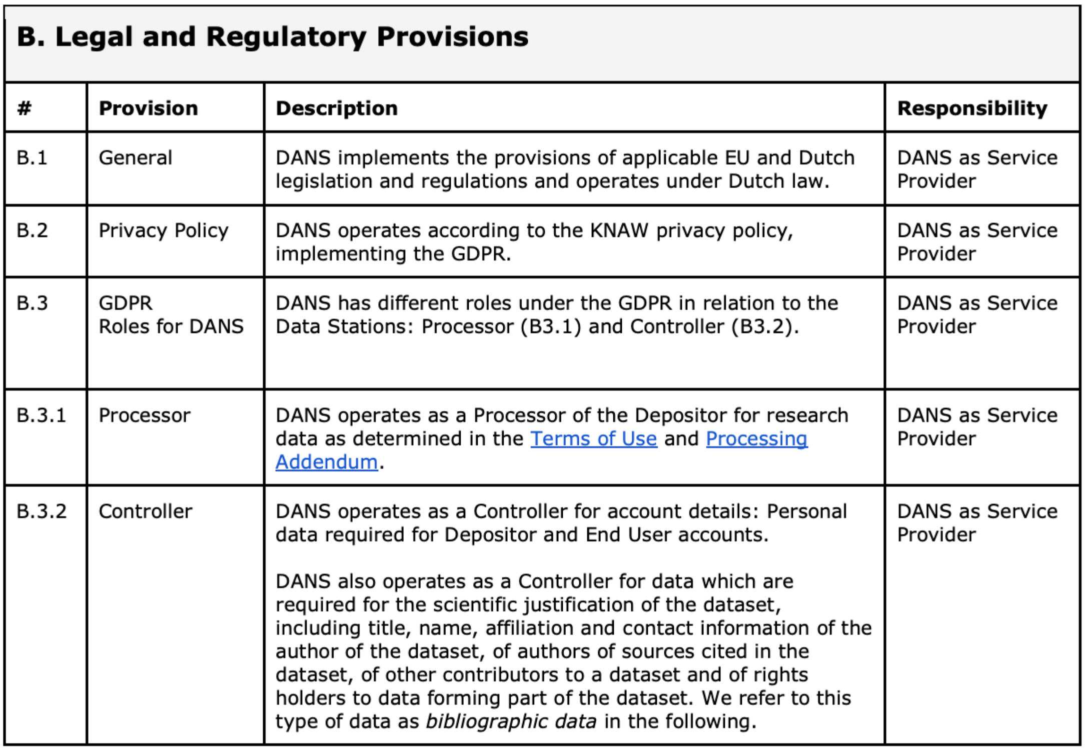
1. Part I: Provides considerations and explanations of the approach followed.
2. Part II: The LTP Policy Template.
3. Part III: Appendices, providing further detail and explanation.
4. Part VI: Specific appendices for EUDAT services B2SHARE and B2SAFE.

4.2.3 An implementation: DANS Data Stations Policy

An example of the implementation of the LTP-Policy Template, concerns the LTP-Policy of the DANS Data Stations (DANS Data Stations Policy[68]). The policy is described and published on a DANS website page[69] that is about "Using the DANS Services".

The DANS Data Stations Policy outlines the policies concerning the Data Stations in more detail and also discusses the preservation plan.

The DANS Data Stations Policy is based on LTP-Policy Template; it has been adjusted to the needs and insights of DANS. Per article, it covers what is being provisioned and who in which role carries responsibility for it, as can be seen in the heading from the screenshot given in Figure 22.



B. Legal and Regulatory Provisions

#	Provision	Description	Responsibility
B.1	General	DANS implements the provisions of applicable EU and Dutch legislation and regulations and operates under Dutch law.	DANS as Service Provider
B.2	Privacy Policy	DANS operates according to the KNAW privacy policy, implementing the GDPR.	DANS as Service Provider
B.3	GDPR Roles for DANS	DANS has different roles under the GDPR in relation to the Data Stations: Processor (B3.1) and Controller (B3.2).	DANS as Service Provider
B.3.1	Processor	DANS operates as a Processor of the Depositor for research data as determined in the Terms of Use and Processing Addendum.	DANS as Service Provider
B.3.2	Controller	DANS operates as a Controller for account details: Personal data required for Depositor and End User accounts. DANS also operates as a Controller for data which are required for the scientific justification of the dataset, including title, name, affiliation and contact information of the author of the dataset, of authors of sources cited in the dataset, of other contributors to a dataset and of rights holders to data forming part of the dataset. We refer to this type of data as *bibliographic data* in the following.	DANS as Service Provider

Figure 22. Screenshot taken from the DANS Data Stations Policy (1st June 2023)

[68] https://dans.knaw.nl/wp-content/uploads/2023/04/DANS-Data-Stations-Policy-20230524.pdf (visited: 1st June 2023)
[69] https://dans.knaw.nl/en/using-data-stations/ (visited: 1st June 2023)

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

4.2.4 EOSC Long-Term Data Preservation Task Force

The EOSC Long-Term Data Preservation Task Force (EOSC TF LT) provides recommendations on the vision and sustainable implementation of long-term data preservation policies and practices, as well as suggestions to later strategy execution.

As indicated in the previous deliverable D4.2, the LTP Policy Template has been brought to the attention of the Task Force to get it promoted and disseminated in a broader context within the EOSC community. The scope and outreach of the template will be further increased this way.

The Task Force has alerted the authors of the LTP template, that a discussion document, named: "Curation & Preservation Levels: CoreTrustSeal Discussion Paper"[70] is ready in which revised Curation and Preservation levels from CoreTrustSeal are discussed because the need for clearer specification of preservation levels has become clear. The current CTS curation levels are included in the LTP Policy Template by reference and should therefore be monitored for changes to keep the LTP Policy Template up-to-date. However, to date, this has not been substantiated anywhere.

4.3 The Digital Preservation Service: an implementation

4.3.1 Introduction

This part of the deliverable describes the Digital Preservation Service (DP Service). This is a service that implements the technical specifications that were created within this task 4.3. It describes what is technically needed to enable dataset transfer from a short-to-midterm web repository to a long-term preservation archive that is supported by the LTP policy applicable locally. The objective is to transfer datasets from the EUDAT B2SHARE web repository and archive them into the DANS Data Vault, which is the LTP archive component of the DANS Data Station. Input to the Data Vault goes through Dataverse[71] instances, in particular the discipline-oriented Data Stations as well as a special Data Station that only has the function of passing data sets from third parties (not in the discipline-oriented Data Stations) to the Data Vault.

All digital objects that make up a dataset should be transferred, but also the metadata and other related artefacts, like authors, should be part of the archival package that is to be deposited. In this part of the deliverable, we will discuss the use-cases, the technical specifications document and a proof-of-concept (POC) implementation of such a service that should adhere to the LTP policies, of both the data-node (B2SHARE) as well as the service-node (DANS Data Vault).

The technology readiness level (TRL) of the service is somewhere between TLR 2 ("technology concept formulated") and TLR 3 ("experimental proof-of-concept"). This is caused by, among other things, the readiness levels of existing systems that are part of the DP Service and the maturity level of the DP Service components itself. The service implementation should be regarded as proof of the validity of the specifications in Appendix 2.

4.3.2 Use cases

Within this project, we have identified the following two use cases on which the DP Service is based:

[70] https://doi.org/10.5281/zenodo.6908019

[71] https://dataverse.org/ (visited: 11th May 2023)

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

1. **Single dataset record archiving**
   A single dataset will be transferred to an out-sourced LTP-Archive on demand, according to the repository LTP Policy, by an authenticated repository dataset owner.

2. **(Auto) Archive Community records**
   B2SHARE supports the concept of communities, administering their own metadata schemas and publication requirements. All records within such a community will be auto-archived, according to a community archival agreement.
   Opt-out/in option could be offered to the dataset owner to bypass the community policy.
   This use case is a variant of use case 1 and can quickly be achieved once use case 1 is in place, because it has great overlap with use case 1. The difference is that not the authenticated user, but the web repository itself makes the (auto) archiving request. It will not be discussed in this document, nor will it be implemented in the POC.

Apart from these two use cases, many more can and need to be defined, especially if a higher software maturity level of the DP Service is required. In our proof-of-concept implementation, we looked at the first use case: Single dataset record archiving (out-sourced). A single dataset described on an EUDAT B2SHARE landing page will be transferred to the DANS CTS certified data vault, on demand, by an authenticated repository dataset owner. The service will be carried out according to the policies that are in place. The technical specifications document will be discussed in the next section.

4.3.3 Technical specifications for the DP Service

Technical specifications are required to implement the use cases from the previous section in a standardised way. The specifications for the DP Service (DPS) are based on the use of open web standards and are described in full in Appendix 2. The specification document has been created by the DICE task 4.3. Appendix 2 is a snapshot of the documentation website, of which the content is generated by a GitHub[72] repository. The latest version of the specifications is available from this website. The DSP specifications document can also be used by other parties as a guide or example on how to implement such an archival service, by using open web standards only.

The specifications are in line with the Event Notifications in Value-Adding Networks[73] project, which details a profile for using Linked Data Notifications with ActivityStreams2 payloads in value-adding networks. The profile lists examples of LDN+AS2 messages (see Activity Streams 2 section) between web repositories and LTP archives, on which our example payloads are also based.

Another project with which this specification is in line is the COAR Notify[74] project. COAR Notify is a repository and stand-alone (peer) review services interoperability project.

All payloads in the DPS specifications include the COAR Notify context file (@context). This context file defines commonly used namespaces in the profile and are listed in Appendix 2. It also includes the COAR Notify vocabulary, which may sooner or later define terms that can be used within the DP Service.

[72] https://github.com/Dans-labs/ddps-docs (visited: 11th May 2023)
[73] https://www.eventnotifications.net/ (visited: 11th May 2023)
[74] https://notify.coar-repositories.org/ (visited: 11th May 2023)

The open web standards that are involved in the core implementation of the Digital Preservation Service are listed below. In the next section motivations are discussed why we're using these open web standards in the DP Service.

Linked Data Notification

W3C Linked Data Notification[75] (LDN) is an HTTP-based notification (push) protocol. It will be used for repository/archive communication.

It is a messaging system that implements the concepts of Senders, Receivers and Consumers and can be seen as an email system for machines. These messages, or notifications should be expressed in RDF and can contain any data.

Senders and consumers must implement an LDN inbox. The resource's Inbox URL can be (auto)discovered through a relation in the HTTP Link header or body of the resource.

Activity Streams 2.0

W3C Activity Streams 2.0[76] (AS2) provides a foundational vocabulary for messaging about activities that involve web resources. A message profile will be used for LDN notifications payload exchanged by repositories and archives.

Within this documentation, the Linked Data Notifications (LDN) with ActivityStreams2 (AS2) payloads will be referred to as: 'LDN+AS2 notifications'.

The notification payloads should use JSON-LD as default syntax, but other RDF syntaxes may also be used.

Signposting

Signposting is a REST/HATEOAS "follow your nose" (navigational) approach to make the scholarly web more friendly to machines; it leverages IETF RFCs and IANA-registered link relation types. Typed links are used to allow machines to uniformly navigate scholarly artefacts irrespective of the repository they reside in. Implementations of the framework into repositories allow machines to determine which content resources are associated with the data object it describes. For resources of any media type, these Typed Links are provided in HTTP Link headers[77], or for HTML resources like landing pages, they may additionally be provided in HTML link[78] elements. This way the links can be auto discovered by machines to support machine interoperability.

The FAIR Signposting Profile[79] is a lightweight, yet powerful approach to increase the FAIRness of scholarly objects. It can be used by repositories as a means to allow archives to determine which web resources need to be retrieved in response to an on-demand archiving request.

4.3.4 Technical motivation

An archiving service can be accomplished in different ways, with different techniques. In this section we motivate the chosen techniques.

[75] https://www.w3.org/TR/ldn/ (visited: 11th May 2023)
[76] https://www.w3.org/TR/activitystreams-core/ (visited: 11th May 2023)
[77] http://tools.ietf.org/html/rfc5988 (visited: 11th May 2023)
[78] https://developer.mozilla.org/en-US/docs/Web/HTML/Element/link (visited: 11th May 2023)
[79] https://signposting.org/FAIR/ (visited: 11th May 2023)

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

Asynchronous

The intended communication style among nodes is point-to-point, requiring no centralized hubs. Interactions among nodes (Service Nodes and Data Nodes) are necessarily asynchronous because certain notification patterns do not require a response ("fire and forget") and, in patterns that do, such as requesting an action, the time between a request and the announcement of the Action Result is unpredictable, as the recipient may complete tasks at its own pace. Pushing any data at any moment to an archive might cause resource problems at the server side. Therefore, it is push-oriented, with only the relevant nodes being updated about new information as it becomes available.

Lightweight

This approach is lightweight. It does take relatively little computing resources to implement, both from the client as from the server side.

Multi-purpose

The LDN+AS2 notifications approach can also be used for other purposes with the same investment. For example, the peer reviewing service COAR Notify, which is also strongly aligned with the Event Notifications in Value-Adding Networks Profile.

Open Web standards

The DP Service specification is built solely on Open Web Standards. Because of its obvious benefits like larger audience and community, forward compatibility with browsers and cost savings; no patents or licensing.

Overall, embracing Open Web standards ensures a more inclusive, compatible, secure, and innovative web ecosystem, benefiting both developers and end users.

4.3.5 Digital Preservation Service

In this section the proof-of-concept for the DP Service, that was created within task 4.3, will be described. It implements the DP Service technical specifications. These specifications are intended as a stand-alone document in practice and are available from Appendix 2.

It follows use case 1: Individual record archiving (outsourced). Use case 2 can be derived from use case 1. However, use case 2 targets a (B2SHARE) community and not a single user.

If implemented, this should be reflected in the policies applicable to the relevant community. This could include auto-archiving of a dataset that will be deposited to a specific community.

We will start with a short discussion of the status of the B2SHARE Signposting module, and then look at the different components that make up the DP Service. These components and their relations are discussed in detail in the “Requirements and Specifications” in Appendix 2.

4.3.5.1 Status B2SHARE Signposting module

The B2SHARE Signposting module has been implemented on the EUDAT B2SHARE production environment and was released on 2023-04-26 by CSC[80] in Finland. Preliminary work has been carried out in the context of the EOSC-hub[81] project that has ended in December 2020.

[80] https://www.csc.fi/about-us (visited: 11th May 2023)
[81] https://www.eosc-hub.eu/ (visited: 11th May 2023)

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

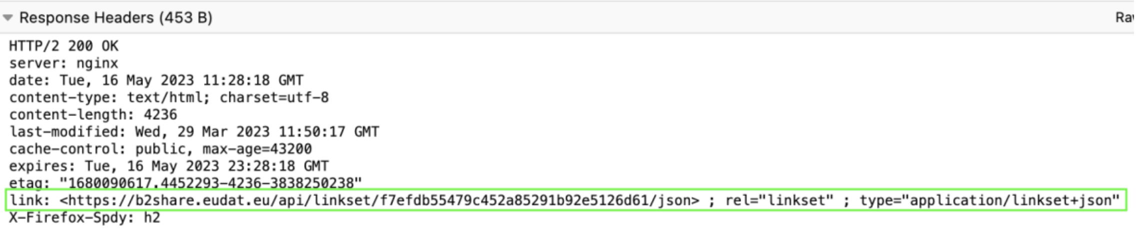
The B2SHARE Signposting module creates a Link Set endpoint[82] based on the metadata and data objects of the dataset described on a landing page in B2SHARE. Link Sets are specified in RFC9264[83]. The Link Set URL for a B2SHARE record, is published, according to the Signposting specifications, in the HTTP response headers for a GET request in the 'Link' header field of the landing page URL, in the following format:

```
Link: <https://b2share.eudat.eu/api/linkset/<<record_id>>/json>;
rel="linkset"; type="application/linkset+json"
```

In which the "*<<record_id>>*" is a placeholder for the B2SHARE record identifier.

Figure 23 shows the HTTP response headers for a GET request of a B2SHARE landing page; https://b2share.eudat.eu/records/f7efdb55479c452a85291b92e5126d61

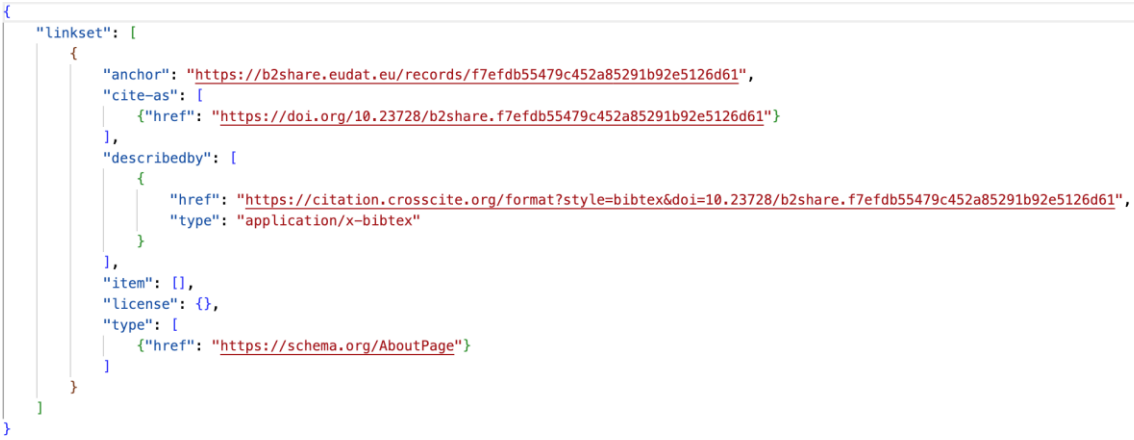
The Link header field is highlighted in the green box. The URL value will provide the serialized Link Set in JSON format.



```
HTTP/2 200 OK
server: nginx
date: Tue, 16 May 2023 11:28:18 GMT
content-type: text/html; charset=utf-8
content-length: 4236
last-modified: Wed, 29 Mar 2023 11:50:17 GMT
cache-control: public, max-age=43200
expires: Tue, 16 May 2023 23:28:18 GMT
etag: "1680090617.4452293-4236-3838250238"
link: <https://b2share.eudat.eu/api/linkset/f7efdb55479c452a85291b92e5126d61/json> ; rel="linkset" ; type="application/linkset+json"
X-Firefox-Spdy: h2
```

Figure 23. B2SHARE HTTP response headers including the Link Set URL in the "Link" field

The (JSON) response of the Link Set URL (i.e.: https://b2share.eudat.eu/api/linkset/f7efdb55479c452a85291b92e5126d61/json) is shown in Figure 24.



```
{
    "linkset": [
        {
            "anchor": "https://b2share.eudat.eu/records/f7efdb55479c452a85291b92e5126d61",
            "cite-as": [
                {"href": "https://doi.org/10.23728/b2share.f7efdb55479c452a85291b92e5126d61"}
            ],
            "describedby": [
                {
                    "href": "https://citation.crosscite.org/format?style=bibtex&doi=10.23728/b2share.f7efdb55479c452a85291b92e5126d61",
                    "type": "application/x-bibtex"
                }
            ],
            "item": [],
            "license": {},
            "type": [
                {"href": "https://schema.org/AboutPage"}
            ]
        }
    ]
}
```

Figure 24. B2SHARE serialised Link Set response

4.3.5.2 Repository Dummy Service Component

Despite all the effort that was put into the B2SHARE Signposting module, some resources are still missing from the serialised Link Set which makes this module actually unsuitable for use

[82] https://b2share.eudat.eu/api/linkset/cc7b9f5b8e6042f8b3335c28e3a5ee4a/json (visited: 11th May 2023)
[83] https://www.rfc-editor.org/rfc/rfc9264.html (visited: 11th May 2023)

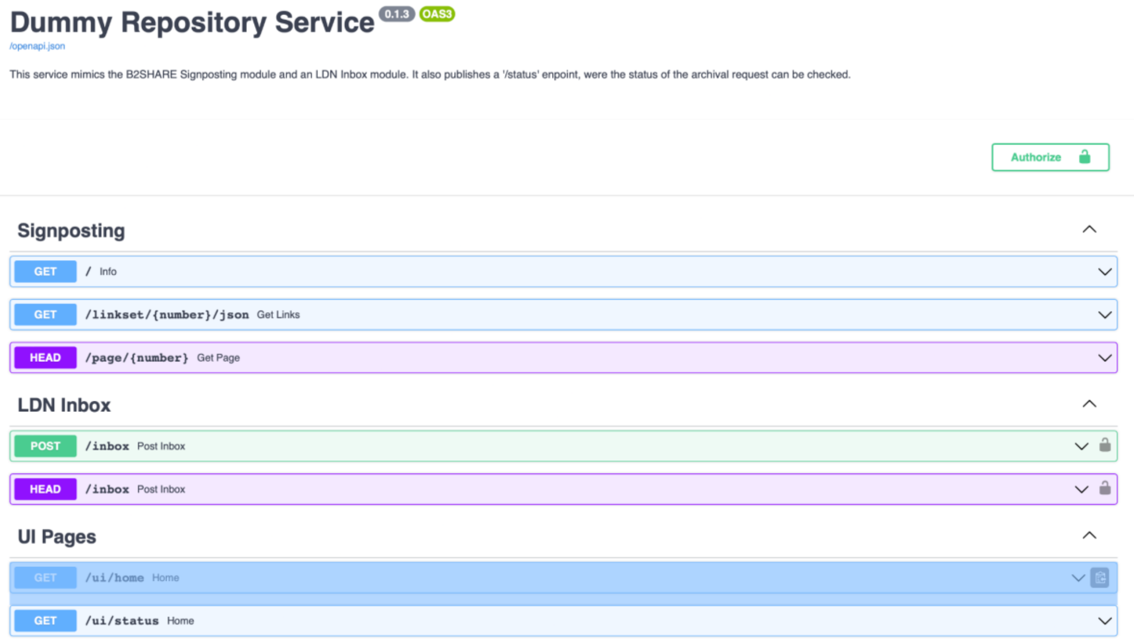
within the DP Service. Even more, developing an LDN Inbox module on B2SHARE was also not within reach of the project, but still mandatory for running a successful POC.

We decided to create a "dummy" Repository Service[84] component, that mimics Signposting and implements an LDN Inbox that can be used in the proof-of-concept to communicate with.

In addition to this, the dummy service also publishes a "status" endpoint (/ui/status), from which the current status of the archival request can be obtained (i.e. 'LTP Request Pending' , 'LTP In Progress' & 'LTP Archived'). These statuses should otherwise be updated by the B2SHARE web-interface itself on the landing page, as can be seen from the sequence diagram (figure 8) in the Appendix 2. From the LDN point of view, this component acts as the "Data node".

The dummy service implements the OpenAPI Specification 3[85] (OAS), which is an HTTP (REST) API written in Python programming language using FastAPI[86].

Figure 25 shows the OpenAPI user interface of the dummy service from which the different endpoint groups can be recognised as "Signposting", "LDN Inbox" and "User Interface pages" respectively.



Dummy Repository Service

/openapi.json

This service mimics the B2SHARE Signposting module and an LDN Inbox module. It also publishes a '/status' enpoint, were the status of the archival request can be checked.

Signposting

GET / Info

GET /linkset/{number}/json Get Links

HEAD /page/{number} Get Page

LDN Inbox

POST /inbox Post Inbox

HEAD /inbox Post Inbox

UI Pages

GET /ui/home Home

GET /ui/status Home

Figure 25. OpenAPI user interface of the dummy service. Exposing a Signposting, LDN Inbox and Status endpoint

4.3.5.3 *Archival Bot Component*

The Archival bot[87] is a middleware application written in python. It takes care of the LDN+AS2 notifications communication between the Web repository and the LTP Archive. This is the service node seen from an LDN perspective.

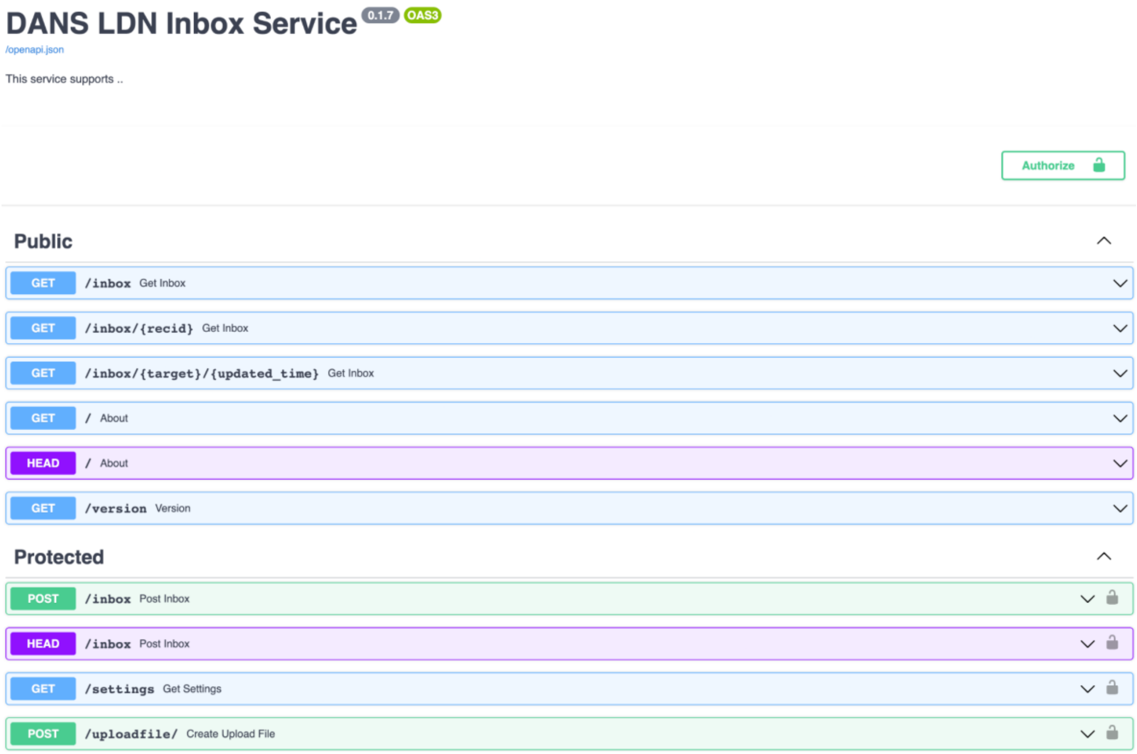
[84] https://github.com/Dans-labs/dummy-repository-service (visited: 11th May 2023)
[85] https://swagger.io/specification/ (visited: 11th May 2023)
[86] https://fastapi.tiangolo.com/ (visited: 11th May 2023)
[87] https://github.com/Dans-labs/ldn-inbox-service (visited: 11th May 2023)

This middleware component could be implemented in several ways, depending on the architecture used. It could for instance be a plugin or module on the archive.

The archival bot that was developed for the POC contains three functional containers that together make up the Archival Bot[88] component. First there is the mandatory LDN Inbox API (Figure 26) to send and receive LDN+AS2 notifications. Secondly a Signposting 'module' was included. This part is necessary to discover the Link Set of a landing page and collect all resources involved in the archival request.

The Archival bot retrieves data and metadata files that are available from the Link Set (Signposting). These resources are packed into a BagIt[89] file packaging archival format, which is supported by the Data Vault.

After creation of such an archive package, the Archival bot should deposit it to the LTP Archive, that allows depositing archival packages via the repository deposit protocol SWORD2[90]. The result is an archive URI that should be published on the landing page in the web repository. This way repository users know that the current dataset has been offered for archiving and also know where to find it (which archive, location).



DANS LDN Inbox Service 0.1.7 OAS3

/openapi.json

This service supports ..

Authorize

Public

GET /inbox Get Inbox

GET /inbox/{recid} Get Inbox

GET /inbox/{target}/{updated_time} Get Inbox

GET / About

HEAD / About

GET /version Version

Protected

POST /inbox Post Inbox

HEAD /inbox Post Inbox

GET /settings Get Settings

POST /uploadfile/ Create Upload File

Figure 26. OpenAPI user interface of the LDN Inbox service of the archival bot

4.3.5.4 DANS Data Vault

The DANS Data Vault is a long-term preservation archive. This is implemented as a collection of files stored on tape via SURF's Data Archive service. Some components of the vault are currently still in transition from DANS Easy[91] and therefore not all functionality is yet available. One of

[88] https://github.com/Dans-labs/archivalbot (visited: 11th May 2023)
[89] https://datatracker.ietf.org/doc/html/rfc2629 (visited: 11th May 2023)
[90] https://sword.cottagelabs.com/previous-versions-of-sword/sword-v2/ (visited: 11th May 2023)
[91] https://easy.dans.knaw.nl/ui/home (visited: 11th May 2023)

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

them that is still missing is the dd-vault-catalog[92], which is a catalog of the contents of the DANS Data Vault. Therefore, at the time of writing this document, it is not possible to lookup the archived dataset in the Vault externally. The archival deposits have therefore been carried out on the test environment and on local (development) deployments of the DANS Vault.

4.3.6 Discussion

By implementing the POC for the DP Service, we demonstrated that archival requests from a web repository to an out-sourced LTP archive can be carried out by using open web standards like Signposting and LDN+AS2 notifications.

These open web standards play a major role in different aligned projects, like CAOR Notify and the Event Notifications in Value-Adding Networks project. This is also the reason why Dataverse has announced to implement an experimental Linked Data Notification API / inbox[93] in their latest release. This allows for receiving LDN+AS2 messages indicating a link between an external resource and a Dataverse dataset. The motivation is to support a use case where Dataverse administrators may wish to create back-links to the remote resource. This has great similarities with the specifications and the POC described in this document.

The technology readiness level (TRL) of the DP Service is somewhere between TLR 2 (“technology concept formulated”) and TLR 3 (“experimental proof-of-concept”). Unfortunately, no more resources were available in the project, to bring the DP Service to a higher readiness level. For instance, no Error handling could be implemented due to time constraints.

Other aspects also suffered from this. The Link Set conversion to the archival package is another example. The Archival bot needs to know how the Link Set format needs to be converted, or mapped, to a specific archival package format. The DANS Data[94] Volt requires a BagIT package that is deposited by SWORD2. So, the Archival bot must know how to map the Link Set items to the (metadata) scheme of the archival format. In the POC, we have created an archival packager application that converts a Link Set file to a deposit BagIt file packaging format. This was only possible because the Archival bot has knowledge about the "describedby" types it can expect. For instance, in our example, it knows how to extract, or parse, metadata from the "application/x-bibtex" mime type found in the Link Set. Other mime types, like "application/vnd.datacite.datacite+json" were not implemented in the Archival bot and will therefore not be recognised if published in the Link Set. In this sense, the Archival bot behaves as a rule engine and could also be implemented as such.

4.3.6.1 Achievements

This section contains an overview of the achievements of the work done in Task 4.2 of the DICE project regarding the Digital Preservation Service:

- Defined the technical specifications for the DP Service (Appendix 2).
- Implemented (partial) Signposting B2SHARE module.
- Developed a Proof-of-concept for the DP Service at TRL 2, 3.

4.3.6.2 Future work

Within the time constraints of the project, we could not implement all aspects of the POC.

[92] https://dans-knaw.github.io/dd-vault-catalog/ (visited: 11th May 2023)
[93] https://guides.dataverse.org/en/latest/api/linkeddatanotification.html (visited: 11th May 2023)
[94] https://dataverse.org/ (visited: 11th May 2023)

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

The B2SHARE related issues should be tackled by EUDAT in the future, because these improvements will not only benefit the POC that is created within this project task, but will also increase the FAIRness (signposting) and accessibility (LDN-inbox) of B2SHARE. This has been communicated with EUDAT.

In addition to this, the technical specifications listed in Appendix 2, will be brought to the attention of the ongoing FAIRCORE4EOSC project, to the task developing tools and services for archival, reference, description and citation of research software artifacts, with the Software Heritage universal source code archive, using the CodeMeta standard, and the Software Heritage intrinsic identifiers (SWHID).

In the next paragraphs, we list aspects which need further improvement.

Shapes Constraint Language (SHACL)[95] validation: To create a more secure and reliable service, all the notifications should be validated. We suggest using SHACL for this. SHACL is a “schema” language for validating RDF graphs against a set of conditions. This way the system can check whether an incoming message is valid and valuable to the network. It may check for required namespaces, required AS2 types and certain values that should be supplied in the LDN message.

Improve LDN Inbox Error Handling: These specifications prescribe the nature of HTTP responses for cases when a notification is successfully received in an LDN Inbox. However, what should an LDN-inbox reply when it rejects the incoming notification? At the moment the specification does not offer a way to indicate why it is rejecting the `as:Offer`. Obviously, this error handling can be carried out in combination with the proposed SHACL validation. More on LDN Inbox error handling can be found in the Event Notifications in Value-Adding Networks specifications.

Improve the B2SHARE Signposting module: The Signposting module for B2SHARE does not seem to fully implement the FAIR Signposting profile. For instance, “authors” and “item” links (links to the data objects) are currently still missing from the serialised Link Set. This seems to be the case for every record.

Especially the omission of the “item” links makes it impossible to implement an archival service on top of it, because these are mandatory to create an archival package. CSC has been informed on this issue.

B2SHARE Archive button: The user interface (UI) of B2SHARE remained unchanged due to time limitations. As a result, an archiving status button was not implemented, which would initiate a Long-Term Preservation (LTP) request for the dataset on the landing page or, if already archived, a location to the archive.

B2SHARE LDN Inbox module: Just like the Signposting module adds value to the B2SHARE repository for increasing FAIRness, an LDN Inbox will also add value to it, not only for this project. In this way the system will be prepared for future technologies. As an example of this we mention the Dataverse experimental Linked Data Notification API. This API can be enabled to conduct experiments with LDN+AS2 notifications.

[95] https://www.w3.org/TR/shacl/ (visited: 11th May 2023)

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

5 Report on enabling sensitive data workflow by adapting standard interoperability frameworks to connect the endpoints (T4.4)

5.1 Introduction

One of the goals in task 4.4 was to further develop Secure B2SHARE and have it deployed in sensitive data infrastructure provided by CSC and TSD (Norwegian name: Tjenester for Sensitive Data), the research platform for working with sensitive data at the University of Oslo. Once deployed, work would continue by identifying possible ways for other services to utilize data stored in sensitive data infrastructures. In essence, Secure B2SHARE would be used to interact with the sensitive data infrastructure and if possible, to enable processing of sensitive data in services deployed outside of the sensitive data infrastructure without compromising data security and sensitivity of the data. During the work, it was discovered that supporting processing of sensitive data outside of the sensitive data infrastructure where data has been stored, is technologically possible, but difficult due to the nature of sensitive data; trust must be established between the data provider and data processing service provider, and in some cases, there are policies in place which prevent forming this trust. The data provider cannot provide the sensitive data outside of sensitive data infrastructure without compromising the promises made to data owners and legislative requirements for storing and processing of sensitive data. In some cases, exporting sensitive data outside of the sensitive data infrastructure is possible, but in these cases it is ultimately up to the data owner to make sure that processing of the data complies with legislation and that the original purpose of sensitive data gathering is honoured.

In this chapter we report work done in task 4.4, by describing first the Secure B2SHARE concept and then describing how the two Secure B2SHARE instances realize the concept on two distinct sensitive data infrastructures at CSC and at TSD. We describe how EUDAT B2FIND is used to improve discoverability and through this promote use of sensitive data stored in a Secure B2SHARE instance. We also describe one type of integration we designed for processing of sensitive data in a service located outside of the sensitive data infrastructure where sensitive data is stored.

5.2 Secure B2SHARE concept

The Secure B2SHARE concept has three distinct components:

1. B2SHARE[96]
2. Secure Data Submission service (SDS)
3. An Authorization service.

The dataset owner uploads files in SDS, creates datasets and describes metadata for the datasets in B2SHARE, and manages authorization to datasets in the Authorization Service.

Datasets created in B2SHARE always only refer to files previously uploaded through SDS. Files themselves are stored in Secure Storage, and not in B2SHARE. Researchers can find datasets with search functionality provided by B2SHARE, or through metadata discovery services such as B2FIND.

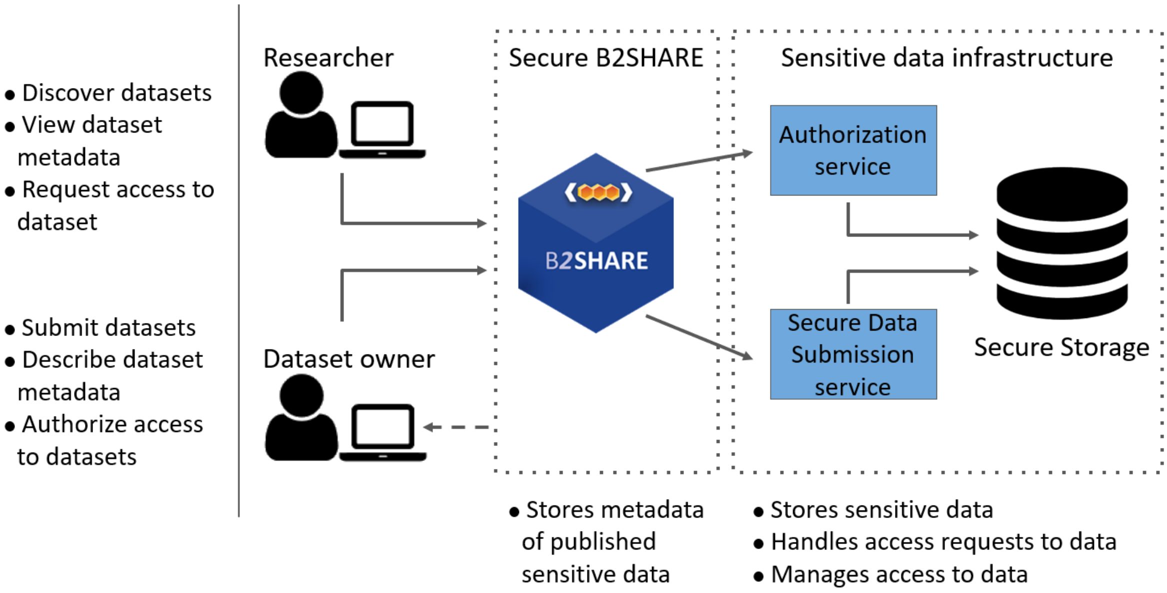
When a researcher discovers an interesting dataset, an access request must be made. The data owner (or a representative) reviews the access request and either rejects or accepts it. If

[96] B2SHARE is a catalog service for publishing and sharing of research data

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

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 doesn't specify how access to sensitive data should be implemented, as this very much depends on the sensitive data infrastructure Secure B2SHARE is implemented on. There currently are two instances of Secure B2SHARE; one at TSD[97] in Norway and one at CSC[98] in Finland. Both realize the Secure B2SHARE components on provider specific sensitive data infrastructure.



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

Figure 27. Secure B2SHARE concept

5.3 Secure B2SHARE TSD instance

The uploading component of Secure B2SHARE @ TSD is located in a portal that is accessible only from the inside of the secure TSD network environment, as it is intended to be used for uploading of data that already is located in secure storage. As access to this internal web portal requires that a user is connecting from inside of TSD, they will have logged in using two-factor authentication for a quite high level of security. Furthermore, it is possible to limit uploading access to only users that are granted export and/or publishing rights at the project level.

A versioned dataset will be created for the uploaded data copy, which is made immutable and assigned an identifier. The dataset's assigned identifier, along with some metadata, will be transmitted to a message queue inside of TSD that federates to an external (to the TSD network environment) message broker.

Through information sent in messages relayed via these message brokers, an external helper service will perform the task of creating and updating catalog records in the publicly available Secure B2SHARE instance with metadata and the dataset identifier required for access requests, through use of the B2SHARE Application Programming Interface (API).

[97] The University of Oslo's Services for sensitive data research platform

[98] CSC – IT Center for Science, a Finnish state and higher education institutions owned non-profit company providing information technology solutions

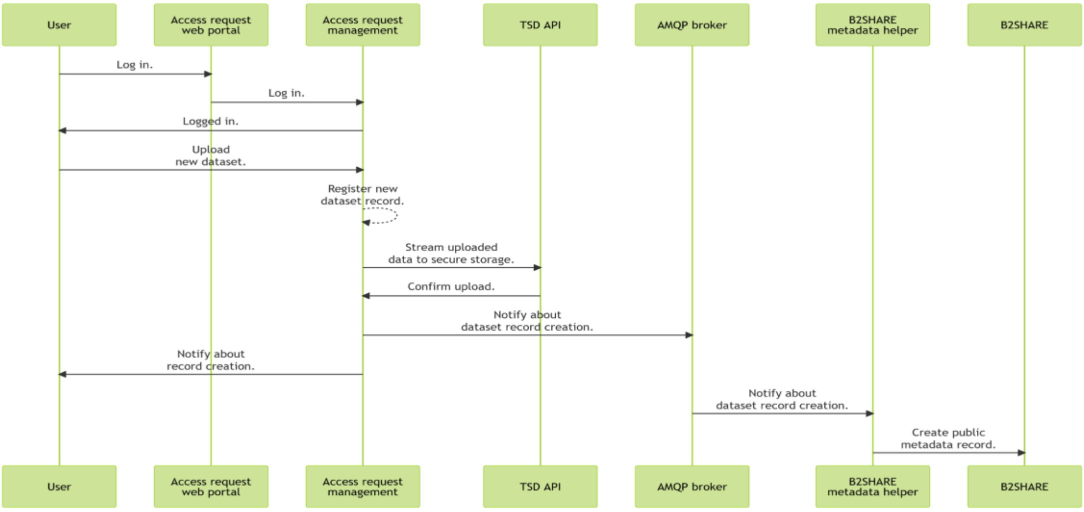

User
Access request web portal
Access request management
TSD API
AMQP broker
B2SHARE metadata helper
B2SHARE
Log in.
Log in.
Logged in.
Upload new dataset.
Register new dataset record.
Stream uploaded data to secure storage.
Confirm upload.
Notify about dataset record creation.
Notify about record creation.
Notify about dataset record creation.
Create public metadata record.
User
Access request web portal
Access request management
TSD API
AMQP broker
B2SHARE metadata helper
B2SHARE

Figure 28. Sensitive data upload process at Secure B2SHARE TSD instance

TSD realizes the Authorization service of Secure B2SHARE with Nettskjema[99] component, which provides the interface for submitting dataset access requests. Nettskjema supports several external authentication schemes for verification of identity, such as ID-porten (Norwegian national identity), Norwegian academic identity federation (FEIDE), TSD user login and Educloud user login (another research platform developed by the University of Oslo). eIDAS[100] support in ID-porten is expected to bring cross-borders assurance of identity verification for citizens of many European countries.

When a person wishes to request access to a dataset, the identifier will automatically be passed to the web form that they then authenticate to fill out when redirected via B2SHARE's web interface. The web form data is transmitted to TSD and the data owner is notified that an access request has been made. Data owner can reject or approve the access request. In case data owner approves the access request, data owner utilizes TSD specific access controls to authorize access to sensitive data for the requester.

[99] Nettskjema is a survey / data collection service developed by the University of Oslo

[100] eIDAS stands for electronic Identification, Authentication and Trust Services and is a EU regulation for electronic identification interoperability between the EU/EEC states.

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

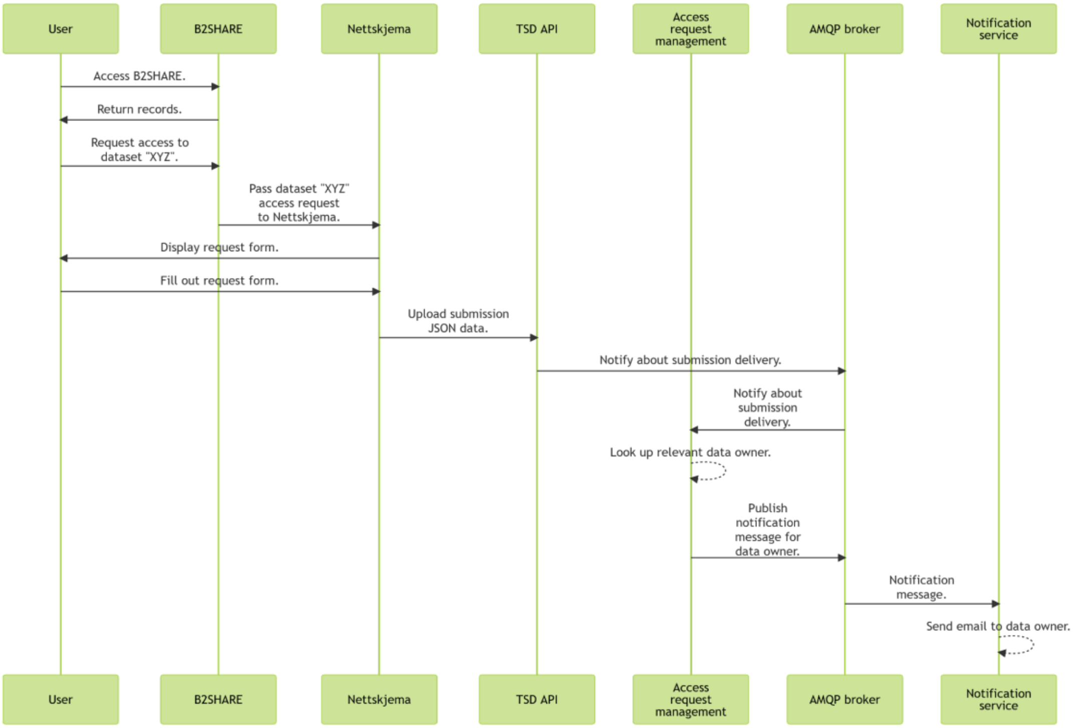

User
B2SHARE
Nettskjema
TSD API
Access request management
AMQP broker
Notification service
Access B2SHARE.
Return records.
Request access to dataset "XYZ".
Pass dataset "XYZ" access request to Nettskjema.
Display request form.
Fill out request form.
Upload submission JSON data.
Notify about submission delivery.
Notify about submission delivery.
Look up relevant data owner.
Publish notification message for data owner.
Notification message.
Send email to data owner.
User
B2SHARE
Nettskjema
TSD API
Access request management
AMQP broker
Notification service

Figure 29. Access request process at Secure B2SHARE TSD instance

5.4 Secure B2SHARE CSC instance

Secure B2SHARE instance at CSC is still under development. The following description depicts the functionality of the instance as it is when the instance is ready. In its current state, Secure B2SHARE instance at CSC is not properly integrated with SD Submit and SD Apply services of CSC's Sensitive Data Services (SDS), which means that datasets cannot be created to Secure B2SHARE instance at CSC. This also means that access requests cannot be sent from SecureB2SHARE UI. Further development is needed to address these shortcomings.

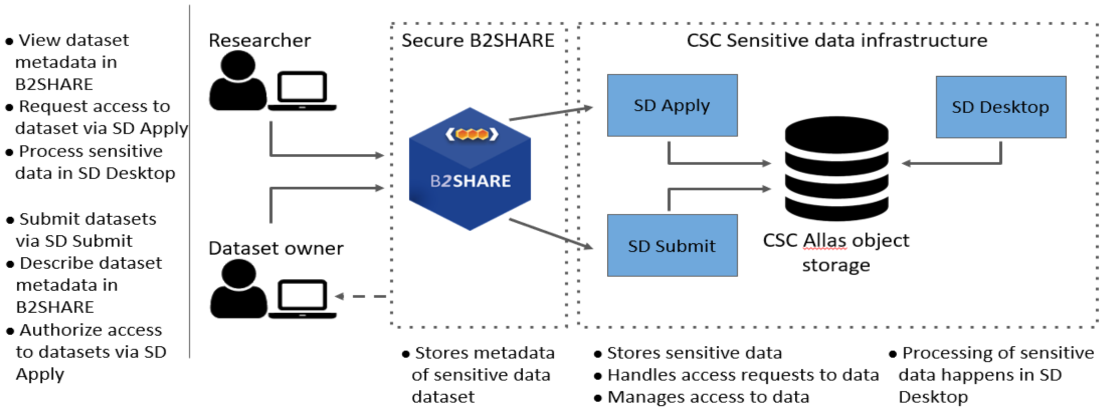

• View dataset metadata in B2SHARE
• Request access to dataset via SD Apply
• Process sensitive data in SD Desktop
• Submit datasets via SD Submit
• Describe dataset metadata in B2SHARE
• Authorize access to datasets via SD Apply
Researcher
Dataset owner
Secure B2SHARE
B2SHARE
CSC Sensitive data infrastructure
SD Apply
SD Desktop
SD Submit
CSC Allas object storage
• Stores metadata of sensitive data dataset
• Stores sensitive data
• Handles access requests to data
• Manages access to data
• Processing of sensitive data happens in SD Desktop

Figure 30. Secure B2SHARE components in CSC deployment

In order to include files containing sensitive data in a dataset that is to be published in CSC's Secure B2SHARE instance, the owner of data must first upload sensitive data through CSC Sensitive Data Services (SDS) SD Submit service. Before uploading, the data owner must encrypt the data with a CSC defined public key and upload the data with SFTP protocol using e.g. a certain

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

naming schema as instructed by SD Submit. Once the data is uploaded, the owner can define what kind of licenses and rules must be agreed, if someone wants to process the data. Data owner can also define people that can process (ultimately reject or approve) access requests to the data. After this, a draft metadata record is created in Secure B2SHARE and the sensitive data owner uploaded to SD Submit is associated with this metadata record draft in Secure B2SHARE.

Data owner then uses Secure B2SHARE to describe metadata of the dataset and publish the metadata record. By this published metadata record, the dataset can be discovered; people can browse and search the Secure B2SHARE instance for datasets and dataset metadata can be harvested to dataset aggregation services such as EUDAT B2FIND to increase findability. Sensitive data of the dataset is never openly available. In order to gain access to the sensitive data of the dataset, a user must make an access request in CSC Sensitive Data Services SD Apply service.

Access request can be initiated in SecureB2SHARE after which the user is redirected to SD Apply service. User must authenticate and agree to comply with terms and license(s) defined for processing data of this dataset. After this, the access request can be submitted for review, where data owner or people data owner has defined as possible reviewers, can process the access request. In case the access request is accepted, the user who requested the access can access the sensitive data of this dataset via CSC Sensitive Data Services SD Desktop service.

In SD Desktop, user utilizes a remote desktop connection via web browser to process data in a secure environment otherwise disconnected from the internet. Sensitive data of the dataset cannot be exported from the secure environment. Results of data processing can be exported, but only after a review where content of the results are checked for any sensitive data. If the results are properly anonymized they can be exported by CSC Sensitive Data Services administrators.

5.5 Secure B2SHARE interoperability for sharing of metadata records

EUDAT B2FIND service is used as a metadata aggregation service to improve discoverability of datasets on Secure B2SHARE instances. B2FIND must be configured to fetch metadata of sensitive dataset from a Secure B2SHARE instance. Once configured, B2FIND fetches metadata via OAI-PMH[101] protocol from the OAI-PMH endpoint exposed by the Secure B2SHARE instance. B2FIND fetches metadata records from a Secure B2SHARE instance using EUDAT Core metadata schema definition. B2FIND performs conversion from EUDAT Core format to its own internal format, after which metadata records can be discovered from B2FIND web portal.

In general, anyone can harvest metadata records from a Secure B2SHARE instance via OAI-PMH. The OAI-PMH interaction with a B2SHARE instance is the same that applies to a Secure B2SHARE instance. B2SHARE categorizes metadata records by communities. A metadata record always belongs to one community. OAI-PMH verb ListSets can be used to list communities. OAI-PMH verb ListRecords can be used to list all records of a B2SHARE instance and using setSpec modifier with ListRecords verb, one can list only records that belong to a certain community. A response to ListRecords displays metadata of metadata records. A single metadata record can be fetched with GetRecord verb.

[101] Open Archives Initiative Protocol for Metadata Harvesting

While metadata harvesting improves discoverability of the datasets, access to sensitive data has to be explicitly requested via means provided specific to the sensitive data infrastructure Secure B2SHARE instance is deployed to.

5.6 Secure B2SHARE adapter for Galaxy portal

Improving integration and interoperability between services has been one goal of this project. Integration between Secure B2SHARE and Galaxy portal[102] was seen as a useful demonstrator of this, so an adapter that allows users of Galaxy portal to fetch sensitive data (after receiving proper authorization) from a Secure B2SHARE instance for processing in Galaxy portal was designed.

Due to difficulties with data processing policies related to CSC's Secure B2SHARE instance (sensitive data that is once ingested into the infrastructure is not allowed to leave the infrastructure without anonymization), it was decided that the first implementation of this adapter would only support a regular B2SHARE instance, without support for fetching sensitive data. Further development would need to be done to support fetching data from a Secure B2SHARE instance, but it goes without saying, that if policies of the instance, or rather the infrastructure that Secure B2SHARE is realized on, prevent export of sensitive data, it cannot be exported to a 3rd party service. Interesting topic for further research would be to find out if using technologies that allow to perform computation using encrypted data without decrypting it first could be used to perform analysis using sensitive data on a 3rd party service without compromising sensitivity promises of the infrastructure that is storing the sensitive data.

Secure B2SHARE adapter for Galaxy portal utilizes Linkset definition from FAIR Signposting specification to fetch resources of a B2SHARE dataset and possibly other datasets and data related to the dataset. Linkset support for B2SHARE has been developed in task 4.3 of the DICE project.

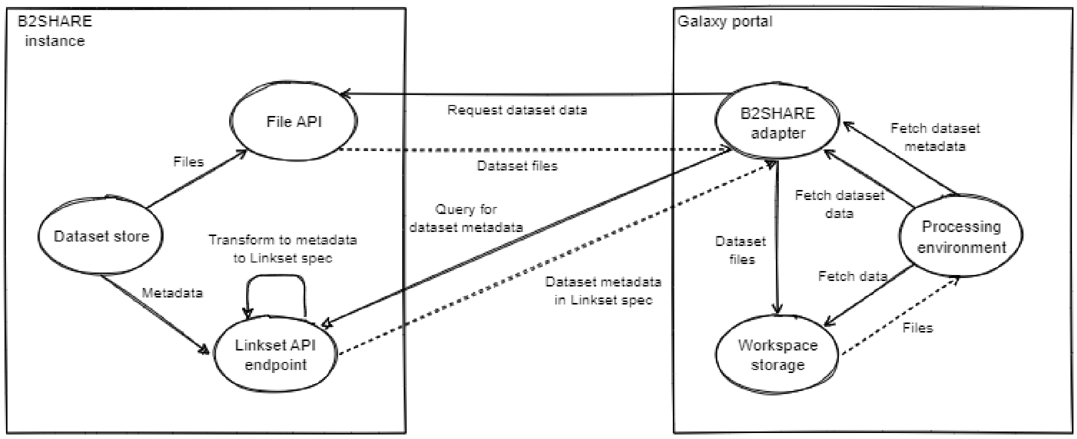


Figure 31. Secure B2SHARE adapter for Galaxy portal

The Adapter works by first fetching dataset metadata from a B2SHARE instance using B2SHARE Linkset API endpoint, based on user supplied B2SHARE dataset identifier. After this, the adapter parses the Linkset response and presents user with a list of files that are part of the requested dataset. User can select which files to import into Galaxy portal, after which adapter fetches the selected files via File API endpoint of B2SHARE.

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

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

Implementation of the adapter would be based on the already existing adapters available for Galaxy portal (https://github.com/galaxyproject/galaxy/tree/dev/lib/galaxy/files/sources).

Unfortunately, implementation of the adapter could not be completed during DICE project.

5.7 Conclusion and future work directions

This section described the work for integrating B2SHARE with sensitive data infrastructure services and secure data storage services. Secure B2SHARE has been introduced as a new concept and technology for supporting sharing of metadata about sensitive datasets which are being processed and/or temporarily stored in sensitive data processors, without publishing the actual data. Secure B2SHARE has been deployed on two sensitive data infrastructures: TSD (Norway), and CSC Sensitive Data Services (Finland). The TSD Secure B2SHARE instance is currently in production, while the CSC instance integration is still in progress.

In addition, an integration adaptor has been designed for integrating Secure B2SHARE with the Galaxy portal for allowing users of Galaxy portal to fetch sensitive data (after receiving proper authorization) from a Secure B2SHARE instance for processing in Galaxy portal. Due to the strict technological needs for implementing the adaptor in a secure and GDPR compliant way, the implementation of the adapter could not be completed during DICE project.

Future work includes:

- Implement full integration of Secure B2SHARE instance at CSC with SD Submit and SD Apply services of CSC’s Sensitive Data Services (SDS)
- Complete the implementation of the Secure B2SHARE adaptor for Galaxy.

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

6 References

[1] J. Nordling, "Developing the B2INST service for registering and persistently identifying instruments," doi: 10.5281/zenodo.6411786., Feb. 23, 2022.

[2] M. Stocker, "Persistent Identification of Instruments," *Data Science Journal,* Vols. vol. 19, no. 1, no. DOI: https://doi.org/10.5334/dsj-2020-018, p. 18ff, 2020.

[3] R. Krahl, Metadata Schema for the Persistent Identification of Instruments, doi: 10.15497/RDA00070, Mar. 30, 2022.

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

APPENDIX 1: Configurations and commands for the data transfer use case

1 Data transfer between B2SAFE and HPC using HTTP Rest API protocol

For the usage of the HTTP Rest API the following authorization information is needed:

- Authorization header needs to be provided on each request.
- Authorization is basic authentication username/password combination to the backend.

The username and password must be provided from the B2SAFE system that will be used.

The given credentials can then be automatically used on the computing platform by creating a configuration file (.netrc)

The configuration file must contain the following login information:

```
machine <b2safe.domain.org>
login <username>
password <password>


machine <eud-res01.domain.org>
login <username>
password <password>


machine <eud-res02.domain.org>
login <username>
password <password>
```

We are going to concentrate on using the HTTP-API interface with command line tool and library (cURL)[103] for transferring data with URLs.

List file or collections (recursively)

```
curl -n -i 'https://b2safe.domain.org:8443/collections/eudat.fi/
            home/username/TestData.txt
curl -n -i 'https://b2safe.domain.org:8443/collections/eudat.fi/
            home/username/collection1?recursive'
```

Create a collection

```
curl -n -i 'https://b2safe.domain.org:8443/collections/eudat.fi/
            home/username/collection1' -X PUT
```

Upload file

```
curl -n -i 'https://b2safe.domain.org:8443/objects/eudat.fi/
            home/ariyo/gustavelund/TestData.txt' -T TestData.txt
```

Delete file

```
curl -n -i 'https://b2safe.domain.org:8443/objects/eudat.fi/
            home/ariyo/TestData.txt' -X DELETE
```

[103] https://curl.se/

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

Upload file recursively

There is no direct command for recursive uploading of files to B2SAFE system.

We need to resort into scripting to help the recursive upload. A typical script (Curl_Script.sh) used in this use case is shown below.

```
Curl_Script.sh
     <code begins>
     #!/bin/bash
     #  Curl_Script.sh
     #
     #  Created by Chris Ariyo on 30.9.2021.
     # If no parameters given: Fail:
     #
     if [ $# -lt 2 ]
     then
             echo "Input the necessary parameters for upload!"
             echo "$0 <collection name> <source directory for files>"
             exit 1
     else
             # collection name below
             collection=$1
             # source directory
             srcdir=$2
             for fullfile in ${srcdir}/*
             do
                     basename "${fullfile}"
                     file="$(basename - ${fullfile})"
                     echo ""
                     curl -n -i https://b2safe.domain.org:8443/
                                collections/eudat.fi/home/username/
                                ${collection}/${file} -T ${file}
                     echo ""
             done
     echo "Done"
     fi
     <code ends>
```

2 Data transfer between Object Storage and HPC

For the usage of the rclone command line file transfer tool to access Allas a configuration file (rclone.conf) needs to be created with the necessary credentials for the connection to the object storage:

```
Rclone.config:

     [allas]
     type = swift
     env_auth = true

     [s3allas]
     type = s3
     provider = Other
     env_auth = false
     access_key_id = ...
     secret_access_key = ...
```

```
    endpoint = a3s.fi
    acl = private
```

In order to use Allas in Puhti or Mahti, first load the module allas:

```
module load allas
```

Allas access for a specific project can then be enabled:

```
allas-conf
allas-conf project_name
```

The allas-conf command prompts for your CSC password (the same that you use to login to CSC servers). It lists your Allas projects and asks you to define a project (if not already defined as an argument). allas-conf generates a rclone configuration file for the Allas service and authenticates the connection to the selected project.

You can only be connected to one Allas project at a time in one session. The project you are using in Allas does not need to match the project you are using in Puhti or Mahti, and you can switch to another project by running allas-conf again.

Authentication information is stored in the shell variables OS_AUTH_TOKEN and OS_STORAGE_URL and is valid for up to eight hours. However, you can refresh the authentication at any time by running allas-conf again. The environment variables are available only for that login session, so if you start another shell session, you need to authenticate again in there to access Allas.

Allas client software options for Puhti and Mahti:

- a-tools for basic use: (Swift, optionally S3) Quick and safe: a-commands
- Advanced functions with rclone: (Swift) Advanced tool: rclone
- A wide range of functionalities: (Swift) Swift client
- S3 client and persistent Allas connections: (S3) S3 client

This contains instructions for using Allas with Rclone in the Puhti and Mahti computing environments. Rclone provides a very powerful and versatile way to use Allas and other object storage services. It is able to use both the S3 and Swift protocols (and many others), but in the case of Allas, the Swift protocol is preferred. It is also the default option on the CSC servers.

Below are the Rclone syntax:

```
rclone subcommand options source:path dest:path
```

The most frequently used Rclone commands:

- rclone copy – Copy files from the source to the destination, skipping what has already been copied.
- rclone sync – Make the source and destination identical, modifying only the destination.
- rclone move – Move files from the source to the destination.
- rclone delete – Remove the contents of a path.
- rclone mkdir – Create the path if it does not already exist.
- rclone rmdir – Remove the path.
- rclone check – Check if the files in the source and destination match.
- rclone ls – List all objects in the path, including size and path.

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

- rclone lsd – List all directories/containers/buckets in the path.
- rclone lsl – List all objects in the path, including size, modification time and path.
- rclone lsf – List the objects using the virtual directory structure based on the object names.
- rclone cat – Concatenate files and send them to stdout.
- rclone copyto – Copy files from the source to the destination, skipping what has already been copied.
- rclone moveto – Move the file or directory from the source to the destination.
- rclone copyurl – Copy the URL's content to the destination without saving it in the tmp storage.

Create buckets and upload objects

In the case of Rclone, create a bucket:

```
rclone mkdir allas:2000620-raw-data
```

Upload a file using the command rclone copy:

```
rclone copy file.dat allas:2000620-raw-data/
```

List buckets and objects

List all the buckets belonging to a project:

```
rclone lsd allas:
     0 2019-06-06 14:43:40         0 2000620-raw-data
```

List the content of a bucket:

```
rclone ls allas:2000620-raw-data
     677972 file.dat
```

Download objects

Use the same rclone copy and rclone copyto commands to download a file:

```
rclone copy allas:2000620-raw-data/file.dat
```

If you include a destination parameter in the download command, Rclone creates a directory for the download:

```
rclone copy allas:2000620-raw-data/file.dat doh
```

Synchronizing a directory

For example, a folder named mydata has the following structure:

```
ls -R mydata
     mydata/:
```

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

```
        file1.txt  setA  setB
    mydata/setA:
        file2.txt
    mydata/setB:
        file3.txt  file4.txt
```

An example of using sync (note that the destination parameter requires the folder name (mydata)):

```
rclone sync mydata allas:2000620-raw-data/mydata
```

More information available at https://docs.csc.fi/data/Allas/

3 Using cURL to publish data on B2SHARE

We are using the training B2SHARE to test this use case, trng-b2share.eudat.eu.

We used the curl command[104] to test the publication of data in the training B2SHARE, trng-b2share.eudat.eu[105]

To use this tool, we need to export a few environment variables:

```
Export ACCESS_TOKEN=
'7O28DlvgCatQV0pkS6jLw947tbo123oztkU4dPw6fnqmJ8inOYAi7dYhF0d04'
export B2SHARE_HOST='trng-b2share.eudat.eu'
```

Manipulation and publication of research data object on B2SHARE

- List all communities

  -
    ```
    curl ”https://$B2SHARE_HOST/api/communities/”
    ```
- Get community schema

  -
    ```
    curl ”https://$B2SHARE_HOST/api/communities/
          $COMMUNITY_ID/schemas/last”
    ```
- List all records

  -
    ```
    curl ”https://$B2SHARE_HOST/api/records/”
    ```
- List records per community

  -
    ```
    curl ”https://$B2SHARE_HOST/api/records
          ?q=community:$COMMUNITY_ID”
    ```
- Search records

  -
    ```
    curl ”https://$B2SHARE_HOST/api/records/
          ?q=$QUERY_STRING&page=1&size=100&sort=mostrecent”
    ```
- Search drafts

  -
    ```
    curl "https://$B2SHARE_HOST/api/records/
          ?drafts&access_token=$ACCESS_TOKEN"
    ```
- Get specific record

  -
    ```
    curl ”https://$B2SHARE_HOST/api/records/
          47077e3c4b9f4852a40709e338ad4620”
    ```
- Create a draft record

  -
    ```
    curl -X POST -H "Content-Type:application/json"
    -d '{"titles":[{"title":"My dataset record"}],
    "creators":[{"creator_name": "John Smith"}, {"creator_name":
    "Jane Smith"}], "descriptions":[{"description": "A simple
    description",
                     "description_type": "Abstract"}],
    "community":"e9b9792e-79fb-4b07-b6b4-b9c2bd06d095",
    ```

[104] https://curl.se/

[105] https://trng-b2share.eudat.eu/

```
"open_access":true}'
https://$B2SHARE_HOST/api/records/
?access_token=$ACCESS_TOKEN
```

- Upload file into draft record

    -
      ```
      curl -X PUT -H 'Accept:application/json'
      -H 'Content-Type:application/octet-stream'
      --data-binary @$FILE_NAME
      ”https://$B2SHARE_HOST/api/files/
       $FILE_BUCKET_ID/
       $FILE_NAME?access_token=$ACCESS_TOKEN”
      ```
- Delete file from draft record

    -
      ```
      curl -X DELETE -H 'Accept:application/json'
      ”https://$B2SHARE_HOST/api/files/
       $FILE_BUCKET_ID/
       FileToBeRemoved.txt?access_token=$ACCESS_TOKEN”
      ```
- List files of record

    -
      ```
      curl ”https://$B2SHARE_HOST/api/files/
            $FILE_BUCKET_ID?access_token=$ACCESS_TOKEN”
      ```
- Update draft record metadata

    -
      ```
      curl -X PATCH
      -H 'Content-Type:application/json-patch+json'
      -d '[{"op": "add", "path":"/keywords",
            "value": ["keyword1", "keyword2"]}]'
      "https://$B2SHARE_HOST/api/records/
       $RECORD_ID/draft?access_token=$ACCESS_TOKEN”
      ```

    -
      ```
      curl -X PATCH
      -H 'Content-Type:application/json-patch+json'
      -d '[{"op": "replace", "path":"/titles/0/title",
            "value": ["The new title"]}]'
      "https://$B2SHARE_HOST/api/records/
       $RECORD_ID/draft?access_token=$ACCESS_TOKEN”
      ```
- Add externally referenced files to draft record

    -
      ```
      curl -X PATCH -H 'Accept:application/json-patch+json' -d '["op":
      "add", "path": "/external_pids", "value": "[{\"ePIC_PID\":
      \"prefix/suffix-of-file\", \"key\": \"filename\"},{\"ePIC_PID\":
      \"prefix/suffix-of-file-2\", \"key\": \"filename-2\"}]'
      “https://$B2SHARE_HOST/api/records/$RECORD_ID/draft?access_token=
      $ACCESS_TOKEN”
      ```
- Submit draft record for publication

    -
      ```
      curl -X PATCH
      -H 'Content-Type:application/json-patch+json'
      -d '[{"op": "add", "path":"/publication_state",
            "value": "submitted"}]' “https://$B2SHARE_HOST/api/records/
       $RECORD_ID/draft?access_token=$ACCESS_TOKEN”
      ```

More information available at https://documentation.eudat.eu/b2share/httpapi/

APPENDIX 2: Digital Preservation Service; Requirements and Specifications

The latest version of the Digital Preservation Service (DPS) specification can be found at: https://dans-labs.github.io/ddps-docs/

Snapshot date: 29th May 2023

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

Digital Preservation Service (DPS)

Living Document (WIP), 16 May 2023

This work is carried out in the context of the DICE project (Data Infrastructure Capacities for EOSC), funded by the EU's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207. It is developed by Task 4.3: "Long Term Preservation", as part of the working package 4 (WP4): "Integration with other services & platforms", and published in the joint WP4 Deliverable: "D4.3 Final integration with other services & platforms”.

This version is taken from (2023-05-22):
https://dans-labs.github.io/ddps-docs/

Introduction

Within the DICE project, task 4.3 created a Long-Term Preservation (LTP) template and accompanying guidance that data services and repositories can use to formulate their own LTP policy. The LTP policy should clarify to the users what is or can be guaranteed by the service, for how long and by whom. The policy template contains articles that distinguish between outsourced services and services hosted in-house.

This file documents the functional requirements and technical specifications for a service that implements such an LTP policy in which the archive is outsourced; the Digital Preservation Service (DPS).

It is meant as a guide on how to implement an example long-term preservation Service between a short to midterm data (web)repository service and a long-term preservation (LTP) archive.

The implementation adheres to documented community conventions for the use of W3C Linked Data Notifications (LDN) and Activity Streams 2 (AS2) to integrate repository systems with long-term (LTP) archives, in a distributed, resilient and web-native architecture.

The standards used, and the application profile documented here, are implementations of the generic patterns described by [Event Notifications in Value-Adding Networks], that details a profile for using Linked Data Notifications with ActivityStreams2 payloads in value-adding networks.

The Digital Preservation Service

Use cases

1. Single record archiving

A single dataset will be transferred to a LTP-Archive on demand, according to the repository LTP Policy, by an authenticated repository dataset owner.

2. (Auto) Archive Community records

All records within a repository community could be auto-archived, according to a community archival agreement.

Opt-out/in option could be offered to the dataset owner.

This use case is a variant of use case 1 and is not worked out in this document, but it could easily be derived from the previous use case.

Technical motivation

Asynchronous

The intended communication style among nodes is point-to-point, requiring no centralized hubs. Interactions among nodes (Service Nodes and Data Nodes) are necessarily asynchronous because certain notification patterns do not require a response ("fire and forget") and, in patterns that do, such as requesting an action, the time between a request and the announcement of the Action Result is unpredictable, as the recipient may complete tasks at its own pace. Pushing any data at any moment to an archive might cause resource problems at the server side. Therefore, it is push-oriented, with only the relevant nodes being updated about new information as it becomes available.

Lightweight

This proposal is lightweight. It does take relatively little compute resources to implement, both from the client as from the server side.

Multi-purpose

The LDN+AS2 notifications approach can also be used for other purposes with the same investment. For example the peer reviewing service COAR Notify, which is also strongly aligned with the [Event Notifications in Value-Adding Networks] Profile.

Open Web Standards

This specification is built using Open Web Standards only. Because of its obvious benefits like larger audience and community, forward compatibility with browsers and cost savings; no patents or licensing.

Open Web Standards

Below is a list of the open web standards that will be used for the core implementation of the Digital Preservation Service.

Linked Data Notification

W3C Linked Data Notification (LDN) is an HTTP-based notification (push) protocol. It will be used for repository/archive communication.

Activity Streams 2.0

W3C Activity Streams 2 (AS2) provides a foundational vocabulary for messaging about activities that involve web resources. A message profile will be used for LDN notifications payload exchanged by repositories and archives.

Within this documentation, the Linked Data Notifications (LDN) with ActivityStreams2 (AS2) payloads will be referred to as: 'LDN+AS2 notifications'.

The notification payloads should use JSON-LD as default syntax, but other RDF syntaxes may be used.

Signposting

Signposting is a REST/HATEOAS "follow your nose" (navigational) approach to make the scholarly web more friendly to machines; it leverages IETF RFCs and IANA-registered link relation types. Typed links (HTTP Link header and/or HTML) are used to allow machines to uniformly navigate scholarly artefacts irrespective of the repository they reside in. The FAIR Signposting Profile is a lightweight, yet powerful approach to increase the FAIRness of scholarly objects. It will be used by repositories as a means to allow archives to determine which web resources need to be retrieved in response to an on-demand archiving request.

Namespaces used

Within this document, the following namespace prefix bindings are used:

Prefix	Namespace	Name
as	https://www.w3.org/ns/activitystreams#	W3C ActivityStreams 2.0
ldp	http://www.w3.org/ns/ldp#	W3C Linked Data Platform (LDP) Vocabulary
sorg	https://schema.org/	Schema.org
ietf	http://www.iana.org/assignments/relation/	The Internet Engineering Task Force (IETF)

In our example LDN+AS2 notifications we use JSON-LD as syntax, in which we don't explicitly write the prefixes.

DPS Architecture

Main components

Web Repository

This is the short- to midterm repository system. In the DPS this repository system is used for depositing, storing and disseminating datasets / digital objects.

The **Web Repository** must support Signposting and have an LDN inbox for handling LDN+AS2 notifications.

In the Event Notifications in Value-Adding Networks specification, this is also referred to as the Data Node.

LTP Archive

A Long-term Preservation Archive where datasets/digital objects can be preserved and archived. The archive must provide a deposit endpoint. Depending on your system, this can either be internal or external. For example a REST API or SWORD endpoint.

Archival Bot

Middleware server that takes care of the LDN+AS2 notifications communication between the **Web repository** and the **LTP Archive**.

It must support Signposting and have an LDN inbox.

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

This middleware component can be implemented in several ways, depending on your architecture. It might be part of the LTP Archive application itself (built-in) or as a plug-in (internal). It may also run as a micro-service that interacts with the Archive (external).

The external variant can also be extended with 'rule engine' functionality. Depending on certain business rules, it can send the archival deposit to different endpoints in the required format. In the Event Notifications in Value-Adding Networks specification, this is also referred to as the Service Node.

High level Architecture Overview

The diagram below (Figure 32) displays the components of the preservation service and how they relate and interact with each other. The notation used is not a formal one and is intended to be self-explanatory.

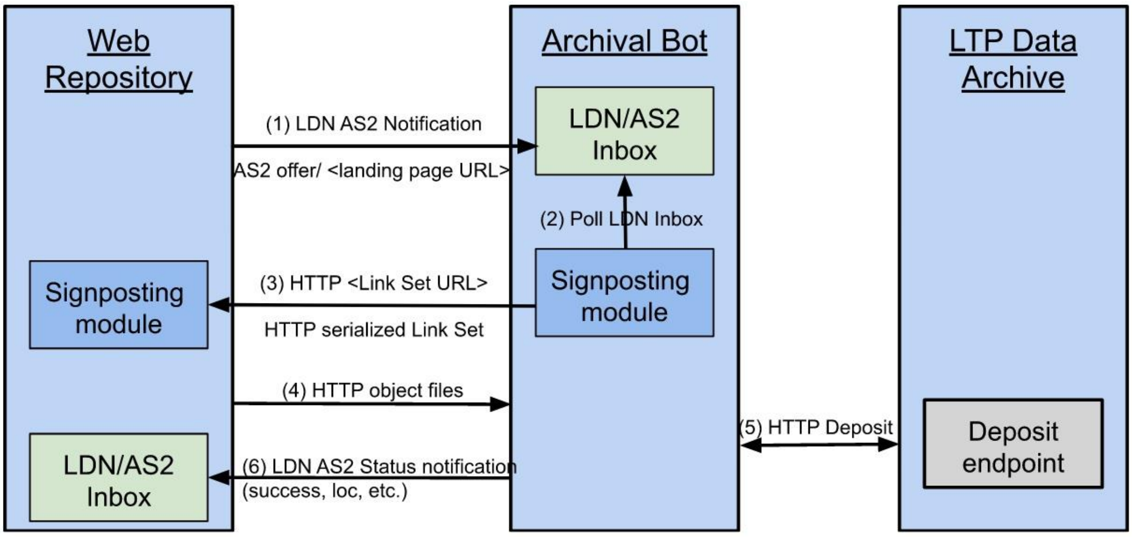


Figure 32. Context diagram of the Digital Preservation Service

1. An LTP request will be sent from a **Web Repository** landing page to the **Archival Bot**'s inbox, conveying the URL of the landing page of the dataset.
2. The **Archival Bot** will poll the LDN Inbox for such requests.
3. The **Archival Bot** visits the landing page URL, discovers a Link Set provided via Signposting, and retrieves it.
4. The **Archival Bot** parses the Link Set to obtain URLs for object files and metadata, and then retrieves those.
5. The **Archival Bot** deposits an archival package containing metadata and object files to the **LTP Archive**.
6. The **Archival Bot** reports back on the status of the deposit.

Sequence Diagram Use Case #1

The sequence diagram (Figure 33) below shows the interactions involved in use case #1. The numbers in the diagram correspond to the numbers given in the description of the interactions below.

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

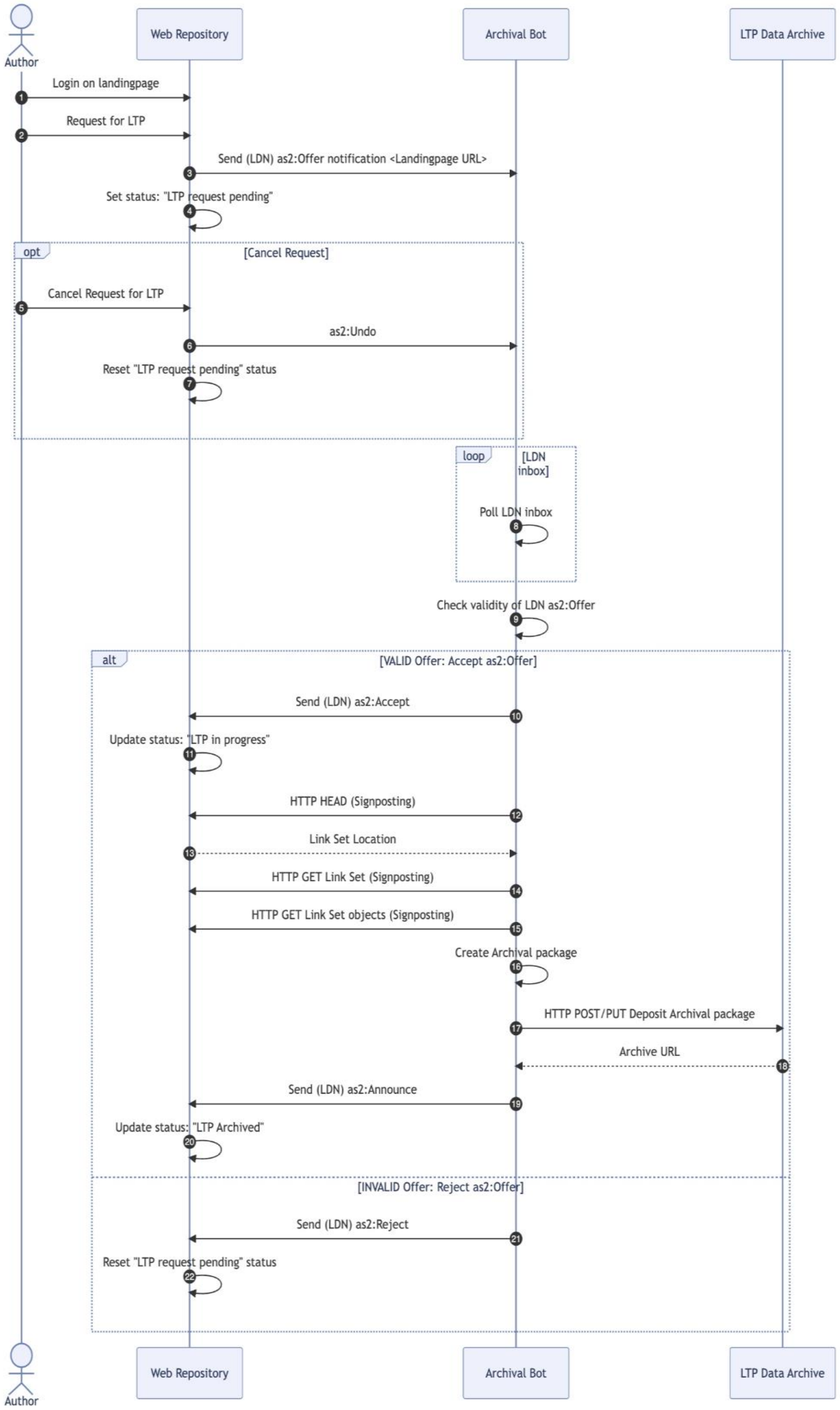

Author
Web Repository
Archival Bot
LTP Data Archive
Login on landingpage
Request for LTP
Send (LDN) as2:Offer notification <Landingpage URL>
Set status: "LTP request pending"
opt
[Cancel Request]
Cancel Request for LTP
as2:Undo
Reset "LTP request pending" status
loop
[LDN inbox]
Poll LDN inbox
Check validity of LDN as2:Offer
alt
[VALID Offer: Accept as2:Offer]
Send (LDN) as2:Accept
Update status: "LTP in progress"
HTTP HEAD (Signposting)
Link Set Location
HTTP GET Link Set (Signposting)
HTTP GET Link Set objects (Signposting)
Create Archival package
HTTP POST/PUT Deposit Archival package
Archive URL
Send (LDN) as2:Announce
Update status: "LTP Archived"
[INVALID Offer: Reject as2:Offer]
Send (LDN) as2:Reject
Reset "LTP request pending" status
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Web Repository
Archival Bot
LTP Data Archive
Author

Figure 33. Sequence diagram of the Digital Preservation Service

LDN+AS2 notification payloads

The LDN payloads that are involved in the sequence diagram above, have been composed of the specifications in both Event Notifications in Value-Adding Networks and COAR Notify.

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

For a profound understanding of this application profile one should also look into these specifications. Prior to each example payload, the requirements of the properties used are listed in a table. The JSON-LD properties @id and @type (mapped to 'id' and 'type' in the notification by the @context property) represent the mandatory identifier for the activity and the activity type respectively.

The activity identifier (@id) is distinct from the notification identifier, which is the URI minted by the LDN Receiver when the LDN+AS2 notification that describes the activity is received in its LDN Inbox. Notice that all payloads include the COAR Notify context file (@context). This context file defines commonly used namespaces in this profile as listed in table 1, and also includes the COAR Notify vocabulary. The activitystreams context file also includes namespace prefixes used in the examples.

1. The **Author** logs in into the landing page of the dataset that it owns on the **Web Repository** (authorized user).
2. The authorized **Author** presses an archival-button, to start a Long-Term Preservation (LTP) request for this dataset on the landing page.
3. The **Web Repository** (Data Node) sends a Linked Data Notification (LDN) Offer `as:Offer` to the LDN inbox `ldp:inbox` of the **Archival Bot** (Service Node). The payload MUST hold the landing page URL `sorg:AboutPage` of the scholarly object/dataset. This is by convention in both Event Notifications and COAR Notify. Also, providing the PID as a `ietf:cite-as` relation, together with the landing page URL is mandated by both specifications. This way all the associated resources can be found and retrieved by using FAIR Signposting discovery methods (12).

as:Offer (3):

In this example, the request for an LTP archival request is initiated by the authorized author, 'Some Author' (identified in the payload as the actor). The origin identifies the system that sends the message on behalf of the actor.

From the Event Notifications specification it is mandatory to supply at least one AS2 core type, which is `as:Document`.

Requirements	Properties
Required	`@id`, `@type`, `as:actor`, `as:object`
Optional	`as:origin`, `as:target`

```
{
    "@context": [
        "https://www.w3.org/ns/activitystreams",
        "https://purl.org/coar/notify"
    ],
    "id": "urn:uuid:78a30582-ef0f-11ed-a05b-0242ac120003",
    "type": "Offer",
    "actor": {
        "id": "https://orcid.org/0000-0003-4405-7546",
        "name": "Some Author",
        "type": "Person"
    },
    "origin": {
        "id": "https://b2share.eudat.eu",
        "name": "EUDAT B2SHARE Web Repository",
        "inbox": "https://b2share.eudat.eu/inbox/",
        "type": "Service"
    },
    "object": {
        "id":
"https://b2share.eudat.eu/records/1c42a67a73e9424b8192ba65c81077e1",
        "ietf:cite-as":
"https://doi.org/10.23728/b2share.1c42a67a73e9424b8192ba65c81077e1",
        "type": [
            "sorg:AboutPage",
            "sorg:Dataset",
            "Document"
        ]
    },
    "target": {
        "id": "https://archivalbot.data-stations.nl/",
        "inbox": "https://archivalbot.data-stations.nl/inbox/",
        "name": "DANS Archival Bot",
        "type": "Service"
    }
}
```

4. The **Web Repository** flags the status on the landing page of the scholarly object/dataset to "LTP Request Pending".
5. Option: The **Author** cancels the LTP request, by pressing the "Cancel LTP" button on the landing page.
6. Option: In response to action (5), the **Web Repository** sends a Linked Data Notification (LDN) Undo `as:Undo` to the LDN inbox `ldp:inbox` of the **Archival Bot** (Service Node).
7. Option: In response to action (5), the **Web Repository** cleares the LTP status/flag on the landing page, on behalf of the **Author**.

as:Undo (6):

In this payload, the as:object holds the original as:Offer message that needs to be undone.

Requirements	**Properties**
Required	`@id`, `@type`, `as:actor`, `as:object`
Optional	`as:origin`, `as:target`

```
{
    "@context": [
        "https://www.w3.org/ns/activitystreams",
        "https://purl.org/coar/notify"
    ],
    "id": "urn:uuid:82eaace6-3bf3-4b8b-b168-15d1a46fb668",
    "type": "Undo",
    "actor": {
        "id": "https://orcid.org/0000-0003-4405-7546",
        "name": "Some Author",
        "type": "Person"
    },
    "origin": {
        "id": "https://b2share.eudat.eu",
        "name": "EUDAT B2SHARE Web Repository",
        "inbox": "https://b2share.eudat.eu/inbox/",
        "type": "Service"
    },
    "object": {
        "id": "urn:uuid:78a30582-ef0f-11ed-a05b-0242ac120003",
        "type": "Offer",
        "actor": {
            "id": "https://orcid.org/0000-0003-4405-7546",
            "name": "Some Author",
            "type": "Person"
        },
        "origin": {
            "id": "https://b2share.eudat.eu",
            "name": "EUDAT B2SHARE Web Repository",
            "inbox": "https://b2share.eudat.eu/inbox/",
            "type": "Service"
        },
        "object": {
            "id":
"https://b2share.eudat.eu/records/1c42a67a73e9424b8192ba65c81077e1",
            "ietf:cite-as":
"https://doi.org/10.23728/b2share.1c42a67a73e9424b8192ba65c81077e1",
            "type": [
                "sorg:AboutPage",
                "sorg:Dataset",
                "Document"
            ]
        },
        "target": {
            "id": "https://archivalbot.data-stations.nl/",
            "inbox": "https://archivalbot.data-stations.nl/inbox/",
            "name": "DANS Archival Bot",
            "type": "Service"
        }
    },
    "target": {
        "id": "https://archivalbot.data-stations.nl/",
        "inbox": "https://archivalbot.data-stations.nl/inbox/",
        "name": "DANS Archival Bot",
        "type": "Service"
    }
}
```

8. The `as:Offer` in the LDN `ldp:inbox` of the **LTP Data Archive** is picked up by the **Archival Bot**
9. The Archival Bot checks the validity of the as:Offer. It checks against known rules, like registered domain, RDF format/SHACL validation, etc. It will either accept (10) or reject (19) the `as:Offer`.
10. The **Archival Bot** sends an `as:Accept` notification to the LDN `ldp:inbox` of the **Web Repository**.

***as:Accept (10)*:**

This payload, the `as:object` holds the original `as:Offer` message that is accepted and has a reference to the URI of the initial request (`as:inReplyTo`)

Requirements	**Properties**
Required	`@id`, `@type`, `as:actor`, `as:object`
Recommended	`as:context`, `as:inReplyTo`
Optional	`as:origin`, `as:target`

```
{
    "@context": [
        "https://www.w3.org/ns/activitystreams",
        "https://purl.org/coar/notify"
    ],
    "id": "urn:uuid:0cd58f07-69aa-4dd3-bc19-7b7de0f3550a",
    "type": "Accept",
    "actor": {
        "id": "https://archivalbot.data-stations.nl/",
        "inbox": "https://archivalbot.data-stations.nl/inbox/",
        "name": "DANS Archival Bot",
        "type": "Service"
    },
    "inReplyTo": "urn:uuid:78a30582-ef0f-11ed-a05b-0242ac120003",
    "context":
"https://b2share.eudat.eu/records/1c42a67a73e9424b8192ba65c81077e1",
    "object": {
        "id": "urn:uuid:78a30582-ef0f-11ed-a05b-0242ac120003",
        "type": "Offer",
        "actor": {
            "id": "https://orcid.org/0000-0003-4405-7546",
            "name": "Some Author",
            "type": "Person"
        },
        "origin": {
            "id": "https://b2share.eudat.eu",
            "name": "EUDAT B2SHARE Web Repository",
            "inbox": "https://b2share.eudat.eu/inbox/",
            "type": "Service"
        },
        "object": {
            "id":
"https://b2share.eudat.eu/records/1c42a67a73e9424b8192ba65c81077e1",
            "ietf:cite-as":
"https://doi.org/10.23728/b2share.1c42a67a73e9424b8192ba65c81077e1",
            "type": [
                "sorg:AboutPage",
                "sorg:Dataset",
                "Document"
            ]
        },
        "target": {
            "id": "https://archivalbot.data-stations.nl/",
            "inbox": "https://archivalbot.data-stations.nl/inbox/",
            "name": "DANS Archival Bot",
            "type": "Service"
        }
    },
    "target": {
        "id": "https://b2share.eudat.eu",
        "name": "EUDAT B2SHARE Web Repository",
        "inbox": "https://b2share.eudat.eu/inbox/",
        "type": "Service"
    }
}
```

11. The **Web Repository** updates the status of the landing page from "Request Pending" to "Long-Term Archiving in progress".
12. The **Archival Bot** will retrieve the Link Set URL from the **Web Repository** landing page, by either HTTP header or HTML of the landing page URL, as is described by Signposting.
13. The Link Set location is returned by the **Web Repository**.
14. The **Archival Bot** retrieves the serialized Link Set from the **Web Repository**.
15. The **Archival Bot** retrieves the content resources from the **Web Repository** that are listed in the Link Set.
16. The **Archival Bot** creates the archival package with all content resources.
17. The **Archival Bot** deposits the archival package to the **LTP Data Archive** deposit endpoint.

18. The **LTP Data Archive** responds with the Archive URL to the **Archival Bot**.
19. **Archival Bot** sends an `as:Announce` notification to the LDN `ldp:inbox` of the **Web Repository** to inform about the creation of the archive artifact and its URI.

as:Announce (19):

In this payload, the `as:object` announces the creation of the archived artefact, by indicating a "memento" link relation between the landing page and the archived artefact.

Requirements	Properties
Required	`@id`, `@type`, `as:actor`, `as:object`, `as:inReplyTo`
Recommended	`as:context`
Optional	`as:origin`, `as:target`

```
{
    "@context": [
        "https://www.w3.org/ns/activitystreams",
        "https://purl.org/coar/notify"
    ],
    "id": "urn:uuid:4a6b8761-3365-4379-a2cf-1fb012f1c2d8",
    "type": "Announce",
    "actor": {
        "id": "https://archivalbot.data-stations.nl/",
        "inbox": "https://archivalbot.data-stations.nl/inbox/",
        "name": "DANS Archival Bot",
        "type": "Service"
    },
    "inReplyTo": "urn:uuid:78a30582-ef0f-11ed-a05b-0242ac120003",
    "context":
"https://b2share.eudat.eu/records/1c42a67a73e9424b8192ba65c81077e1",
    "object": {
        "id": "urn:uuid:CF21A499-1BDD-4B59-984A-FC94CF6FBA86",
        "type": "Relationship",
        "subject":
"https://b2share.eudat.eu/records/1c42a67a73e9424b8192ba65c81077e1",
        "relationship": "http://www.iana.org/assignments/relation/memento",
        "object":    "https://www.persistent-identifier.nl/urn:nbn:nl:ui:13-wk-
epni"
    },
    "target": {
        "id": "https://b2share.eudat.eu",
        "name": "EUDAT B2SHARE Web Repository",
        "inbox": "https://b2share.eudat.eu/inbox/",
        "type": "Service"
    }
}
```

20. The **Web Repository** updates the status of the landing page from 'Long-Term Archiving in progress' to 'Long-Term Archived', including a link to the archived artefact.
21. **Archival Bot** sends an `as:Reject` notification to the LDN `ldp:inbox` of the **Web Repository**.

as:Reject (21):

This payload, the `as:object` holds the original `as:Offer` message that is rejected by the Archival Bot and has a reference to the URI of the initial request (`as:inReplyTo`)

Requirements	Properties
Required	`@id`, `@type`, `as:actor`, `as:object`
Recommended	`as:context`, `as:inReplyTo`
Optional	`as:origin`, `as:target`

```
{
    "@context": [
        "https://www.w3.org/ns/activitystreams",
        "https://purl.org/coar/notify"
    ],
    "id": "urn:uuid:2a74a606-42c8-4e26-88b9-0f14e7b79d82",
    "type": "Reject",
    "actor": {
        "id": "https://archivalbot.data-stations.nl/",
        "inbox": "https://archivalbot.data-stations.nl/inbox/",
        "name": "DANS Archival Bot",
        "type": "Service"
    },
    "inReplyTo": "urn:uuid:78a30582-ef0f-11ed-a05b-0242ac120003",
    "context":
"https://b2share.eudat.eu/records/1c42a67a73e9424b8192ba65c81077e1",
    "object": {
        "id": "urn:uuid:78a30582-ef0f-11ed-a05b-0242ac120003",
        "type": "Offer",
        "actor": {
            "id": "https://orcid.org/0000-0003-4405-7546",
            "name": "Some Author",
            "type": "Person"
        },
        "origin": {
            "id": "https://b2share.eudat.eu",
            "name": "EUDAT B2SHARE Web Repository",
            "inbox": "https://b2share.eudat.eu/inbox/",
            "type": "Service"
        },
        "object": {
            "id":
"https://b2share.eudat.eu/records/1c42a67a73e9424b8192ba65c81077e1",
            "ietf:cite-as":
"https://doi.org/10.23728/b2share.1c42a67a73e9424b8192ba65c81077e1",
            "type": [
                "sorg:AboutPage",
                "sorg:Dataset",
                "Document"
            ]
        },
        "target": {
            "id": "https://archivalbot.data-stations.nl/",
            "inbox": "https://archivalbot.data-stations.nl/inbox/",
            "name": "DANS Archival Bot",
            "type": "Service"
        }
    },
        "id": "https://b2share.eudat.eu",
        "name": "EUDAT B2SHARE Web Repository",
        "inbox": "https://b2share.eudat.eu/inbox/",
        "type": "Service"
}
```

22. The **Web Repository** cleares the LTP status on the landing page, because the `as:Offer` was not accepted by the **Archival Bot** because the business rules in (9) were not met.

Acknowledgements

Aligned Projects

- Event Notifications in Value-Adding Networks (including feedback on this specification).
- COAR Notify Protocol

Data Infrastructure Capacities for EOSC

The specifications for long-term-archiving notifications is carried out in the context of the DICE project (Data Infrastructure Capacities for EOSC), funded by the EU's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

It is developed by Task 4.3: "Long Term Preservation", as part of the working package 4 (WP4): "Integration with other services & platforms", and published in the joint WP4 Deliverable: "D4.3 Final integration with other services & platforms".