DICE

Data Infrastructure Capacity for EOSC

Deliverable D4.2

Pilots for the integration with other services & platforms

Responsible Partner:	Forschungszentrum Jülich
Status-Version:	Final – v2.0
Date:	13/04/2022
Distribution level (CO, PU):	Public

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

Title of Deliverable:	Pilots for the integration with other services & platforms
Due Date of Delivery to the EC	31.03.2022
Actual Date of Delivery to the EC	13.04.2022

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 Sven Bingert, GWDG (Section 3.2) Göksenin Cakir, GWDG (Section 3.1.3.1 and 3.1.3.2) T4.3 Long Term Preservation policy: Doorn, P. – KNAW - DANS Sanden, M. van de - SURF Kraaikamp, E. - KNAW-DANS Piggelen, H. van - SURF Steinhoff, W. - KNAW-DANS Indarto, E. - KNAW-DANS T4.4 Sensitive Data: Abdulrahman Azab, SIGMA
Reviewer(s):	Tonello, N. – BSC Testi, D. - CINECA
Recommended/mandatory readers:	WP5 Integration with community platforms

Abstract:	This deliverable includes the pilot use cases for the integration of data services with computing platforms of T4.1 the integration of the integrity check for PIDs of T4.2 the long-term preservation policies for B2SHARE and B2SAFE of T4.3 and the sensitive data risk analysis of T4.4
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

Document Description

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	08.12.2021	First draft version	FZJ
V0.11	25.01.2022	Added chapters for each task	FZJ
v0.2	23.02.2022	Sub deliverable T4.3	DANS
v0.3	18.02.2022	Rearranged T4.1 part	CSC
v0.4	23.02.2022	B2DROP use case	FZJ
v0.5	25.02.2022	Added all T4.1 parts	CSC
v0.6	28.02.2022	Finalised T4.1 parts	CSC
V0.7	04.03.2022	Sensitive data risk analysis	SIGMA
V0.8	08.03.2022	First version of T4.2	GWDG
V0.9	10.03.2022	First full draft with all 4 task contributions	FZJ
V1.0	15.03.2022	Finalised T4.2 contribution Added Executive Summary and Conclusions	GWDG FZJ
V1.1	21.03.2022	Finalised T4.4 contribution	SIGMA
V1.2	29.03.2022	Update taking into account the DICE internal reviews	FZJ, CSC, DANS, GWDG
V1.3	08.04.2022	Update T4.4 contribution taking into account the DICE internal reviews	SIGMA
V2.0	13.04.2022	Final version ready for submission	PMT

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.

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

API	Application Programming Interface
AUP	Acceptable Use Policy
BIT	Business Intelligence Team
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
DDPS	DICE Digital Preservation Service
DKRZ	Deutsches Klimarechenzentrum GmbH
DoW	Description of Work
DPA	Data Processing Agreement
DPS	Data Privacy Statement
DTR	Data Type Registry
EC	European Commission
EOSC	European Open Science Cloud
EOSC TF LT	EOSC - Task Force - Long Term Data Preservation
ePIC	Persistent Identifier Consortium for eResearch
ETHZ	Eidgenössische Technische Hochschule Zürich
EU	European Union
EUDAT	European Data Infrastructure
EUDAT CDI	EUDAT Collaborative Data Infrastructure
EUDAT ltd	EUDAT ltd
FZJ	Forschungszentrum Juelich Gmbh
GA	Grant Agreement to the project
GDPR	General Data Protection Regulation
GHR	Global Handle Registry
GRNET	National Infrastructures for research and technology
GWDG	Gesellschaft für Wissenschaftliche Datenverarbeitung mbh Göttingen
INFN	Istituto Nazionale di Fisica Nucleare
IT4I	Vysoka Skola Banska - Technicka Univerzita Ostrava
KIT	Karlsruhe Institut für Technologie
KNAW-DANS	Koninklijke Nederlandse Akademie van Wetenschappen
KPI	Key Performance Indicator
LHS	Local Handle System
LTP	Long-term preservation
MPG	Max Planck Gesellschaft zur Foerderung der Wissenschaften e.V.
OLA	Operational Level Agreement
PID	Persistent Identifier
SIGMA	SIGMA2
SLA	Service Level Agreement
SMF	Service Management Framework
SNIC	Uppsala Universitet
SURF	SURFsara BV
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 4.2 "Pilots for the integration with other services & platforms" of the DICE projects compiles the contributions of the four tasks of WP4:

- Task 4.1 "Compute and Analysis" reports on the status of piloting the integration of data services with computing platforms
- Task 4.2 "Discovery and Referencing" addresses the integrity of PID infrastructure and of PID metadata and describes the TypeAPI service, which will improve the use of datatypes in the PID landscape
- Task 4.3 "Long Term Preservation of Data" delivers a Long-Term Preservation (LTP) Policy Template for the EUDAT Services B2SHARE and B2SAFE, which is generic and can be used by other repositories to compose their LTP policies, too
- Task 4.4. "Sensitive Data" informs on the "Sensitive data risk analysis"

The next deliverable of WP4 is due in project month 30 and will have the final contributions from all four tasks. Its name is D4.3 "Final integration with other services & platforms" and it will inform about the final integration of data services with computing platforms (T4.1), the integration of PID Graph resources in B2FIND (T4.2), the implementation of the LTP policy for B2SHARE in one CTS certified archive (T4.3), and the enabling of sensitive data workflow by adapting standard interoperability frameworks to connect the endpoints (T4.4).

1 Introduction

D4.2 “Pilots for the integration with other services & platforms” is the second of three deliverables of the WP4 “Integration with other services & platforms” of the DICE project.

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

For task 4.1 “Compute and Analysis” the contribution in chapter 2 “Pilot use cases for the integration of data services with computing platforms” is the follow-on of the “technical report on the integration of B2-services and object storage service” in D4.1 “Planning for the integration with other services & platforms” from project month 9. Chapter 2 describes pilots for enabling analysis, data replication, and data publication of computing platforms through the integration of data services with these platforms. Due to the nature of the piloting activity, the chapter 2 is very technical, including example commands and configuration files for the enabling of data services on the computing platforms.

Task 4.2 “Discovery and Referencing” addressed two key aspects of PIDs’ integrity in chapter 3 “Integration of the integrity check for PIDs”:

(1) the integrity of the PID infrastructure, and

(2) the integrity of PID metadata.

The (1) is the basis of proper PID resolution, while (2) discusses the content of PID records and the usage of datatypes. We present the Prefix Information Service, which enables a transparent view about the integrity of PIDs for all DICE offered B2HANDLE services and we describe the TypeAPI service, which is a scalable solution for improving the use of datatypes in the PID landscape. This contribution supports a unified view on the status of the integrity of the PIDs, which is important for both PID service providers, as well as service users.

Task 4.3 “Long Term Preservation of Data” delivered a Long-Term Preservation (LTP) Policies Template for the EUDAT Services B2SHARE and B2SAFE. The template is so generic, that it can be used by a wide range of repositories and policy-based data archives to compose their LTP policies. The contribution of task 4.3 in comprises in chapter 4 the introduction to the LTP policy template and guidance on how to apply it to a repository, a short overview of related work, an introduction to cost modelling for LTPs, followed by an outlook to technical work in task 4.3. The LTP policy template itself is the appendix 1 of this deliverable, followed by 6 related appendices: a summary of the OAIS reference model in appendix 2, an overview of monitoring processes for repositories in appendix 3, an overview of available licenses for digital assets in appendix 4, a list of legislations and laws in appendix 5, a differentiation of some often used terms in appendix 6, and last but not least a comparison between the LTP policies of B2SHARE and B2SAFE in appendix 7.

Task 4.4 “Sensitive Data” informs on the “Sensitive data risk analysis” in chapter 5. This is the second report form task 4.4, following the “design of the sensitive data workflow” in deliverable 4.1.

Chapter 5 gives an introduction of measures for securing the processing of sensitive data. It is followed by an introduction to processes and services for processing sensitive data, and the risk analysis for the deployment of the Laniakea cloud platform [1].

2 Pilot use cases for the integration of data services with computing platforms

2.1 Introduction

The aim of task T4.1 “Compute and Analysis” is to integrate data services with computing platforms to enable analysis, data replication, and data publication for computing and cloud computing platforms. This is done by using B2DROP [2] service as a tool to maintain “recipes” for analysis, B2SAFE [3] as a tool to register results, and B2SHARE [4] as a tool to publish datasets.

In this deliverable, this task will express the pilots tried on integration of EUDAT data services and Computing platforms:

- B2DROP
    - Ensure that small data (batch queue scripts etc. similar small objects) can be read from B2DROP to computing environment and that small data can be written back to B2DROP
- B2SAFE
    - Data transfer between B2SAFE and computing platforms, and also data transfer between computing platforms and object storage systems
- B2SHARE
    - Transfer data from computing platforms to B2SHARE for publishing.

2.2 B2DROP use case

B2DROP is an easy-to-use, 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. 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 “Share links”

Share 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 Share 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. Share links do not require additional authentication, and can be used from anywhere where a 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 Share 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.

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

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

2.2.2 Using WebDAV

WebDAV [5] is a much more capable protocol, 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 from all HPC nodes with internet access, typically, these will be the login nodes or dedicated data transfer nodes.

The B2DROP WebDAV interface 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 possible, too.

2.2.3 Challenges in using WebDAV for mounting on HPC

Currently, it is not possible to use B2DROP on HPC systems with WebDAV for mounting B2DROP as a file system, because an entry in */etc/fstab* is required for each WebDAV share to use automatic WebDAV mounts without root permissions [6]. This will not be available on multi-user HPC systems, because you have to specify one mount point for the whole system.

As alternative solution, we have added this for the HDF-Cloud HDF-Cloud resource in Jupyter-JSC [7]. Therefore, users can use their B2DROP files on the HDF-Cloud resources. Users start their JupyterLab in a container, thus every resource can use the same mount path, because they do not share a whole file system and only have access to their own files. There are other solutions to work around these issues. The most promising solution was not updated since May 2020 [8]. Thus, we are considering other possible solutions like using Davix [9] and including WebDAV entries in the file system.

2.3 B2SAFE use case

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

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 regular basis or be triggered by actions like data ingest. The rules interact with external software components, which deliver functionalities such as PID registration. An iRODS zone contains an iCAT-enabled resource server ("iCAT server" for short), which uses a relational database to organize the content of the zone and to maintain iRODS metadata. IRODS zones can be connected to each other for replication or for redundant purposes.

The B2SAFE module offers also rules for integrity checks across iRODS zones [10], 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.

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

B2SAFE offers safe data replication across different data centres. 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 [11] protocol
2. Web Distributed Authoring and Versioning (WebDAV) [5] protocol
3. GridFTP [12] File Transfer Tool

2.3.1 Data transfer between B2SAFE and HPC

This use case will concentrate on using the HTTP-API available in B2SAFE to transfer data between B2SAFE and the HPC systems Mahti [13] and Puhti [14].

Using HTTP-API protocol

The use of this 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 the following requirements should be met:

- Authorization header need 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) which 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 in using the HTTP-API interface with curl.

List file or collections (recursively)

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

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

```
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
```

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 in Appendix 8.

2.3.2 Data transfer between Object Storage and HPC

Accessing CSC – IT Center for science object storage (Allas) in the CSC computing environments will be used for this use case. To transfer data between Allas and the HPC systems Mathi and Puhti we have the following possible tools [15]:

- a-tools for basic use: 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

We are going to concentrate on using rclone [16] command line file transfer tool on HPC to transfer data to and from the object storage system (Allas) in this use case.

Rclone is selected, because it is very common and available in most HPC systems and for different operating systems (OS).

Using rclone command line file transfer tool to access Allas

There is necessary configuration file (rclone.conf) that need 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 my 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.

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.

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.
- 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/:
          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
```

We successfully transfer data to and from Allas to the HPC systems (mahri, puhti).

More information about data transfers between the Allas object store and the HPC systems Mathi and Puhti is available at https://docs.csc.fi/data/Allas/

2.3.3 Better connection of B2SAFE to B2ACCESS

The aim of this test was to improve the usability of B2SAFE by a better integration with B2ACCESS. The possibility to do this relies on the availability of OpenID Connect (OIDC) [17] in the iRODS server, which be available only in the next iRODS version. We will wait with the testing of the integration for the availability of the OIDC plugin in iRODS, because this will tremendously ease and simplicity the integration of the services.

2.4 B2SHARE use case

B2SHARE is the EUDAT service for storing and publishing data sets. In addition to its web-based GUI, B2SHARE offers an HTTP REST API. 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 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. B2SHARE's access tokens follow the OAuth 2.0 standard.

2.4.1 API token generation from the B2SHARE web user interface

The web user interface in B2SHARE allows the user to create a new token and stored into a file, that the user must take care to keep in a safe place to be used later for the authentication.

2.4.2 A publication workflow

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

2.4.3 Using curl as a tool to publish data on B2SHARE

We were using the training B2SHARE to test this use case, trng-b2share.eudat.eu, but the use case will work on the production B2SHARE instance as well.

We used the curl command to test the publication of data in the training B2SHARE, trng-b2share.eudat.eu [18].

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 process and example for the use of this tool for publishing is available in Appendix 9.

- **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://eudat.eu/services/userdoc/b2share-http-rest-api

3 Integration of the integrity check for PIDs

This Chapter contains the deliverable of Task 4.2 of the DICE project, the “Integration of the integrity check for PIDs”. Persistent Identifiers (PIDs) are an important part of the research process and relevant for referencing and locating all kind of (digital) resources. The integrity and reliability of PIDs is achieved by various measures. This deliverable addresses two key aspects of PIDs’ integrity:

- Resolution of PIDs (integrity of the PID infrastructure)
- Content of PIDs and datatypes (integrity of PID metadata)

Both PID service providers, as well as service users need a unified view on the status of the integrity of the PIDs provided. Furthermore, a clear understanding of the service levels, which users can expect, is needed. This transparent view on PID integrity and provided service levels can be the basis of future certification of PID services.

Creating such a transparent view for the integrity of PIDs for all instances of the B2HANDLE service is the first achievement of the work described here. The second achievement we present is a solution for improving the usage of PID datatypes in the PID landscape.

This Section of the document is structured as follows: first the integrity of PID infrastructures is discussed and the Prefix Information Service is introduced, which is useful to get an overview of the integrity of the PID infrastructures. The second part of the Section discusses the standardisation of the metadata stored in the PID records. To make this development available for a broad audience we introduce the “TypeAPI” service, which is useful to discover and use PIDs type standards.

3.1 Integrity of the PID Infrastructure

For a PID infrastructure, transparency about the reliability of the identifiers is of particular importance. The integrity and reliability of PIDs is achieved by various measures taken both on technical, as well as on organizational levels. In the following, we describe these organizational and technical measures, as well as we give some insights into the technical details of the PID infrastructure of the DICE services.

3.1.1 Measures for the PID Infrastructure Integrity

The PID services offered by DICE members of the consortium are branded as B2HANDLE. The B2HANDLE services are generic PID services based on the Handle System [19]. The Handle System is a distributed system, it operates in a distributed manner and has two levels of hierarchy: the *local* handle services and the *global* (root) services.

Local handle services contain the identifier (also known as ‘PID’ or ‘handle’) records under a specific prefix. The global (root) service contains records that describe who controls which prefixes and how the local handle services can be reached. When a PID resolution request or PID maintenance request is issued, first, the global Handle infrastructure (particularly the Global Handle Registry) is contacted. This maps a PID resolution or PID maintenance requests to local services (Local Handle System, LHS). Then, the local service provides the requested information (resolution) or it makes this information available in the local service (maintenance).

As no single provider operates the overall service, the integrity and reliability of the overall service must be addressed on both the global, as well as on the local levels. All these measures together are essential to maintain the integrity of PIDs. We summarize these organizational and technical policies and procedures in in Table 1.

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

Table 1. Organizational and technical measures for the integrity of the PID infrastructure

Scope	Organizational	Technical
Local PID services	• Quality of Service and Policies (QoS&P) document of ePIC. • This describes general Service Level Agreements (SLA), quality of services, policies and workflows for PID services.	• Best practices for service operation. • Monitoring of services and accounting. • Verifying the availability of mirror servers.
Global PID infrastructure	• Multi-Primary Administrator (MPA) Service Agreement with the DONA Foundation [20]. • This credentials an MPA to administer and operate the Global Handle Registry (GHR) and to provide resolution services according to the Foundation Procedures.	• Service operations practices by “Multi-Primary Administrators’ Global Handle Registry (MPA GHR) Service - Operations Manual”. • Testing each MPA GHRs for consistency, reliability, and performance on an ongoing basis • Auditing all the MPA GHRs on daily basis

On the level of local PID services we mention that the B2HANDLE services are Local Handle Systems. All current B2HANDLE providers of DICE committed themselves to follow the technical and organizational rules and procedures of the Persistent Identifiers Consortium for eResearch (ePIC) [21]. The ePIC “Quality of Service and Policies” (QoS&P) [22] document describes general Service Level Agreements (SLA), quality of services, policies and workflows for the local Handle services, as provided by ePIC members or providers of ePIC. All B2HANDLE services currently offered by DICE partners should fulfill these requirements.

On the global level, the integrity of global Handle services is coordinated by the DONA Foundation. The Global Handle Registry is operated in the context of the MPA Service Agreement and in accordance with the Foundation Procedures. This Service Agreement and the Procedures specify that DONA checks the GHR for consistency, reliability, and performance on an ongoing basis. It also describes that DONA verifies the replication of all prefixes and audits all the GHR on a daily basis to ensure its overall integrity.

We continue with giving an overview about the PID services in the context of the DICE project, that several providers have set up based on the Handle System.

3.1.2 Integration of DICE offered B2HANDLE services with community platforms

The DICE partners offer state-of-the-art data management services as building blocks to store, find, access and process data in a consistent and persistent way.

DICE's resource provisioning is accompanied by enhancing the current service offering in order to fill the gaps still present to the support of the entire research data lifecycle. The solutions are provided to increase data reusability and data quality, as well as they support long term preservation and the management of sensitive data.

To help the integration of DICE offered data services with community platforms and engage research communities in the exploitation of DICE services, the European Commission has made available an innovative funding instrument known as Virtual Access (VA). The VA funding

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

mechanism [23] makes it possible for providers to be compensated fairly and equitably while transparently offering new users their services at no cost until the end of the project.

In the context of the DICE project, several B2HANDLE providers are making available PID services based on the Handle System. These providers are: Gesellschaft für wissenschaftliche Datenverarbeitung mbh Göttingen (GWDG), Greek Research and Technology Network (GRNET) and SURF. At the time of writing, they host 9 PID prefixes (the term ‘prefix’ is also known as ‘PID namespace’ or ‘naming authority’). We summarize all DICE VA instances Table 2.

Table 2. Integrated B2HANDLE instances (DICE Virtual Access)

B2HANDLE (Prefix)	Provider	Project / Community / Organization
21.11150	GWDG	Leibniz-Institut für Alternsforschung
21.11151	GWDG	HZB
21.12144	SURF	BSC
21.12145	SURF	EOSC-SYNERGY WP4 LAGO
21.12146	SURF	BSC
21.12149	SURF	IT41
21.12150	SURF	UCL Compbiomed
21.15107	GRNET	University of Belgrade
21.15108	GRNET	Gallo-Roman Museum

Unfortunately, the Handle System does not provide any easy way that would allow users to get more useful information about the prefixes, like which servers are responsible for a prefix, or how many mirrors are registered for a particular PID. We continue with describing a possible solution.

3.1.3 The Prefix Information Service

In this Subsection we introduce the Prefix Information Service, which can be used to get an overview of the integrity of the PID infrastructures.

Such an overview is important for both PID service providers, as well as for service users. In case of the service providers, they need a unified view on the status of the integrity of the PIDs provided and thus, all components and services of its PID infrastructure. From the end user point of view, a clear understanding of the service levels, which users can expect, is needed. This transparent view on PID integrity and the provided service levels could be a first step for certification of PID services in the future.

3.1.3.1 *Replication and PID infrastructure integrity*

In the previous Subsection we gave an overview about the B2HANDLE services provided by DICE partners. Each of those PID prefixes is assigned to one or more local handle services. The handle services can be primary (for maintenance) or mirror (for backup and/or additional resolution). A generic handle setup foresees a single primary. Since all current B2HANDLE providers of DICE committed themselves to follow the ePIC “Quality of Service and Policies” [22], they all set up a more robust, more reliable, and more complex PID replication procedure. Typically, one primary server stores the PID records and two additional mirror servers (provided by other ePIC partners) replicate the PID records from the primary server. As the mirrors have all PID records, they are transparently used for resolution. Thus, they allow redundancy for resolution, provide a full backup of all PIDs and could be used for disaster recovery purposes. For these purposes every service, which is responsible for a particular prefix, must be registered with the Global Handle Registry. This happens by adding a dedicated HS_SITE handle value (one for each service) to the

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

prefix. The HS_SITE value contains the technical description for handle clients that they know how to reach the local handle services (IP address, port, protocols, etc). This particular information registered in the Global Handle Registry and the current status of the local handle services must match. This is crucial for the integrity of the PID infrastructure.

To check the integrity of the PID infrastructure and the setup of the DICE B2HANDLE services provided via DICE, we designed the Prefix Information Service. It understands the HS_SITE technical descriptions and gives a descriptive overview, which is easily understandable to the users. Next, we describe the service architecture and how the service works.

3.1.3.2 *Architecture*

In order to assist the users in obtaining the prefix related information that they are looking for, we designed a web service with a simple user interface [24]. This allows users to easily and efficiently get such information about the prefixes, like which servers are responsible for a prefix, or how many servers are there or how many mirrors are registered for a particular PID.

The service contains a backend service and a frontend. The backend is a Python FastAPI application and is responsible to collect the information about the integrity of a prefix handle and its servers. Figure 1 depicts the architecture of the backend.

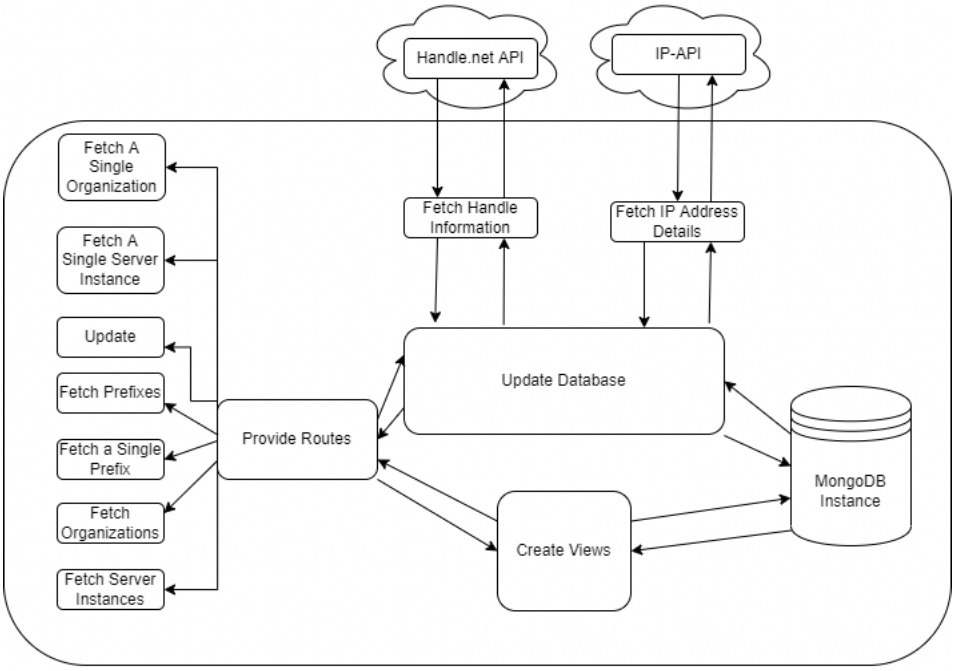


Figure 1. Architecture of the service backend

As a first step, the backend contacts the global registry (sends a GET request to its REST API) and if it was successful, it processes the prefix information. The backend service caches all necessary data in a MongoDB instance.

Afterwards, some statistical information of the prefix, namely, the number of IP addresses, the number of the subnets are calculated and added to the prefix information. This information is calculated on the fly and is stored in the database together with the prefix information.

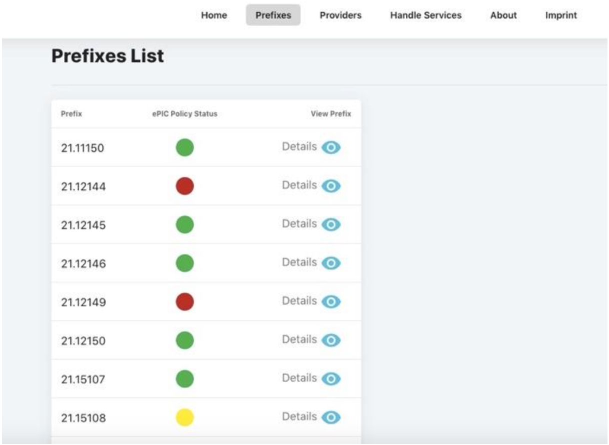
As a third step, more details about the service providers are collected, based on the servers' IP addresses. For this, an external service, the 'ip-api' [25] is contacted for each server and various details about the networks are fetched. These details include: Organization name, City, Country, Country Code, Continent, Continent Code, Region, Region code, Region Name, Internet Service Provider (ISP) name.

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 backend application provides routes that can be consumed by our front-end service or other external services. The major routes relevant to our application: Update, Fetch Prefixes / Fetch a Single Prefix, Fetch Providers / Fetch a Single Organization, Fetch Server Instances / Fetch A Single Server Instance. The names of these routes describe their functionality.

3.1.3.3 *Web User Interface*

The frontend is a JavaScript application and is written in VueJS. That way, it can be used for both desktop as well as mobile web applications. It offers a data-binding and data-driven model to handle HTML DOM. It also observes the changes in the UI elements and then makes calculations based on those changes to restructure the view and the UI elements, without the overhead of extra coding.



Prefixes List

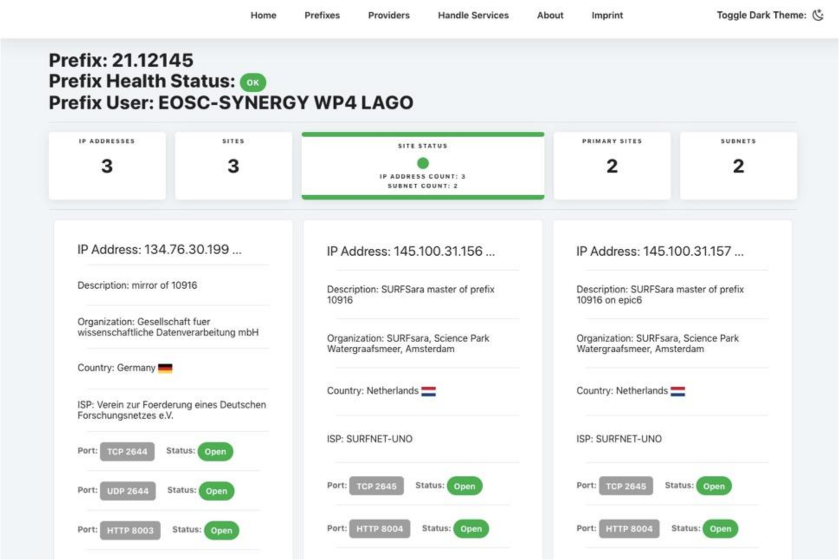
Prefix	ePIC Policy Status	View Prefix
21.11150	🟢	Details
21.12144	🔴	Details
21.12145	🟢	Details
21.12146	🟢	Details
21.12149	🔴	Details
21.12150	🟢	Details
21.15107	🟢	Details
21.15108	🟡	Details

Figure 2. Integrity of the DICE offered B2HANDLE services

Figure 2 shows a screenshot about the DICE B2HANDLE prefix instances and their status. The status is colour coded and based on some metrics calculated by our service. Such parameters that influence the reported status of a prefix can be for example, the number of servers (eg. IP addresses) and the number of subnets (eg. somehow independent internet connections for backup services). The following colour codes are possible for a prefix entry:

- Green: at least three servers (IP addresses) and at least two subnets
- Yellow: three servers (IP addresses) with one subnet only, or two servers only (IP addresses) with two subnets
- Red: one or two servers (IP addresses) with one subnet only
- Grey: It signifies the lack of any IP address or subnet information in the prefix output

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



Prefix: 21.12145
Prefix Health Status: OK
Prefix User: EOSC-SYNERGY WP4 LAGO

IP ADDRESSES	SITES	SITE STATUS	PRIMARY SITES	SUBNETS
3	3	● IP ADDRESS COUNT: 3 SUBNET COUNT: 2	2	2

IP Address: 134.76.30.199 ...

Description: mirror of 10916

Organization: Gesellschaft fuer wissenschaftliche Datenverarbeitung mbH

Country: Germany

ISP: Verein zur Foerderung eines Deutschen Forschungsnetzes e.V.

Port: TCP 2644 Status: Open

Port: UDP 2644 Status: Open

Port: HTTP 8003 Status: Open

IP Address: 145.100.31.156 ...

Description: SURFSara master of prefix 10916

Organization: SURFsara, Science Park Watergraafsmeer, Amsterdam

Country: Netherlands

ISP: SURFNET-UNO

Port: TCP 2645 Status: Open

Port: HTTP 8004 Status: Open

IP Address: 145.100.31.157 ...

Description: SURFSara master of prefix 10916 on epic6

Organization: SURFsara, Science Park Watergraafsmeer, Amsterdam

Country: Netherlands

ISP: SURFNET-UNO

Port: TCP 2645 Status: Open

Port: HTTP 8004 Status: Open

Figure 3. Prefix related details as shown by the 'Prefix' subpage

A prefix subpage displays prefix related details. Figure 3 shows such a prefix page. It provides information about the number of subnets, the number of IP addresses and various details related to the network connectivity of the handle service instance. For each service, the following information is displayed:

- Organization name (as fetched from the 'ip-api' service),
- Which ports are open (along with a button to verify if they are open),
- In which country and/or city the provider of the PID service is located,
- Which Internet Service Provider (ISP) provides the access to the IP address,
- Whether it serves a primary or a mirror site.

This way, the service provides a transparent view for each prefix and the calculated status of the integrity of the PID infrastructure behind it.

3.2 Integrity of PID Metadata

In this Subsection we address the second key aspect of Persistent Identifiers' integrity: the integrity of PID metadata. Although there are several standards for PID metadata, the Handle system [19] allows free choice of type. Types here are the keys of the key-value pairs that are stored in a handle record. The free choice of types (or datatypes) opens up new fields of application but carries certain risks. In a first step, scientific communities have to agree on common types. This is a time-consuming process that inhibits the actual work. In addition, different communities may define types that mean the same thing but are named differently. This leads to a proliferation of types in the scientific landscape. To prevent this and to support the scientific communities, an RDA Working Group "Data Type Registries" was founded by a group of interested researchers and stakeholders. This working group discussed the topic intensively and published a recommendation [26]. A first prototype of a data type registry, as a result of the recommendation, was put into operation in February 2014.

Currently, it has to be clarified how many of these registries are useful and necessary to serve the (European) scientific community. The ePIC consortium [21] is also currently running a data type registry [27]. In addition to the RDA working group PIDInst [28], there are currently other projects and use cases that have registered data types and use them to create PIDs (c.f. Figure 4)

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

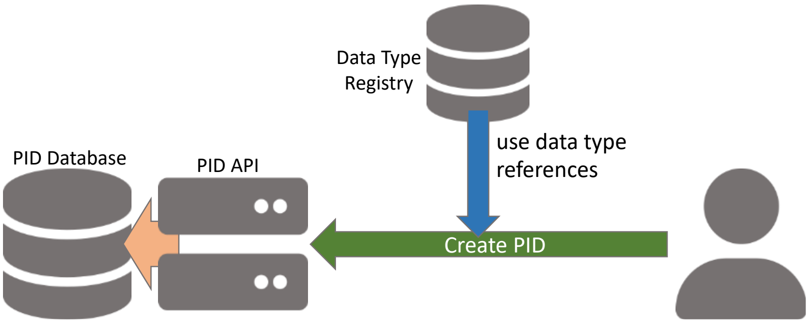
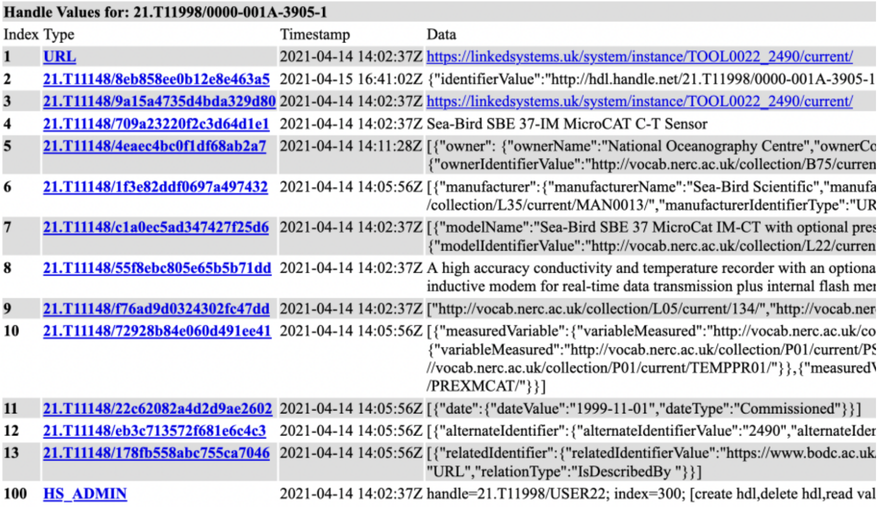


Figure 4. Simplified representation for the use of data types. The PID API can be for example the Handle API.

The exact use of registered data types has not yet been conclusively clarified. There are various approaches, especially for hierarchically structured data types. Hierarchical data types are data types that have a parent-child relationship to each other. For example, authors are a list of authors which in turn have a name and an identifier. The names then further consist of a first name and a last name. The possibility of using hierarchical data types allows the reuse of data types and the optimized structuring of the data in the PID. The registered data types are described in the registration with a text and can be defined by further restrictions (e.g., string or number). In addition, a Json validation scheme is stored for each data type, including the parent types.

Figure 5 shows an example of the use of data types. The otherwise usual plain names of types, such as URL or Author, are replaced by the references to the respective types. This use of data types increases the machine readability of PIDs. In automated processes, the type references can thus be resolved, and the information used to interpret the values. The values either expressed as a JSON object if the type is a derived (means hierarchical) type or a string otherwise (c.f. Figure 5).



Handle Values for: 21.T11998/0000-001A-3905-1

Index	Type	Timestamp	Data
1	URL	2021-04-14 14:02:37Z	https://linkedsystems.uk/system/instance/TOOL0022_2490/current/
2	21.T11148/8eb858ee0b12e8e463a5	2021-04-15 16:41:02Z	{"identifierValue":"http://hdl.handle.net/21.T11998/0000-001A-3905-1
3	21.T11148/9a15a4735d4bda329d80	2021-04-14 14:02:37Z	https://linkedsystems.uk/system/instance/TOOL0022_2490/current/
4	21.T11148/709a23220f2c3d64d1e1	2021-04-14 14:02:37Z	Sea-Bird SBE 37-IM MicroCAT C-T Sensor
5	21.T11148/4eaec4bc0f1df68ab2a7	2021-04-14 14:11:28Z	[{"owner": {"ownerName":"National Oceanography Centre","ownerCo {"ownerIdentifierValue":"http://vocab.nerc.ac.uk/collection/B75/curren
6	21.T11148/1f3e82ddf0697a497432	2021-04-14 14:05:56Z	[{"manufacturer":{"manufacturerName":"Sea-Bird Scientific","manufa /collection/L35/current/MAN0013/","manufacturerIdentifierType":"UR
7	21.T11148/c1a0ec5ad347427f25d6	2021-04-14 14:02:37Z	[{"modelName":"Sea-Bird SBE 37 MicroCat IM-CT with optional pres {"modelIdentifierValue":"http://vocab.nerc.ac.uk/collection/L22/curren
8	21.T11148/55f8ebc805e65b5b71dd	2021-04-14 14:02:37Z	A high accuracy conductivity and temperature recorder with an optiona inductive modem for real-time data transmission plus internal flash me
9	21.T11148/f76ad9d0324302fc47dd	2021-04-14 14:02:37Z	["http://vocab.nerc.ac.uk/collection/L05/current/134/","http://vocab.ner
10	21.T11148/72928b84e060d491ee41	2021-04-14 14:05:56Z	[{"measuredVariable":{"variableMeasured":"http://vocab.nerc.ac.uk/co {"variableMeasured":"http://vocab.nerc.ac.uk/collection/P01/current/PS //vocab.nerc.ac.uk/collection/P01/current/TEMPPR01/"}},{"measured /PREXMCAT/"}}]
11	21.T11148/22c62082a4d2d9ae2602	2021-04-14 14:05:56Z	[{"date":{"dateValue":"1999-11-01","dateType":"Commissioned"}}]
12	21.T11148/eb3c713572f681e6c4c3	2021-04-14 14:05:56Z	[{"alternateIdentifier":{"alternateIdentifierValue":"2490","alternateIder
13	21.T11148/178fb558abc755ca7046	2021-04-14 14:05:56Z	[{"relatedIdentifier":{"relatedIdentifierValue":"https://www.bodc.ac.uk "URL","relationType":"IsDescribedBy "}}]
100	HS_ADMIN	2021-04-14 14:02:37Z	handle=21.T11998/USER22; index=300; [create hdl,delete hdl,read val

Figure 5. Example PID using data types registered in the data type registry

3.2.1 The TypeAPI

Data type registries are important step towards standardization and optimized use of PIDs. However, registration of data types brings new challenges. The first challenge for the user is to find the appropriate registry. Which registry is responsible for which science branch and where to find the appropriate type? Another aspect is the interpretation of the PID data. How to effectively resolve the references to the registered data types. For these tasks the TypeAPI was developed as part of the DICE contribution. The TypeAPI provides different endpoints (c.f. Figure

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) and allows configurable access to information from different data type registries. There are two groups of endpoints, data types and PIDs, which are explained in more detail below. The endpoint data type is used to search and discover types and to query type information. The PID endpoint is used for integrity checking of already created PIDs.

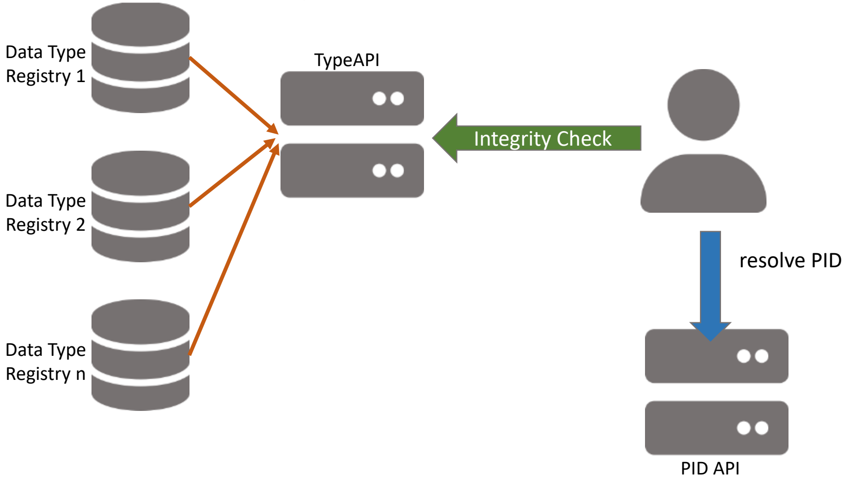


Figure 6. Simplified representation for the use of data types

3.2.1.1 *Searching Types*

The search endpoint "/dtype/search" allows searching and faceted searching for type descriptions in selected registers. The return value can be either JSON or HTML. This searches across all connected registries and can thus unify the use of types across different science domains. The search is not only limited to the description field in the data type registry, but all information such as the provenance information or the creator of the type is searched.

3.2.1.2 *Obtaining the type description*

If the PID of the type is known you can use the endpoints "/dtype/JSON" and "/dtype/HTML" to retrieve information about a specific type.

3.2.1.3 *Validation Schema*

The endpoints "/dtype/schema/JSON" and "/dtype/schema/HTML" allow to get the respective schema of a type in the two formats. This allows to implement own validation procedures.

3.2.1.4 *PID Integrity Check*

For integrity check of a PID the endpoints "/pid/content/*" can be used. The PID (prefix and suffix) is passed. The API checks the given types and compares the references with available references to the attached registries. If there is a match, the references are supplemented with the plain names of the types. In addition, the values in the PID to the types are validated with respective schemas from the registry. On the one hand, PIDs containing type references can be displayed in a human-readable way, and on the other hand, PIDs can be checked even if the data types originate from different registries.

3.2.1.5 *JSON LD*

In addition to returning the PID as a JSON or HTML object, the API also allows the PID to be returned as a JSON-LD object [29], which can be used as a container for linked data. Using JSON-LD allows web services and web applications that prefer to exchange their data in JSON to easily connect to the Semantic Web and collaborate more smoothly by using globally unique labels for logically ordered terms.

3.2.2 Implementation

The TypeAPI is implemented using Uvicorn [30] managed by Gunicorn [31] for high-performance FastAPI web applications in Python 3.6 and above with performance auto-tuning. Following list of additional Python models are required.

- fastapi
- jinja2
- typing
- pydantic
- jsonschema
- aiofiles
- urllib3

The code is maintained using the CI/CD functionalities of the gitlab service. This allows to automize the build of docker images which eventually are pushed to the gitlab container registry. The current implementation is accessible under http://typeapi.pidconsortium.net (the API endpoints) and http://typeapi.pidconsortium.net/docs/ (the documentation).

3.2.3 Summary

A unified view on the status of the integrity of the PIDs provided is important for both PID service providers, as well as service users. With this deliverable we addressed two key aspects of PIDs' integrity: (1) the integrity of the PID infrastructure, and (2) the integrity of PID metadata. The (1) is the basis of proper PID resolution, while (2) discusses the content of PID records and the usage of datatypes.

We presented the Prefix Information Service, which enables a transparent view about the integrity of PIDs for all DICE offered B2HANDLE services. This is the first achievement of our work.

We also described the TypeAPI service, which is a scalable solution for improving the use of datatypes in the PID landscape. This is our second achievement.

As next steps, we plan to suggest the Prefix Information Service for uptake by non-DICE providers that would allow a transparent view also about non-DICE offered PID services. Further endpoints of the TypeAPI are planned to improve and simplify the use of PID types. For example, further return formats are possible or a more detailed output for the validation of PIDs via the TypeAPI. We will also foster the integration of the TypeAPI in dedicated use-cases.

4 Long-Term Preservation Policy Report for EUDAT Services (in particular B2SHARE and B2SAFE)

4.1 Introduction

This chapter contains the deliverable of Task 4.3 of the DICE project, the formulation of Long-Term Preservation (LTP) Policies for the EUDAT Services B2SHARE and B2SAFE. The report consists of two main parts: Part I, this chapter, provides the considerations and methodology followed, and provides explanations of choices made; Part II, the appendices 1 to 7, consists of the Long-Term Preservation Policy Template and is accompanied by a number of additional appendices. The Policy Template intends to be generic, so that it can be used by a wide range of repositories and policy-based data archives to compose their LTP policies.

In section 4.2 "LTP Policy Template Introduction" we describe the aims of this task, and explain how the LTP Policy Template is applicable for B2SHARE, B2SAFE and other EUDAT services. The template has a modular structure consisting of sections following the functional components of ISO Standard 14721:2012 [32], the Open Archival Information System (OAIS) reference model, from which LTP Service Providers can select the applicable articles, compliant with the curation level they aim to support.

Section 4.3 concerns the methods followed in the formulation of the LTP Policy Template. Distinction is made between policies for digital preservation services "in-house" and when outsourced to an external LTP archive. As digital preservation has already a considerable history of at least several decades, the DICE project made use of earlier work on LTP principles. Also, the areas to be covered in an LTP Policy according to leading organisations in the area and as formulated by international projects on the subject seeking to formulate good practices and archival standards are described. It is motivated why the Reference Model for an Open Archival Information System (OAIS) [33] was chosen as the framework for the LTP Policy Template.

Section 4.4 explains how the LTP Policy Template can be applied, in particular for the EUDAT Services B2SHARE and B2SAFE. Particular attention is given to the interpretation of the term "designated community", which is often regarded as problematic for generic data services for a broad scientific community. Next, it is discussed how the LTP Policy Template corresponds to four diverging curation levels (as distinguished in the CoreTrustSeal [34] certification), ranging from bit preservation to data-level curation. Obviously, the higher the curation level, the more articles of the template apply. A table clarifies which articles from the template are needed in an LTP policy to satisfy the various levels of curation. Tables are also used to characterise the LTP requirements of the EUDAT services B2SHARE and B2SAFE, and to set forth which articles are applicable for both services.

Section 4.5 explains how LTP policies fit into the general agreement structure of EUDAT, the so-called Service Management Framework (SMF). Two diagrams explain how the LTP function can both be performed in-house (by an EUDAT data service provider running B2SHARE) or by an external LTP service provider. When LTP is outsourced, the LTP Policy should be accompanied by an LTP Agreement between the client and the contractor organisation. A draft template for such an agreement is available, but can only be finalised if the technical implementation of the data transfer between the two organisations is ready (scheduled for a later phase of the DICE project).

Section 4.6 discusses the issue of LTP costs and business models. The DICE project decided not to attempt to formulate an umpteenth cost model, but does provide a summarising overview of cost models developed by a dozen or more leading organisations and projects, including those discerned by the European 4C project [35] ("Collaboration to Clarify the Cost of Curation"). Although there exists considerable consensus about the most important factors contributing to

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 costs of curation and LTP, there appears to be quite diverging ways in how these work out in practice. It is not feasible to formulate a generally applicable template for a cost statement, other than to say that such a statement needs to be part of (or annexed to) an LTP agreement if the digital preservation is outsourced. In the final deliverable, where one use case will be implemented (the long-term archiving of the content of the main B2SHARE instance at DANS), an example of such a cost statement will be available.

Section 4.7 outlines the technical implementation of the DICE Digital Preservation Service and what the technical implications of the LTP Policy Template are. This section looks ahead to work to be done in the next phase of the DICE project. By creating technical specifications and creating a concrete implementation of such a service, all aspects of the LTP Policy Template will have an impact and potentially problematic issues may surface.

The “DICE Digital Preservation Service” (DDPS), as the new service will be called, will enable dataset transfer from a B2SHARE repository to the LTP archive of DANS ("outsourcing" use case). This work continues to build upon the work already done within the EOSC-Hub project and will also serve as proof of putting the LTP Template to practice.

It is expected that this Template is flexible enough to incorporate technical complications that may arise, and they will probably be addressed in the accompanying LTP Agreement mentioned in section 4.5.

Appendix 1 of this deliverable consists of the actual Long-Term Preservation Policy Template. The template consists of seven sections and is structured in line with the Open Archival Information System (OAIS) reference model (ISO Standard 14721:2012 [32]) according to functional elements of digital preservation. After setting forth the objectives, scope and delimitation of an LTP policy, those components are:

- Ingest
- Archival Storage
- Data Management
- Administration
- Preservation Planning
- Access

The sections on Ingest and Access are partly dependent on whether the digital preservation is carried out in-house or is outsourced, and on some more detailed choices made. The template provides alternative formulations for the articles reflecting these situations.

The LTP Policy Template is accompanied by the following appendices:

Appendix 2. Summary of the OAIS Reference Model

Appendix 3. Recurring monitoring processes

Appendix 4. Available Licenses for digital assets uploaded to the EUDAT B2SHARE service

Appendix 5. Legal and Statutory Context and Requirements

Appendix 6. Selected terms used

Appendix 7. LTP comparison between B2SHARE and B2SAFE

4.2 LTP Policy Template Introduction

Task 4.3 of the DICE project concerns Long Term Preservation. The description of work involves the formulation of long-term preservation (LTP) policies for the EUDAT services B2SHARE and

B2SAFE. In this task, we aimed at formulating a model LTP Policy *Template*, which can serve the long-term preservation needs of a broad range of repository services. The template is included in the Appendix 1 of this document.

B2SHARE and B2SAFE, as well as other data services, either in the context of EUDAT or otherwise, can select the sections and articles in the template that apply to their situation. Depending on the context, EUDAT and other data service providers can use the model template to create a full LTP policy document or include a brief LTP paragraph in their Service Level Agreements (SLAs) and Operational Level Agreements (OLAs). For EUDAT services in which LTP plays a minor role or no role at all, LTP paragraphs can be deduced from the template, in order to clarify to users what is or can be guaranteed, for how long, and by whom. The LTP Policy Template is designed to be connected to or included in the "EUDAT Service Management Framework", version 2.5 (19/2/2021).

By following ISO standard 14721:2012 [32], the Open Archival Information System (OAIS) reference model, from which most certification systems (such as CoreTrustSeal) are derived, policy-based data archives implementing this LTP Policy Template will be well positioned to certify their service. In this way, the LTP Policy Template is supporting Trustworthy Digital Repositories and the implementation of the FAIR data principles.

The LTP Policy Template for B2SHARE will be accompanied by a model Long-Term Preservation Agreement (LTP Agreement), in case this task is outsourced to an external, dedicated LTP archive. In the DICE project, the plan is to implement the long-term preservation of one B2SHARE service by outsourcing it to a dedicated long-term archive for research data (i.e. DANS). A draft LTP Agreement template is available, but we decided not to include it in this deliverable, as it still needs to be agreed between the outsourcing B2SHARE service (i.e. CSC on behalf of the EUDAT) and the external LTP service provider (i.e. DANS). This agreement can only be finalized when the technical implementation of the data transfer between both organisations is in place, which is due only at a later phase of the project.

The LTP Policy Template takes the form of a written document from which the applicable articles can be selected. Sometimes there are alternative options that can be chosen.

For B2SHARE service providers, the template can be used both in case the long-term preservation is outsourced, and in case the B2SHARE service provider itself takes care of the LTP service.

In the case of outsourcing of the LTP service, one or more additional contracts will be needed:

- An LTP Agreement needs to be in place if datasets are transferred from a B2SHARE repository to an external LTP Service.
- In case personal data is involved in the transfer of data to an LTP Service, the GDPR requires the signing of a Data Processing Agreement.
- The coverage of costs of the LTP Service is to be specified in a cost statement as an annex to the LTP Agreement.

The LTP Policy Template can also be used for defining an LTP policy for B2SAFE: a more limited selection of articles applies here. Further on in this chapter, a table indicates which articles are recommended to be included in some typical situations. For B2SAFE, a distinction will be made between maintaining the metadata (data documentation) and preserving the actual content. The LTP strategy for B2SAFE will in particular meet the requirement of FAIR Principle [36] A2: "A2 metadata are accessible, even when the data are no longer available".

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 wider application, the LTP Policy Template will be brought to the attention of the EOSC Task Force on Long Term Data Preservation (EOSC TF LTP or EOSC Preservation Taskforce). The Charter of this Taskforce (version 0.5 (08-06-2021)) mentions:

- "The Strategic Research and Innovation Agenda (SRIA) of the EOSC underlines the importance of long-term data preservation, but an explicit strategy has not been formulated. The EOSC TF LTP will provide recommendations for the EOSC board on the vision and sustainable implementation of long-term data preservation policies and practices, as well as suggestions to later strategy execution. It will address the roles and responsibilities of the different stakeholders, the financial aspects of long-term preservation and the necessary service infrastructure."
- "[A] horizontal EOSC preservation policy enables the connection and collaboration on national, community and local level."

We suggest that the LTP Policy Template developed in the DICE project may serve, or at least be input for, such a "horizontal EOSC preservation policy" ambitioned by the EOSC TF LTP.

4.3 Method: Approaches to LTP Policies

For the DICE LTP policies we will make good use of earlier work carried out in the development of digital preservation policies. In the references we will refer to websites and online documents providing good practices and templates for digital preservation policies.

4.3.1 Notational Convention

Throughout this document [square brackets] are used to refer to a templated organisation, like ([LTP Institution]) or a service ([LTP Service]). These should ultimately be replaced by the actual name of the involved institute or service.

4.3.2 Scope: Outsourcing versus in-house LTP

An [LTP Institution] is an organisation which is responsible and liable for long-term preservation. If this organisation also runs and operates the [LTP Service], we will refer to it as "In-house LTP". However, if the [LTP Service] is outsourced to another organisation, this organisation is also required to adopt the LTP Policy and to implement it in the [LTP Service] it offers. We will refer to this situation as "outsourcing".

For example, in the case of EUDAT, the B2SHARE "catch-all" service is an EUDAT service. The EUDAT CDI is responsible, but the service is hosted and operated by one of its members, in this case CSC in Helsinki, Finland. EUDAT is responsible for ensuring the LTP policy: it needs to be adopted by CSC on behalf of EUDAT.

4.3.3 LTP Principles

In January 2007, representatives of four leading preservation organizations[1] formulated ten basic characteristics [37] of digital preservation repositories, which are listed below:

A long-term preservation repository ...

[1] The organizations were:
- The Digital Curation Center (UK)
- Digital Preservation Europe (DPE, a European consortium of libraries, archives and university institutions)
- NESTOR (Germany)
- Center for Research Libraries (North America)

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. (...) commits to continuing maintenance of digital objects for (an) identified community/communities.
2. (...) demonstrates organizational fitness (including financial, staffing, and processes) to fulfil its commitment.
3. (...) acquires and maintains requisite contractual and legal rights and fulfils responsibilities.
4. (...) has an effective and efficient policy [framework][2].
5. (...) acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.
6. (...) ensures the integrity, authenticity and usability of digital objects it holds over time.
7. (...) creates and maintains requisite metadata about the provenance[3] of digital objects it holds and about actions taken on them during preservation.
8. (...) fulfils requisite dissemination requirements.
9. (...) has a strategic program for preservation planning and action.
10. (...) has technical infrastructure adequate to continuing maintenance and security of its digital objects.

Some formulations in these principles raise questions. Including the following:

Ad 1: What is/are "identified [or designated] community/communities"?

Ad 5: What are "stated criteria" for data acquisition?

Ad 7: What is "requisite metadata"?

These questions address respectively for whom (the target audience), what and how (well described) digital objects are preserved. The subject of "designated communities" is discussed in section 4.4 below. The criteria for acquisition are obviously related to the target audience and it also influences the way and degree of detail in which data are described. Anyhow, an LTP Policy should reflect these principles, and organizations offering services to preserve data for the long term should take these principles seriously.

4.3.4 Areas to be covered in an LTP Policy

There are many good practice documents that try to provide guidance on which areas to cover in a Digital Preservation Plan or Long-Term Preservation Policy (LTP Policy), such as those by the British Library and the National Archives in the UK, or those by the Library of Congress and National Archives in the US. There were also recommendations by specific digital preservation projects, of which the InterPARES projects (that started in 1999) [38] was particularly influential. ERPAnet (Electronic Resource Preservation and Access Network) was another leading group that formulated guidelines in the 2000s. Such recommendations, handbooks and other online resources were usefully summarized by the Digital Curation Centre (DCC) in the UK. According to the DCC, a Digital Preservation Policy describes an "organisation's aims and objectives about the long-term care of digital objects", specifying [39]:

- Preservation strategies and acceptable actions
- Decisions about the digital objects (formats, metadata)

[2] See Appendix 6 for the term "policy framework".

[3] Described as "the relevant production, access support, and usage process contexts before preservation".

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

- Standards
- Who the material is being preserved for
- Resourcing
- Responsibilities

Perhaps the most extensive and detailed recommendations for an LTP policy are provided by the reference model for an Open Archival Information System (OAIS, ISO 14721:2012) [40], published in 2012 ("Purple book") [41] and updated in 2020 ("Pink book") [42]. The OAIS model organizes a model archival system in terms of functional entities:

- Ingest
- Data Management
- Archival Storage
- Preservation Planning
- Administration
- Access

A brief overview of the OAIS reference model is provided in Appendix 2. The LTP Policy Template we present here follows the OAIS functional entities.

There are also commercial guidelines and tools available for digital preservation, and companies such as Artefactual Systems Inc., the main developer of the open source digital preservation system Archivematica, also provide guidelines for digital preservation policies and good practices (see, for example, Preservation Planning | Documentation (Archivematica 1.13.2) [43]).

Many organisations with a responsibility for digital preservation, have formulated their digital preservation strategies and LTP policies. In most policies, references are given to the legal and regulatory frameworks in which the organisation operates, and to relevant standards, certifications and other guidelines. Obviously, the Preservation Plan by Data Archiving and Networked Services (DANS, Version 1.0 - May 2018; currently being updated: Preservation plan | DANS [44]), which entails both a preservation strategy and a policy, is important here. A good model of a preservation policy is provided by the UK Data Archive (version 12.00, 26 January 2021), see: UK Data Archive Preservation Policy [45].

To wrap up, perhaps the most fundamental element of any LTP Policy is a mission or dedication by an organisation to keep digital assets alive. And in order to guarantee that, the organisation itself must be sustainable.

4.4 Applying the Long-Term Preservation Policy Template

4.4.1 Outline of the LTP Policy Template

The appendices 1-7 of this document hold the template for long-term preservation policies of research data and related assets, which should at least be applicable for the EUDAT services B2SHARE and B2SAFE.

The basic assumption for this policy is that the EUDAT B2SHARE service assumes responsibility for some key elements of what is normally included in a digital preservation plan, in particular:

- Activities related to the ingest of digital assets uploaded by researchers to B2SHARE, including the metadata describing these assets.

- Activities related to the access of digital assets stored in B2SHARE, including access agreements and licenses, *for as long as the B2SHARE service exists*.

If for any reason the B2SHARE service will be discontinued, the ingest of new materials will stop; however, the continued access to the assets in B2SHARE, which have been archived at [LTP Institution], will be arranged by the [LTP Service] and hence this policy includes a section dedicated to access.

The EUDAT B2SHARE service needs to indicate whether it wants the data it holds to be findable (and hence visible) to the outside world via the [LTP service].

- If this is the case (data findable/visible), the [LTP Service] will make the metadata available through its search mechanism, but it will refer for access to the data in B2SHARE via the B2SHARE persistent identifier.
- If this is not the case the [LTP Service] is a "dark archive" and metadata will not be exposed and hence not be findable in the [LTP Service], unless the B2SHARE service ceases to exist. In that case, the [LTP Service] will make the metadata available through its search mechanism and will provide access to users, according to the access policy.

The data and metadata from B2SHARE will be ingested into the [LTP service] via an automated process, which is described in a separate document (LTP Agreement) and which is to be implemented as part of a demonstrator service in Task 4.3 of the DICE project[4].

The digital objects in EUDAT B2SHARE can be distinguished on three levels:

- Community: a collection of datasets, created by and relevant for a certain group of researchers; this is equivalent to a "collection" in [LTP Service].
- Record: an organized collection of data files belonging to a research project, or created by one researcher or research group, together with their descriptions (metadata); this is the equivalent of a dataset in [LTP Service].
- File: a digital file in which data is stored; it typically has a particular organisation or format, which is usually reflected by a file naming convention (such as the file extension); this corresponds to the "data file" in the [LTP Service]. A Record or Dataset usually contains one or more (data) files, although Records/Datasets may only contain metadata without data files (e.g. as "tombstones" or references to data stored elsewhere).

4.4.2 EUDAT B2SHARE as a designated community or communities

An issue that needs special consideration is the definition of the designated community or communities. The research (or scientific, or scholarly) community is defined as a diverse network of interacting scientists and scholars. It includes many "sub-communities" working on particular scientific/scholarly fields of research, and within particular institutions; although many research communities are distinguished on the basis of disciplinary boundaries, interdisciplinary and cross-institutional activities are also significant. In the context of EUDAT B2SHARE it is significant that research communities contribute to a certain volume of digital research data of relevance for that community.

B2SHARE is described as an EUDAT service for "*researchers, scientific communities and citizen scientists*" to store and share *small-scale research data* (including *software code*) from *diverse contexts*.

[4] This deliverable is due later in the project, hence the LTP Agreement template is not yet included in this document, although a draft is available.

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 is a self-depository system. Researchers (data producers) are responsible for successfully uploading data and for providing documentation.

Only registered users of B2SHARE can create new records and upload data into these records. The current default maximum size for a record is 20 GB and for an individual file 10 GB. The uploading of large numbers of files or the creation of large numbers of records is not considered "fair use" [46] by the B2SHARE service. If researchers want to publish many datasets, they should contact EUDAT through their service request portal [47].

This description makes clear that the service is intended for a varied audience:

- For individual researchers as well as scientific communities.
- For professional and citizen scientists.
- For research data and software of small to moderate size.
- "From diverse contexts", which we interpret as from various disciplines.

The first of the ten preservation principles (see section 4.3) states that an LTP repository should commit "to continuing maintenance of digital objects for identified community/communities". This principle returns in several requirements of the CoreTrustSeal (CTS), which repeatedly stresses that the needs of the "Designated Community" are to be known, respected and served.

It is debatable whether the description above satisfies the community aspect of the principles and requirements sufficiently. Both the principles and the CTS, make the assumption that LTP Repositories serve particular communities, and indeed, many repositories serve, for instance, specific disciplines. Admittedly, particular data types are used in particular ways by particular communities, which may have consequences for their treatment in repositories: they may, for instance, require special metadata elements to be described adequately.

Yet, the B2SHARE service, like many other discipline-independent and institutional repositories, was set up for a broad range of scientific and scholarly uses. The characteristics and needs of a diversity of scientific communities were incorporated in the design and functionality of the service.

Moreover, according to the OAIS Reference Model, the Designated Community for an OAIS service is identified as an "identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time."

According to the same ISO standard 14721:2012, a Consumer is defined as "the role played by those persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail. This can include other OAISes, as well as internal OAIS persons or systems."

It is obvious that the [LTP Service] inherits the researchers and communities served by B2SHARE as users by proxy. Still, this LTP policy is formulated specifically to satisfy the demands of EUDAT B2SHARE, which we may also consider as a designated community here.

4.4.3 The LTP Policy Template and Curation Levels

In the CoreTrustSeal (CTS) Requirements for Trustworthy Digital Repositories [48], CTS distinguishes between four levels of curation that can be performed and should be selected in the certification process. These are:

A. Content distributed as deposited.

B. Basic curation – e.g., brief checking, addition of basic metadata or documentation.

C. Enhanced curation – e.g., conversion to new formats, enhancement of documentation.

D. Data-level curation – as in C above, but with additional editing of deposited data for accuracy.

The relationship between these curation levels and the relevant LTP template sections are summarised in Table 3. The table indicates which LTP Policy articles are needed to comply with the different curation levels. The table only comprises those articles that are directly relevant for digital curation. A distinction of "in-house"(2A) and "outsourced"(2B) LTP is also made available. An empty cell indicates that the article is not needed for this curation level, whereas an 'x' marks that the article is needed for compliance with the corresponding curation level.

Table 3. LTP Policy Template sections and curation levels

Policy Section and Article		Brief description	Curation Level			
2. Ingest			**A: Bit preservation**	**B: Basic Curation**	**C: Enhanced Curation**	**D: Data-level curation**
2A. LTP in-house	**2B. LTP outsourced**					
2.1	2.6	Submission Information Package for Ingest	x	x	x	x
2.2	2.7	Metadata index and Cataloguing		x	x	x
2.3	2.8	Licenses		x	x	x
2.4	2.9	Persistent Identifiers		x	x	x
	2.10	Additional PIDs		x	x	x
2.5	2.11	Quality control	none	basic checks of metadata	enhanced checks of metadata	enhanced checks of data and metadata
2.5.1	2.11.1	Metadata supplied by depositor		x	x	x
2.5.2	2.11.2	Ingest control	none	basic	enhanced	enhanced
	2.11.3	Mapping of metadata		x	x	x
2.5.3	2.11.4	Required metadata fields	none	basic	extended	full
2.5.4	2.11.5	Responsibility for metadata		x	x	x
2.5.5	2.11.6	Notification of deficient metadata			x	x
2.5.6	2.11.7	Data and metadata corrections			x	x
3. Archival Storage						
3.1		Archival actions and chain of provenance	none	basic	extended	full
3.2		Archival storage of AIPs	x	x	x	x
3.4		Integrity and security measures	checksums	basic	extended	extended

4. Data Management						
4.1		Metadata catalogue maintenance		x	x	x
4.2		Archival metadata supporting version control	x	x	x	x
4.3.1		Derived and dissemination formats			x	x
4.3.2		Authenticity and integrity of AIPs	x	x	x	x
4.3.3		Documentation of chain of custody		x	x	x
4.4		Preferred and accepted file formats			x	x
4.5		Obsolescence of file formats				x
4.6.1		Migration of preferred formats				x
4.6.2		Preserving outdated formats				x
4.7		Deletion of datasets and files		x	x	x
4.8		Tombstone records for deleted datasets	x	x	x	x
6. Preservation Planning						
6.1		Monitoring the content of the [LTP Service]	x	x	x	x
6.2		Reviewing and updating list of preferred formats			x	x
6.3		Other monitoring for preservation planning		x	x	x

4.4.4 How to use the LTP Policy Template for B2SHARE and B2SAFE

To understand how the Long Term Preservation policy applies to a service we need to understand the services, the commonalities and differences of the B2SHARE and B2SAFE services and technologies. Both services have been onboarded on the EOSC Service Catalogue[5] and are offered free-at-the-point-of-use via virtual access via the DICE project.

Table 4 describes the B2SHARE and B2SAFE services and capabilities related to aspects of long-term preservation.

Table 4. Service description of B2SHARE and B2SAFE

	B2SHARE	B2SAFE
EOSC Service Catalogue	https://marketplace.eosc-portal.eu/services/b2share	https://marketplace.eosc-portal.eu/services/b2safe
Short description	B2SHARE is a user-friendly, reliable and trustworthy way for researchers, scientific communities and citizen	B2SAFE is a robust and highly available service which allows community and departmental

[5] https://marketplace.eosc-portal.eu/

	B2SHARE	B2SAFE
	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 organisations customised instances and/or access to repositories supporting large datasets.	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.
Features	● Support of metadata descriptions via the EUDAT Core metadata schema ● Registers DOIs for datasets and Handle PIDs for data objects ● Supports versioning ● Harvested by B2FIND and OpenAIRE Explore ● Supports direct upload from B2DROP ● Accessible via a Web GUI and a REST API to support automatic publishing workflows ● Supports community domains ● Allows communities to define metadata extensions, access rules and publishing workflows ● Allows references to externally hosted data objects	● Support for data management policies (e.g. registration of PIDs, cross-site replication, data integrity checks) ● Support for policies customised to community and organisational 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
Curation level	Enhanced	Bit preservation
Designated community	Yes, at time of writing 22 community domains are supported	Yes, in agreement with the customer and contract
Access	Public free-to-use service, depositors need to register an account	Closed service, customers need to request access
Metadata	Yes, supports metadata via EUDAT metadata schema and community specific extensions	Optionally, basic object information is maintained, user descriptive metadata can be provided
PID	Yes, registers DOIs for the landing pages of datasets and Handle PIDs for individual data objects	Yes, PIDs for data objects are automatically registered via the PID data management policy

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	B2SAFE
License	Optionally, licenses can be added to a dataset	No, users are not able to specify a license
Versioning	Yes, versioning is supported	No, options available to implement a versioning policy
Deletion of datasets	Yes, only authorised staff are able to delete published data records	No, data owners are allowed to delete data objects

As can be seen in Table 4, the B2SHARE and B2SAFE services have different aims with respect to curation level. B2SHARE aims at an enhanced curation level, while the aim of B2SAFE is bit preservation, therefore the B2SHARE and B2SAFE service support different capabilities. Table 5 provides a comparison between B2SHARE and B2SAFE in supporting the LTP policy as described in Table 3. The table lists an overview of all the sections and articles provided by the LTP Policy Template and the curation status in both services. Statuses that have been used within this table include:

- Yes - The service supports the LTP policy article.
- No - The service does not support the LTP policy article.
- Partially - The service complies partially with LTP policy article, a short explanation is provided.
- Optionally - The service provides capabilities to comply with the LTP policy article but are optionally, and/or not mandatory for the user and/or designated community to be used, a short explanation is provided.

A more detailed comparison table than Table 5 is provided in Table 9 of Appendix 7.

Table 5. LTP Policy comparison between B2SHARE and B2SAFE

Section and Article	Brief description	Data service			
		B2SHARE		B2SAFE	
		Curation	Status	Curation	Status
		Enhanced		Bit preservation	
2. Ingest					
2.1	Submission Information Package for Ingest	X	Yes	X	Users are able to store AIP packages, but this is not mandatory.
2.2	Metadata index and Cataloguing	X	2 publication workflows are supported, including with a reviewed process		No
2.3	Licenses	X	A license can optionally specified		No
2.4	Persistent Identifiers	X	Yes		Yes
	Additional PIDs		Yes		No
2.5	Quality control	enhanced checks of metadata	enhanced checks of metadata	none	No
2.5.1	Metadata supplied by depositor	X	Yes		No
2.5.2	Ingest control	enhanced	Minimum metadata needs to be specified (e.g. title, creator, checksums)	none	checksums are automatically generated and verified
2.5.3	Required metadata fields	Extended	Yes	none	No
2.5.4	Responsibility for metadata	X	Yes		No
2.5.5	Notification of deficient metadata	X	Optionally via reviewed publication workflow		No
2.5.6	Data and metadata corrections	X	Yes		No

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. Archival Storage					
3.1	Archival actions and chain of provenance	extended	Partially, via version management	none	No
3.2	Archival storage of AIPs	X	Yes	X	Yes
3.3	No responsibility for data stored externally	X	Yes	S	Yes
3.4	Integrity and security measures	extended	Partially, service provided on basis of a OLA and security is managed via the EUDAT ISM process	checksums	Yes
4. Data Management					
4.1	Metadata catalogue maintenance	X	Yes		Partially, the service supports a searchable metadata database, but not required to specify metadata.
4.2	Archival metadata supporting version control	X	Yes	X	No, versioning is not supported
4.3.1	Derived and dissemination formats	X	Yes		No
4.3.2	Authenticity and integrity of AIPs	X	Yes	X	Yes
4.3.3	Documentation of chain of custody	X	Yes		No
4.4	Preferred and accepted file formats	X	No, no preferred file formats specified		No
4.5	Obsolescence of file formats		No		No
4.6.1	Migration of preferred formats		No		No
4.6.2	Preserving outdated formats		No		No
4.7	Deletion of datasets and files	X	Yes	X	No, data owners are allowed to delete data objects

4.8	Tombstone records for deleted datasets	X	No tombstone generated	X	No tombstone generated
6. Preservation Planning					
6.1	Monitoring the content of the [LTP Service]	X	Yes	X	No
6.2	Reviewing and updating list of preferred formats	X	No		No
6.3	Other monitoring for preservation planning				
6.3.1	Security risks	X	Yes		Yes
6.3.2	Technology watch	X	Yes		Yes
6.3.3	Service requirements	X	Yes		Yes
6.4	Roles and responsibilities, confidentiality, liability	X	Yes		Yes
6.5.1	Funding adequate for sustaining	X	The EUDAT CDI guarantees services for a period of 10 years by its members.		The EUDAT CDI guarantees services for a period of 10 years by its members.
6.5.2	Contingency plan	X	Yes, for the duration of the EUDAT CDI Partnership agreement		Yes, for the duration of the EUDAT CDI Partnership agreement and optionally within a contract

From the LTP policy comparison made in Table 5 it can be concluded that the B2SHARE CDI and the B2SAFE services comply to a large extent with the LTP criteria as defined in Table 3 for the aimed curation levels. To become compliant with the LTP policy, the non-compliant articles will be assessed in more detail in the next phase of the project.

4.5 Agreement structure within EUDAT

The EUDAT Collaborative Data Infrastructure (or EUDAT CDI) is one of the largest infrastructures of integrated data services and resources supporting research in Europe. It is sustained by a network of more than 20 European research organisations, data and computing centres that in September 2016 have signed a partnership agreement to maintain the EUDAT CDI for the next 10 years and in 2018 have supported the establishment of the limited liability company, EUDAT Ltd.

The EUDAT CDI partnership agreement includes a Service Management Framework (SMF) which defines the principles, policies, agreement framework and structured processes governing the EUDAT collaborative data infrastructure (CDI). This structure is schematically shown in Figure 7. Current EUDAT agreement structure.

The purpose of the SMF is to clarify the operational constituents, roles and responsibilities of the *Service Provider* and to ensure a high quality of the service delivery to the *Customers* and their users.

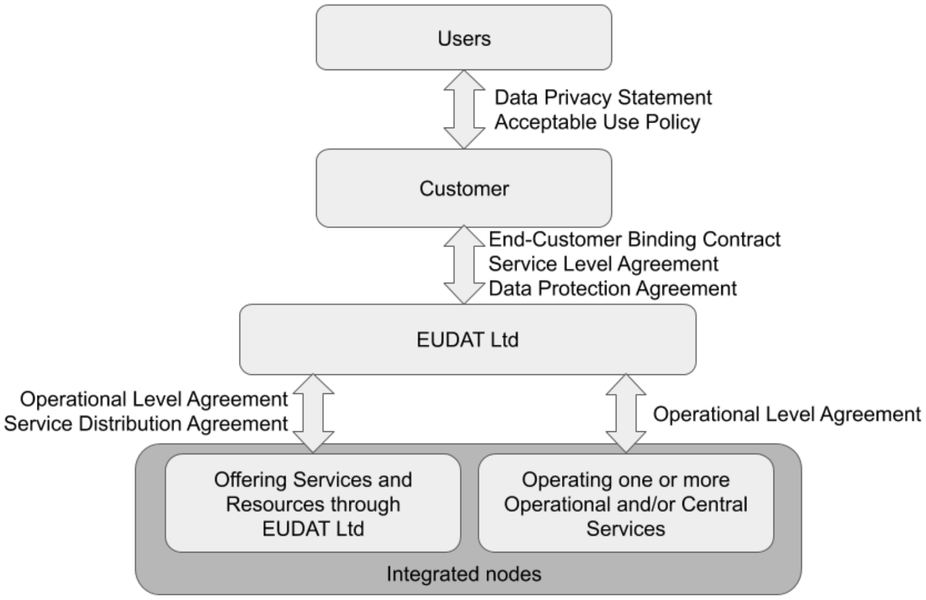

Users
Data Privacy Statement
Acceptable Use Policy
Customer
End-Customer Binding Contract
Service Level Agreement
Data Protection Agreement
EUDAT Ltd
Operational Level Agreement
Service Distribution Agreement
Operational Level Agreement
Offering Services and Resources through EUDAT Ltd
Operating one or more Operational and/or Central Services
Integrated nodes

Figure 7. Current EUDAT agreement structure

The agreement structure within EUDAT consists of agreements between a customer and EUDAT Ltd which are supported by agreements between EUDAT Ltd and the members who are hosting and operating the services and resources provided through EUDAT Ltd.

- Customer: The customer is the organization and/or mandated person with whom EUDAT Ltd has a binding contract for the delivery of the services and resources. The customer is also the entity that deposits data in the EUDAT service. The contract is underpinned by a Service Level Agreement (SLA) describing the service level targets and responsibilities between EUDAT Ltd and customers. Within the EUDAT CDI infrastructure and services at different levels personal data is being managed. To ensure that this personal data is managed according to the General Data Protection Regulation (EU) 2016/679 (GDPR) the Data Processing Agreement (DPA) is used.
- *Service provider*: To underpin the customer contracts EUDAT Ltd has agreements with level 1 CDI members in provisioning of services and resources through EUDAT. To allow the marketing and reselling of services to customers, EUDAT Ltd has Service Delivery Agreements with the service providers offering services through EUDAT. Underpinning the SLA's with customers, but also the provisioning of EUDAT central services (e.g. B2SHARE) and the operational tools, EUDAT Ltd has Operational Level Agreements.
- *User*: The user is the actual user of the service, who has an account to make use of the service. Access and usage of the service is underpinned by an Acceptable use Policy (AuP) and Data Privacy Statement (DPS). The AuP sets out the terms under which you may access our Services and applies as soon as you access and/or use. The DPS describes what personal data and how this is being handled within EUDAT services. Therefore the DPS are service specific.

The scope of the Long Term Preservation task within DICE is to extend the EUDAT agreement framework with a Long Term Preservation policy and agreement supporting the delivery of the B2SHARE and B2SAFE services through EUDAT Ltd.

- **Long Term Preservation Policy (LTP Policy)** describes preservation policy to ensure the long-term preservation and accessibility of electronic information while ensuring the highest level of authenticity possible.
- **Long Term Preservation Agreement (LTP Agreement)** supplements the SLA and OLA agreements describing the preservation level, on the basis of the LTP Policy, agreed either between the customer and EUDAT Ltd and/or between EUDAT Ltd and the organisation providing the LTP service.

Depending on the capabilities and aim of the service and service provider, a service provider can decide to support the full LTP policy by itself, or decide to make use of an external service provider to outsource long-term preservation of the data. To be able to do so needs to be organised in the customer contracts, unless this is possible on another legal basis (for instance: the data is in the public domain).

Independently if a service provider delegates the responsibility for the long-term preservation to an external service provider, the service in which the data is initially deposited needs to comply with LTP policy. When a service provider makes use of an external LTP service and service provider the LTP policy defines additional requirements the external LTP service provider has to comply with too.

In the following two diagrams the agreement structure, including the LTP policy and agreement, is explained. Figure 8 shows a preservation setup in which the service provider takes full responsibility for the LTP policy. In the context of the EUDAT B2SHARE CDI service, the service is offered as public service by EUDAT Ltd. and the service is hosted and operated by CSC. Therefore, EUDAT is responsible to comply with the LTP policy, to supply the enhanced curation level towards the users, while CSC is responsible for the execution of the LTP.

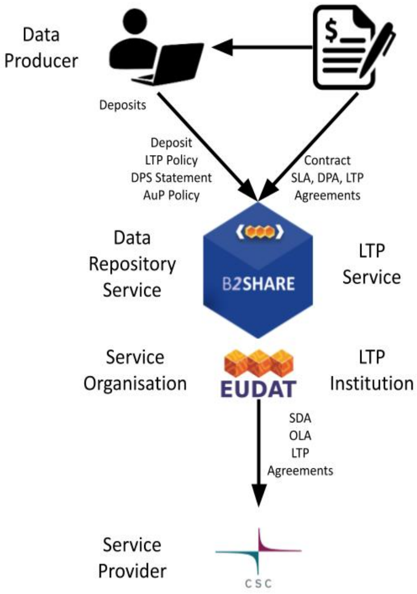


Figure 8. Data Repository Service is LTP Service

Figure 9. Long preservation is delegated by Data Repository Service to external LTP Service. It shows the long-term preservation set up with an external LTP service provider, for which EUDAT has an LTP agreement with DANS for the long-term preservation of data records. Data records are deposited in B2SHARE and preserved in the Data Vault of DANS.

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

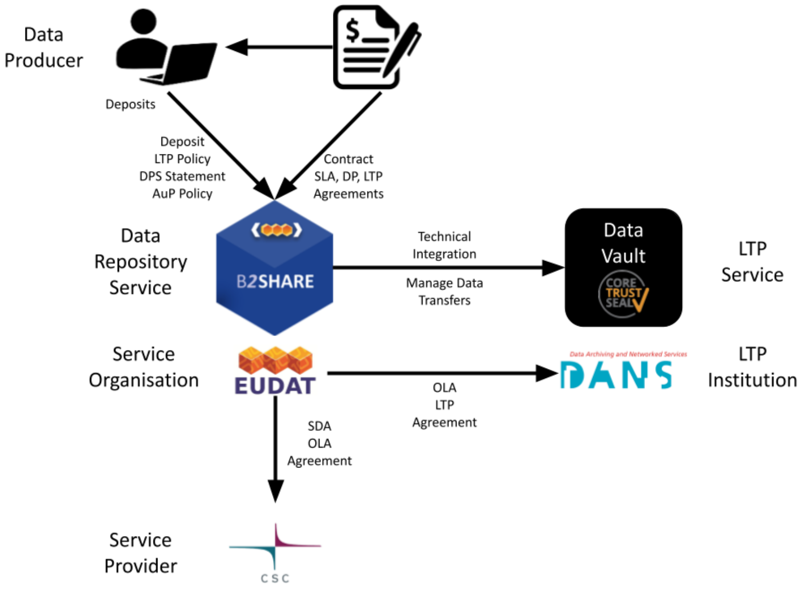


Figure 9. Long preservation is delegated by Data Repository Service to external LTP Service

4.6 Cost modelling for LTP

4.6.1 LTP Costs and business models in the DICE DoW

In the DoW of Task 4.3 it was expected that the majority of the costs for long-term preservation would occur during the ingest of research data into the LTP Service, during which the checking of metadata quality could be an important cost factor.

In the case of B2SHARE, the responsibility for the quality of the data descriptions and documentation lies primarily with the communities owning the data. In the LTP Policy Template, the amount of checking performed by the LTP Service depends on the level of curation offered by the Service, as reflected in Table 3. Storage costs are a second main factor in digital preservation, and these obviously return yearly. As B2SHARE is mainly geared towards the preservation of “long-tail data” (which are relatively small or modest in size), the storage costs so far are quite overseeable.

In the case of B2SAFE, which is dealing with considerably bigger data volumes, the costs of storage for the long-term are considerable, and the core questions to be answered for an LTP strategy are in the business model and in the selection of the data that need archiving. Without a business model in which storage costs are covered, any LTP policy for B2SAFE is bound to fail. “How to recover the costs for the preservation of data in the long-term? And who pays for what?” were questions posed in the DoW.

In the DICE DoW of Task 4.3 we foresaw that the LTP policy for B2SAFE would consist of a “written document with one or more model contracts for the coverage of the costs. In the limited time frame and funds for the project, recommendations on testing the implementation of such a strategy annex business model is the maximum attainable”.

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

4.6.2 Towards cost models for B2SHARE and B2SAFE

Over the years, a variety of studies, projects and reports on the costs of long-term preservation have been performed or produced[6]. It is unfeasible to provide an extensive overview of the literature on the subject here. Many different cost models have been developed and described. We summarize the most important and well-known of them in Table 6.

Table 6. Overview of cost models for digital preservation

ID	Name	Acronym	Owner	Authors (year)
1	Test bed Cost Model for Digital Preservation	T-CMDP	National Archives of the Netherlands	Kejser et al. (2011)
2	NASA Cost Estimation Tool	NASA-CET	National Aeronautics & Space Administration (NASA)	Hendley (1998)
3	LIFE3 Costing Model	LIFE3	University College London and the British Library	Wheatley et al. (2009)
4	Keeping Research Data Safe	KRDS	Charles Beagrie Ltd	Stanger (2011)
5	Cost Model for Digital Archiving	CMDA	Data Archive and Networking Services (DANS)	Palaiologk et al. (2012)
6	Cost Model for Digital Preservation	CMDP	Danish National Archives and the Danish Royal Library	Kejser et al. (2012)
7	DP4LIB Cost Model	DP4LIB	The German National Library	DP4Lib (2013)
8	PrestoPRIME Cost model for Digital Storage	PP-CMDS	The PrestoPRIME Project	PrestoPRIME (2011)
9	Total Cost of Preservation	CDL-TCP	California Digital Library	University of California (2013)
10	Economic Model of Long-Term Storage	EMLTS	David Rosenthal	Rosenthal et al. (2011/2012)
11	Digital Curation Sustainability Model	DCSM	4C Project	Grindley (2015)
12	ENSURE Cost Model for Long-Term Digital Preservation	ENSURE	ENSURE (EC FP7 project)	Xue et al. (2011)
13	Dutch Cost Model for Digital Preservation	DCMDP	Nationale Coalitie Digitale Duurzaamheid (NCDD)	Uffen et al. (2017, 2019)

Adapted (and extended) from the Summary of Cost Models [49] by the 4C [50]

There is a certain correspondence on the cost categories the various models discern, and which factors are contributing to the total costs of digital preservation, but at the same time there appears to be a great variety of the ways in which these costs are calculated in practice. The actual costs will largely depend:

- on the volumes of content to be preserved

[6] An overview of cost projects for digital preservation is provided by the EU-funded 4C-project: https://www.4cproject.eu/community-resources/related-projects/

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

- on the curation level that is offered by the LTP service provider
- on the organisational context, funding situation and charging policy of the LTP service provider
- on whether or not the digital preservation is outsourced (and which choice is made with respect to provision of access, including user support)

Writing a more or less generic cost template or business model for either B2SHARE or B2SAFE would be to add to the range of choice that is already available. Providing one or more model contracts for the coverage of the costs also seems impractical, as such contracts need to be specific for the organisations providing the LTP service.

Therefore, we refer potential LTP service providers for B2SHARE or B2SAFE to the cost models that have been proposed as summarized in Table 6. In the implementation of the long-term preservation of the B2SHARE content by an external provider (i.e. DANS), a cost statement will be part of the LTP agreement, which will be annexed to the LTP policy. This agreement will be made between the institution responsible for the B2SHARE service on behalf of the EUDAT Consortium Consortium (i.e. CSC) and the LTP service provider (i.e. DANS). Given the dependencies mentioned in the bullets above, a general model for such an annex is not practical to provide, but a draft cost statement for this particular situation is available, albeit that the exact specification is still under negotiation. Therefore, the cost statement will not be part of this deliverable, nor will it be part of the next deliverable within this task.

4.7 Technical implications related to the LTP Policy Template

Further developments within Task 4.3 will focus on the technical implementation of the LTP policy for B2SHARE data transfer to a CTS certified archive.

By creating technical specifications and creating a concrete implementation of such a service, all aspects of the LTP Policy Template will have an impact and potential issues may surface.

The “DICE Digital Preservation Service” (DDPS), as the new service will be called, will enable dataset transfer from a B2SHARE repository to the LTP archive of DANS ("outsourcing" use case). This work continues to build upon work already done within the EOSC-Hub project and will also serve as proof of putting the LTP Policy Template to practice.

Starting points:

- The action of archiving of a dataset in the repository is completely done by the archiving facility
- The triggering of the archiving is done either by a user of the repository or the archiving facility using an automated process

Some issues that may occur and should be signaled by the LTP Template include:

- Triggering the archiving of a record:

  The repository must be able to trigger the archiving of a given record. This requires the design of an API that allows requesting this action.

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

- Archiving status:

  The repository must be able to determine the status of archiving for a given record. The status can be retrieved from the vault facility and displayed on the landing page of the record in the repository. An API request call must be available to do this.

- Metadata support for archiving status:

  The archiving status of a record must be available through the REST API of the repository itself, either with the exact status itself or a link that allows to request the status of archiving via the vault facility when for example the JSON representation of the dataset is requested.

- Mapping of metadata fields:

  Mapping metadata from the source system to the target system may cause problems. For instance, "License" type. Problems may arise if this is optional in the source system, but mandatory in the target system. The same applies to the use of different vocabularies that are not known in both systems. For example, "Discipline".

- Issuing persistent identifiers (PID):

  When a dataset is created in B2SHARE a DOI is minted and assigned to that dataset. However, the targeted archive will probably also issue a PID for the same dataset when ingested into the LTP archive. These PIDs must be updated in such a way that the original DOI or PID can be determined from the archived object's PID and vice-versa.

- Community policy:

  A repository might support communities that either enable or disable archiving of their datasets. This is done via community policies and if enabled, it must be able to configure the service access location and the mode of interaction (i.e. API or functional interface of the archiving facility).

- Limitation of access:

  Archiving facilities should limit their use by configuring the repository services that are allowed to make use of their functionality.

- Scalability:
  The archiving facility should have methods to limit the processing of datasets to avoid flooding the system with archiving requests.

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 Sensitive data risk analysis

This section describes the risk analysis for the University of Oslo Services for Sensitive Data (TSD) [51], CSC sensitive data services as infrastructure services [52], CINECA Openstack infrastructure services [53], the Lanikea cloud platform [54] at CINECA and the Secure B2SHARE [55] service deployments at TSD and CSC. The aim of the risk analysis is to investigate possible barriers, workarounds or impacts especially in non-certified data-sharing settings mainly for the deployment of an encrypted data analysis platform with backend HPC on the sensitive data services. TSD system risk analysis is publicly available [56].

5.1 Data processing in sensitive data e-Infrastructures

Table 7 describes security measures utilized by UiO/TSD and CSC sensitive data services in the context of Secure B2SHARE service.

Table 7 also describes responsibilities related to data processing as defined in General Data Protection Regulation (GDPR). For all of the measures, CSC and UiO/TSD act as data processors (the third-party entity that the data controller has chosen to use and process the data).

Table 7. Data processing and responsibilities

Security measure	Description	Responsibility
Encryption	CSC: Data is stored encrypted at on-premise cloud storage. TSD: Data is encrypted in transit to storage in on-premise cluster.	Data Processor: CSC Data Processor: TSD
Logical access control	CSC: Read access to dataset metadata is public. Write access requires authentication. Service administration, i.e. server access, happens only via intermediate servers. TSD: Access to data requires login with two-factor authentication.	Data Processor: CSC Data Processor: TSD
Traceability	CSC: Access logs are gathered TSD: logins, file import/export, access rights changes, etc., are logged	Data Processor: CSC Data Processor: TSD
Access intrusion	CSC: IDS systems are being used to monitor network access for service administration. TSD: physically isolated network with a very few hosts acting as gateways for entry. Tenants on separate VLANs.	Data Processor: CSC Data Processor: TSD
Infrastructure vulnerability	CSC: Unattended security related updates are distributed daily. Monthly service breaks for regular service and service dependency updates. Usual infrastructure administration processes are in place. TSD: Linux hosts have unattended upgrades. Windows hosts have weekly patching and downtime.	Data Processor: CSC Data Processor: TSD

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

Backup	TSD: backup on tape. Everything under /pXX/data/ is included in the backup except /pXX/data/no-backup CSC:	Data Processor: CSC Data Processor: TSD
Physical access control	CSC: Physical access to servers and hardware is strictly controlled to authorized persons. TSD: Physical access to servers is strictly controlled to authorized persons and all access attempts are logged.	Data Processor: CSC Data Processor: TSD

5.2 Data, processes and support services

This section describes generic activities (Table 8) that University of Oslo Services for Sensitive Data (TSD), CSC sensitive data clouds as infrastructure services and CINECA Openstack infrastructure services offer.

Table 8. Generic processes and services

Provision of a VM	• TSD: Virtual Machines are automatically created in one of several VMWare ESXi clusters by scripts that run on a timer. VMs are configured based on properties and labels specified in our host management system which is a UiO developed software component. • CINECA: manages the Openstack infrastructure. An openstack project tenant is allocated to users. Encrypted volumes and VMs are then created by users in self-provisioning via the Openstack dashboard. CINECA does not have the rights to access the VMs unless otherwise specified. • CSC: Open stack infrastructure. Encrypted volumes and VMs are then created by users in self-provisioning via the Openstack dashboard
Software installation	• TSD: Internal FileSystem repository • CSC: No user installable software. Software related to service operation installed from both internal and external software repositories. • CINECA: Users rely on the VM or volume internal FS repo to install the appropriate software stack for analysis
Upload of datasets	• TSD: REST API • CSC: REST API • CINECA: Users are in charge of uploading data on encrypted volumes exclusively via secure channels
Workflow execution	• TSD: Virtual CPUs or submission to a shared Slurm cluster (separate VLAN) • CSC: No workflow execution. • CINECA: Workflow is entirely executed within the VMs and encrypted volumes of the project
Production of results	• TSD: Forwarded to the project disk area • CSC: Results and outputs of externally executed workflow can be stored via REST API.

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

	• CINECA: Workflow generates the results and outputs
Data storing	• TSD: HNAS (long term), IBM ESS (HPC) • CSC: Data is encrypted and stored at on-premises cloud storage (CEPH based). • CINECA: results are stored in the project volumes. Backup storage is optional.

5.3 Risk Analysis for the deployment of Laniakea cloud platform[7]

Risks may be introduced upon the deployment of a cloud-based computing platform on the top of the sensitive data infrastructure. The potential risks are highly dependent on the actual implementation.

We refer to Table 7 in section 5.1 for the descriptions of roles and responsibilities.

5.3.1 Existing or planned measures to mitigate risks

Encryption

The Data Processor has an integrated system in place that was designed to treat sensitive data with encryption technologies. In particular, will be provided to the Data Controller a cloud service IaaS with the ability to self-provision encrypted volumes via LUKS[8] cyphring technology. The encryption key is handled by the Data Processor system and stored in a protected DB server on a separated intranet. All cloud service network traffic is encrypted via TLS protocol.

Partitioning

The data contains an ID known only by the Data Controller which can be combined with all personal data (personal and health) stored in the Data Controller premises.

Therefore, the data uploaded to the Data Processors systems for processing does not contain references to the identity of the individual person.

Logical access control

Whoever administers the system must provide for logical access control protection mechanisms and in particular:

- The consultation of data processed with electronic tools is allowed after the adoption of two-factor authentication systems based on the combined use of information known to users or of authenticated workstations;
- Password strength control mechanisms are provided;
- The user will be required to change the initial password at the first access;
- Passwords will have a validity period agreed with the Owner.

The Cloud IaaS dashboard is accessed via IdP (keycloak) providing a 2FA authentication mechanism.

Traceability

- The logs will track access to systems and operations performed
- The logs will be kept in an area not accessible to users to ensure their inalterability;

[7] https://laniakea-elixir-it.github.io/

[8] Linux Unified Key Setup is the reference technology for cyphring disks in Linux.

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

- Log files will be subject to periodic backup procedures;
- The log files will be kept for a period of 6 months;
- Access to log files will be allowed only to System Administrators

Data minimization

Each user will be able to upload their own raw data and will be able to install the software necessary for the execution of the data processing.

The uploaded data are only indispensable for the purpose of scientific research.

Access intrusion risks

Data protection mechanisms are provided against threats of intrusion on the physical infrastructure and the action of malicious programs on the Data Processor systems.

On the provided IaaS service, this action is the responsibility of the user / administrator.

Infrastructure vulnerability

The IT infrastructure that hosts the treatment is subject to periodic Vulnerability Assessment and Penetration Testing (VAPT).

On the provided IaaS service, this action is under the responsibility of the administrator / user.

Backup

The infrastructure that the Data Processor makes available is for a research project. For this purpose, the backup is not foreseen as a copy of the input data is kept on Data Controller premises.

Physical access control

Access to data rooms is limited to authorized personnel with personal badges and a video surveillance system is active. An electronic log of access is kept.

Security of communication protocols

The interaction with - and provision of - services that process non-public sensitive is protected via encrypted protocols (https, ldaps, imaps, etc).

5.3.2 Privacy protection management

Contract with the data processor

The appointment of the Data processor by the Data Controller has been defined as required by art. 28 of the GDPR, where the security measures required by the Data Controller are specified.

Privacy protection policy

To support the operational structures that must guarantee compliance with the provisions of the GDPR and adequate levels of IT security, the Data Processor has internally appointed the Data Protection Officer (DPO) and the Chief Information Security Officer (CISO).

Management of privacy protection policies

The Data Processor has adopted an “Organizational Model for the Protection of Personal Data” which describes the procedures, best practices and organizational responsibilities to guarantee the provisions of the GDPR.

Management of security incidents and personal data breaches

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 Data Processor has prepared an operational instruction for the treatment of events relevant to information security, with particular reference to violations of personal data. The procedure aims to:

- minimize the damage resulting from accidents involving information security
- define the operating procedures for the management of security incidents in order to guarantee an immediate and effective response
- monitor these events and learn from them
- document, through formal records, all actions taken to respond to security incidents
- establish the operating procedures and responsibilities relating to the communication from the Data Processor to the Data Controller and of a breach of personal data (data breach), in compliance with the timing defined in the GDPR
- establish the operating procedures and responsibilities relating to the communication to the individual person subject of a data breach relating to data owned by the Data Processor

The Data Processor has defined a CyberSecurity team trained on the procedure for managing the Data Breach.

5.3.3 Action plan

No further risk mitigation plans are envisaged because the level of risk is considered acceptable.

5.4 Risk analysis for secure-B2SHARE

Secure B2SHARE is a composed service for publishing datasets that contain sensitive data. Secure B2SHARE has three distinct components: B2SHARE, Secure Data Submission -service (SDS) and Authorization service. Dataset owner uploads files in SDS, creates datasets and describes metadata for the datasets in B2SHARE and manages authorization to datasets in Authorization service. Datasets created in B2SHARE always only refer to files previously uploaded through SDS. Files themselves are stored in Secure Storage.

Researchers can find datasets with the 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 representative) reviews the access request and either rejects or accepts it. If authorization is granted, Secure B2SHARE notifies Secure Storage that a specific person has been granted access to a specific dataset.

Besides applying general guidelines about 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.

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

Figure 10 describes a layered architecture view of the Secure B2SHARE deployed on TSD specific infrastructure and CSC specific infrastructure. B2SHARE service is used as a UI and metadata store. Users don't need to authenticate in order to view and search for dataset metadata, but for all other actions (creation of datasets, upload of files, authorization requests, etc.) a user must be authenticated. Users are authenticated with B2ACCESS for CSC's Secure B2SHARE deployment and with TSD Auth for TSD's Secure B2SHARE deployment. Other authentication services supporting OpenID Connect (OIDC) or Security Assertion Markup Language 2.0 (SAML

2.0) protocols could be used. Authenticated users can upload data via Secure File Transfer Protocol (SFTP) for TSD's Secure B2SHARE deployment and via HTTPS protocol for CSC's Secure B2SHARE deployment. In CSC's Secure B2SHARE deployment, authorization requests for file access are made through Resource Entitlement Management Service (REMS) and in TSD's Secure B2SHARE deployment through Nettskjema. Other similar services could be used. Secure Storage component of Secure B2SHARE is realized with ePouta [57] component of CSC secure cloud service in CSC specific Secure B2SHARE deployment and by TSD storage system in TSD specific Secure B2SHARE deployment.

Acronyms used in Figure 10:

- B2ACCESS: authorisation and authentication proxy
- TSD Auth: TSD Authentication service
- REMS: Resource Entitlement Management System
- NettSkjema: Web Forms service at UiO
- ePouta: CSC secure cloud service
- TSD: UiO Services for Sensitive Data

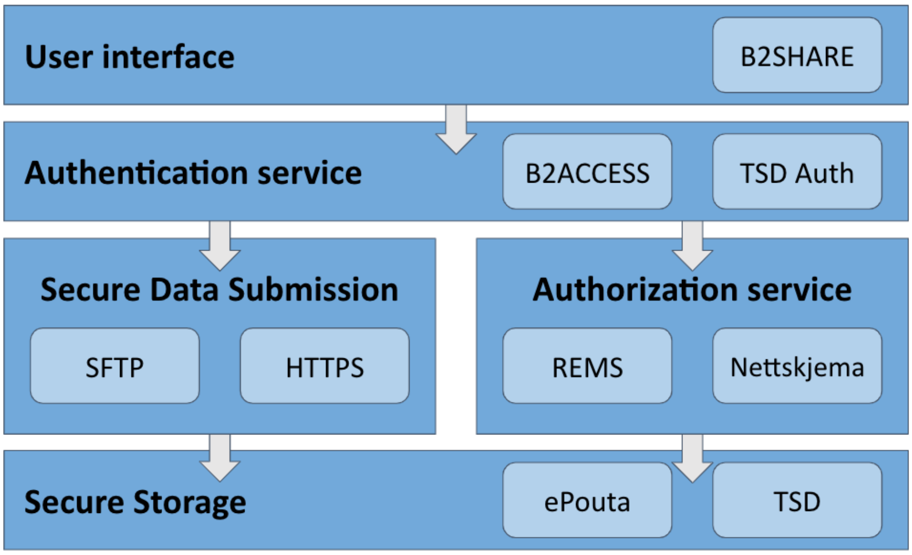


Figure 10: Secure B2SHARE layered architecture diagram.

Risk analysis of Secure B2SHARE deployment on CSC and TSD premises is based on the WISE risk management template (https://wise-community.org/risk-assessment/). The information in the assessment sheets are described as follows:

- Instructions for WISE security risk assessment
- Two approaches are described: Threat centric vs Asset centric
- Assessment for the threat centric approach applied to Secure B2SHARE
- Assessment describes identified risks, existing mitigation activities, impact of identified risks, still existing weakness of each risk and impact, likelihood and overall score of each risk.

The two assessment sheets are:

- Secure B2SHARE@TSD: https://b2drop.eudat.eu/s/iZ3dCzpCpkWL79n
- Secure B2SHARE@CSC: https://b2drop.eudat.eu/s/4kS72prGCDQMNN4

6 Conclusions

The intermediate results of the four tasks of the DICE WP4 are of use beyond the DICE project itself and the current deliverable.

The “Pilot use cases for the integration of data services with computing platforms” of task 4.1 are being used for the interaction of DICE services with several platforms, e.g. Fenix and EuroHPC.

The work of task 4.2 on the integrity check of the PID infrastructure and the PID metadata allows for proper PID resolution and usage of the content of PID records. Together with their (T4.2) work on the usage of datatypes these contributions are used in RDA working groups, e.g. the Persistent Identification of Instruments WG.

Task 4.3 delivered a Long-Term Preservation (LTP) Policies Template, that can be used by a wide range of repositories and policy-based data archives to compose their LTP policies. This contribution will be feed in the EOSC task force on long-term data preservation as input for the horizontal EOSC preservation policy.

The risk analysis of task 4.4 on sensitive data will be provided and discussed with other providers for sensitive data management in EOSC.

The next and final deliverable of WP4 is due in project month 30 and will have again contributions from all four tasks. Its name is D4.3 “Final integration with other services & platforms” and it will inform about the:

- final integration of data services with computing platforms (T4.1)
- integration of PID Graph resources in B2FIND (T4.2)
- implementation of the LTP policy for B2SHARE in one CTS certified archive (T4.3)
- enabling of sensitive data workflow by adapting standard interoperability frameworks to connect the endpoints (T4.4).

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

7 References

[1] [Online]. Available: https://laniakea-elixir-it.github.io/.

[2] „EUDAT B2DROP,“ [Online]. Available: https://eudat.eu/services/userdoc/b2drop.

[3] „EUDAT B2SAFE,“ [Online]. Available: https://eudat.eu/services/userdoc/b2safe.

[4] „EUDAT B2SHARE,“ [Online]. Available: https://eudat.eu/services/userdoc/b2share.

[5] [Online]. Available: http://www.webdav.org.

[6] [Online]. Available: https://b2drop.eudat.eu/remote.php/webdav.

[7] [Online]. Available: https://jupyter-jsc.fz-juelich.de.

[8] [Online]. Available: https://github.com/miquels/webdavfs.

[9] [Online]. Available: https://davix.web.cern.ch/davix/docs/devel/.

[10] „iRODS Technical Overview,“ 2016. [Online]. Available: https://irods.org/uploads/2016/06/technical-overview-2016-web.pdf.

[11] [Online]. Available: https://gitlab.com/noumar/http-api/-/blob/master/DESCRIPTION.md.

[12] [Online]. Available: https://fasterdata.es.net/data-transfer-tools/gridftp/.

[13] „https://research.csc.fi/en/-/mahti,“ [Online]. Available: https://research.csc.fi/en/-/mahti.

[14] „https://research.csc.fi/en/-/puhti,“ [Online]. Available: https://research.csc.fi/en/-/puhti.

[15] [Online]. Available: https://docs.csc.fi/support/faq/how-to-move-data-between-puhti-and-allas/.

[16] [Online]. Available: https://rclone.org.

[17] [Online]. Available: https://openid.net/connect/.

[18] [Online]. Available: https://trng-b2share.eudat.eu.

[19] S. R. S. a. L. L. Sun, "Handle System Namespace and Service Definition," *RFC 3651, DOI 10.17487/RFC3651, https://www.rfc-editor.org/info/rfc3651,* 11 2003.

[20] [Online]. Available: https://dona.net/.

[21] [Online]. Available: https://pidconsortium.net.

[22] „ePIC "Quality of Service and Policies",“ [Online]. Available: https://www.pidconsortium.net/?page_id=904.

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

[23] „DICE Project," [Online]. Available: https://www.dice-eosc.eu/call-service-requests.

[24] C. Göksenin, „PID Statistics Project," *Göttingen Research Online / Data, V1,* Nr. DOI:10.25625/G5PCVI, 2022.

[25] „IP Geolocation API," [Online]. Available: https://ip-api.com/.

[26] L. B. D. &. M. G. Lannom, "RDA Data Type Registries Working Group Output," in *https://doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458*, 2015.

[27] [Online]. Available: https://dtr-pit.pidconsortium.net.

[28] "RDA PIDINST WG (Recommendation in progress)," [Online]. Available: https://www.rd-alliance.org/groups/persistent-identification-instruments-wg.

[29] [Online]. Available: https://www.w3.org/TR/json-ld/.

[30] [Online]. Available: https://www.uvicorn.org.

[31] [Online]. Available: https://gunicorn.org.

[32] [Online]. Available: https://www.iso.org/standard/57284.html.

[33] [Online]. Available: http://www.oais.info/.

[34] [Online]. Available: https://www.coretrustseal.org/.

[35] [Online]. Available: https://www.4cproject.eu/.

[36] [Online]. Available: https://www.go-fair.org/fair-principles/.

[37] [Online]. Available: https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re.

[38] [Online]. Available: : http://www.interpares.org.

[39] [Online]. Available: https://www.dpconline.org/docs/miscellaneous/training/1719-mp-policy/file.

[40] [Online]. Available: https://public.ccsds.org/pubs/650x0m2.pdf.

[41] „CCSDS, Recommendation for Space Data System Practices. Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2, Purple Book," CCSDS Secretariat (NASA), Washington, DC, 06 2012. [Online]. Available: https://public.ccsds.org/pubs/650x0m2.pdf.

[42] „CSDS, Recommendation for Space Data System Practices. Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2, Magenta Book," CCSDS Secretariat (NASA), Washington, DC, 10 2020. [Online]. Available: https://public.ccsds.org/Lists/CCSDS%206500P21/650x0021.pdf.

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

[43] [Online]. Available: https://www.archivematica.org/en/docs/archivematica-1.13/user-manual/preservation/preservation-planning/.

[44] [Online]. Available: https://dans.knaw.nl/en/preservationplan/.

[45] [Online]. Available: https://dam.data-archive.ac.uk/controlled/cd062-preservationpolicy.pdf.

[46] [Online]. Available: https://eudat.eu/services/userdoc/b2share-faq#How_many_files_can_I_deposit_or_records_can_I_create_in_B2SHARE.

[47] [Online]. Available: https://www.eudat.eu/contact-support-request.

[48] [Online]. Available: https://zenodo.org/record/3638211.

[49] [Online]. Available: https://www.4cproject.eu/summary-of-cost-models/16-community-resources/outputs-and-deliverables/108-cost-model-for-digital-preservation-cmdp/.

[50] [Online]. Available: http://www.4cproject.eu.

[51] [Online]. Available: https://www.iso.org/standard/57284.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.

APPENDIX 1: LTP Policy Template

Long-Term Preservation Policy Template

For EUDAT Services (in particular B2SHARE and B2SAFE)

Contents

1. Objectives, scope and delimitation of this policy

1.1. The general aim of this preservation policy is to ensure the authentic and trustworthy preservation and accessibility of digital research data and related outputs, henceforth referred to as “datasets”, for reuse in the long term.

1.2. The specific aims of the preservation policy are to:

- □ Provide authentic and reliable instances of datasets to researchers;
- □ Maintain the integrity and quality of the datasets;
- □ Ensure that datasets are managed throughout their lifecycle (e.g. when migrations or changes in metadata are carried out) in the medium that is most appropriate for the task they perform;
- □ Ensure that the relevant level of information security is applied to each dataset;

1.3. The scope of this policy is limited to [LTP Institution]’s service for long-term preservation, [LTP Service]. It applies exclusively to datasets held in [LTP Service]. This policy does not consider preservation of other materials, such as [LTP Institution]’s web pages, internal and external documents, and digital objects in any other services the [LTP Institution] provides.

1.4. The [LTP Institution] assumes responsibility for the long-term preservation and accessibility of datasets ingested (by individual deposits or bulk data transfers) into its [LTP Service].

1.5. Long-term preservation and providing sustained access to datasets fits within the remit and mission of [LTP Institution].

1.6. The length of the preservation period is … <specify the preservation period in years or “indefinite”>, unless legal obligations prevent this (as described in article 4.7).

1.7. [LTP Service] is designed to be understandable and useful for the designated communities it serves. This community (or communities) of the [LTP Service] consist(s) of … <specify the target audience(s), such as scientific domains, research communities, or research in general>.

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. Ingest

This section distinguishes two situations, to which partly different articles apply:

A. [LTP Service] is provided by the data service itself. In this situation, the ingest takes place via deposits by researchers or research communities into the data service (e.g. B2SHARE), which also acts as [LTP Service]. Researchers or research communities are the depositors.

B. [LTP Service] is provided by an external organisation / separate [LTP Institution]. In this situation, the ingest consists of (bulk) data transfers from the data service (e.g. B2SHARE) to an external [LTP Service]. The data service acts as the depositor on behalf of the researchers and research communities, who originally deposited the datasets. A Long-Term Preservation Agreement (LTP Agreement) will be necessary, a template for which is available as a separate document.[9]

Note that the choice between situation A and B also has repercussions for the ways in which the access to datasets will be arranged (see section 7).

2A. [LTP Service] is provided by the data service itself

2.1. The datasets, including metadata as supplied by the depositor [= researcher or research community], are considered by [LTP service] as the Submission Information Package (SIP) in OAIS terms. This is the "original‟ information to be stored for long-term preservation.

2.2. The datasets are professionally cataloged according to the metadata standard used by [LTP Service].

2.3. [LTP Service] supports a list of licenses for open data sharing, from which a selection can be made by the depositor during ingestion. A list of supported licenses is available in Appendix 4.

2.4. [LTP Service] provides a Persistent Identifier [PID] to each ingested [dataset / file / object] <indicate to which units the PID apply> for long-term reference to their location.

2.5. The quality control of the metadata is maintained as follows

- ☐ The datasets that the [LTP Service] ingests are accompanied by metadata as supplied by the [Depositor], which should be adequate to enable the designated communities to understand and reuse the content for analytical and research purposes.
- ☐ The deposited datasets (including metadata) are checked and validated by [LTP service] according to documented data ingest procedures, including the following checks <indicate what applies>:
    - ☐ Virus scans
    - ☐ Completeness of data (e.g. checksums)
    - ☐ Metadata and additional documentation
    - ☐ Compliance with preferred / accepted formats
    - ☐ Organisation of data (data structure), reformatting
    - ☐ Presence of personal data
    - ☐ Data cleaning
    - ☐ Other: … <specify>

[9] Currently, this document (i.e. the LTP Agreement template) has the status of draft, as it may be subject to changes once the technical implementation of the data transfer between the repository service (B2SHARE) and the [LTP Service] is defined, and this deliverable is due only at a later stage in the project.

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

- □ In order to guarantee minimal metadata quality, only records that comply with the following basic Dublin Core Metadata Terms[10] can be successfully ingested into [LTP Service] <indicate which terms minimally apply>:
  - □ Title
  - □ Creator
  - □ Description
  - □ Date (created)
  - □ Rights
  - □ Audience
  - □ … <specify more terms if applicable>
- □ The completeness and accuracy of the metadata accompanying the deposited dataset is the responsibility of the depositor.
- □ [LTP Service] may notify the depositor of deficiencies and inaccuracies in the metadata (during or after ingestion).
- □ Alternatively, if datasets or metadata contain deficiencies, [LTP Service] can make alterations post ingest. Hereby distinction is made between minor and major alterations to digital assets:
  - Major: when a digital object in a dataset is changed, this will result in a new version and therefore a new dataset with a new PID in [LTP service]. The new and the old version are cross-referenced in their respective descriptive metadata, and the new version will be the default version for access.
  - Minor: when there is a change (addition or edit) in the metadata, descriptive documents or supplementary files, this is documented in the (administrative) metadata, no new dataset is created and no new persistent identifier is minted.

2B. [LTP Service] is provided by an external organisation / separate [LTP Institution]

2.6. The datasets, including metadata as supplied by depositor [= data service], are considered by [LTP service] as the Submission Information Package (SIP) in OAIS terms. This is the “original‟ information to be stored for long-term preservation in the [LTP Service].

2.7. The datasets are professionally catalogued according to the metadata standard used by [LTP Service].

2.8. Licenses for open data sharing attributed to the datasets are ingested “as is”, and are preserved and supported unchanged by [LTP service]. A list of supported licenses is available in Appendix 4.

2.9. Persistent Identifiers [PIDs] minted by the [data service] are ingested into the [LTP Service] “as is”, and continue to refer to the original location of the data (see section 7 on Access for changes in the case [data service] ceases to exist).

2.10. [LTP Service] will allocate additional PIDs for reference to archived copies of the data sets in [LTP Service] (see section 7 on Access for alternatives in referencing for access to datasets).

[10] See DCMI terms: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

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

2.11. The quality control of the metadata is maintained as follows:

- □ The datasets that the [LTP Service] ingests are accompanied by metadata as supplied by the [Depositor], which should be adequate to enable the designated communities to understand and reuse the content for analytical and research purposes.
- □ The deposited datasets (including metadata) are checked and validated by [LTP service] according to documented data ingest procedures, including the following checks <indicate what applies>:
  - □ Virus scans
  - □ Completeness of data (e.g. checksums)
  - □ Metadata and additional documentation
  - □ Compliance with preferred / accepted formats
  - □ Organisation of data (data structure), reformatting
  - □ Presence of personal data
  - □ Data cleaning
  - □ Other: … <specify>
- □ Metadata fields of the [data service] are mapped with the fields of the [LTP Service] metadata schema:
  - Corresponding fields (from the same metadata scheme) are copied
  - Similar fields (from different metadata schemes) are mapped wherever possible
  - Fields that cannot be mapped are stored with the dataset as additional documentation (in a separate metadata blob file).
- □ In order to guarantee minimal metadata quality, only datasets that comply with the following basic Dublin Core Metadata Terms[11] can be successfully ingested into [LTP Service] <indicate which terms minimally apply>:
  - □ Title
  - □ Creator
  - □ Description
  - □ Date (created)
  - □ Rights
  - □ Audience
  - □ … <specify more terms if desired>
- □ The completeness and accuracy of the metadata ingested from [data service] into [LTP Service] is the responsibility of the [data service].
- □ [LTP Service] may notify the [data service] of deficiencies and inaccuracies in the metadata (during or after ingestion).
- □ Alternatively, if datasets or metadata contain deficiencies, [LTP Service] can make alterations post ingest. Hereby distinction is made between minor and major alterations to digital assets:
  - Major: when a digital object in a dataset is changed, this will result in a new version and therefore a new dataset with a new PID in [LTP service]. The new and the old version are cross-referenced in their respective descriptive metadata, and the new version will be the default version for access.
  - Minor: when there is a change (addition or edit) in the metadata, descriptive documents or supplementary files, this is documented in the (administrative) metadata, no new dataset is created and no new persistent identifier is minted.

[11] See DCMI terms: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

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. Archival Storage

3.1. The ingested datasets (data and metadata) from B2SHARE constitute the basis for the Archival Information Package (AIP) in OAIS terms. Archival actions related to the custody of AIPs, which may affect the chain of provenance, will be documented, including changes in <indicate what applies>:

☐ … the archival system (upgrades, new systems)
☐ … documentation standards
☐ … license definitions
☐ … file formats (format conversions and updates)
☐ … storage location (migrations to other facilities)
☐ … other <specify other archival actions influencing the chain of provenance>

3.2. The archival storage function receives AIPs from the ingest function and adds them to the permanent storage facility, and oversees the management of this storage, including media monitoring and refreshment. This ensures that AIPs will be retrievable and can be disseminated over time.

3.3. [LTP Service] does not assume responsibility for the preservation of data that are hosted externally, unless this is specifically agreed with an external storage provider, to be specified in an agreement (as in article 3.5).

3.4. The [LTP Institution] is committed to take all necessary precautions to ensure the physical integrity and security of the datasets stored in its [LTP Service], including <indicate which measures apply>:

☐ Periodic technology-vulnerability scans
☐ SLAs with external IT providers, e.g. for data storage
☐ Procedures for file fixity checking (checksums), verifying that data have not been altered or corrupted
☐ A confidentiality statement for [LTP Service] staff (as described in article 6.6.4)
☐ Periodic security audits
☐ Other measures: … <specify>

3.5. [LTP Institution] may outsource the physical data storage to an external storage provider, provided it has a contract and accompanying Service Level Agreement (SLA) with its provider, which include <indicate what applies>:

☐ Periodic technology-vulnerability scans
☐ SLAs with external IT providers, e.g. for data storage
☐ Procedures for file fixity checking (checksums), verifying that data have not been altered or corrupted
☐ A confidentiality statement for [LTP Service] staff (as described in article 6.6.4)
☐ Periodic security audits
☐ Other measures: … <specify>

4. Data Management

4.1. [LTP Service] maintains a catalogue of metadata which uses a standard schema for describing its contents. This database is searchable by internal and external finding aids.

4.2. The Archival Information Package includes administrative metadata which support archival operations, including change or version control of datasets.

4.3. [LTP Institution] guarantees the timeliness, authenticity and integrity of the datasets in its [LTP Service] for future access and reuse.

- □ The datasets (and metadata) ingested (including exact and integral copies thereof) are considered as „original" and authentic.
  - ● If preservation and/or dissemination copies in derived formats are made (e.g. preserved formats or lower resolution video formats for web viewing), these will be stored alongside the originals.
  - ● Provenance information is supplied for all copies in preservation and dissemination formats that can be traced back to the original datasets as ingested.
- □ Authenticity and integrity of Archival Information Packages (AIPs) imply that once a dataset is ingested into the [LTP Service], the AIP cannot be changed or removed by the user.
  - ● Only authorized staff of [LTP Institution] have the right to modify the format and/or functionality of an archived object, if this is deemed necessary to facilitate the digital sustainability, distribution or re-use of the information it contains.
  - ● [LTP Institution] ensures that any alteration to the preserved version of any part of a dataset will only take place under controlled conditions and will be accurately documented. Hence, the content stored in [LTP Service] is of a static nature.
  - ● When datasets are updated, changed or extended, the resulting new versions are considered and treated as new assets.
- □ The chain of custody of datasets archived in the [LTP Service] is documented through metadata, which guarantee that all archival actions are explicit, complete, correct and current.

4.4. [LTP Service] maintains (or conforms to) a documented list of <indicate what applies>:.

- □ … preferred or archival formats, about which [LTP Service] is dedicated to offer long-term guarantees in terms of future findability, accessibility and reusability[12].
- □ … accepted formats, which are preserved in the format as ingested, for which the findability and accessibility are guaranteed, but for which the future reusability can not be guaranteed.

4.5. [LTP Service] monitors the potential obsolescence of file formats and has conversion policies and procedures in place to take action when required.

[12] The Interoperability of datasets depends on characteristics of the data that are partly beyond the control of [LTP service] and can therefore not be guaranteed by [LTP Service]. The Interoperability of research data is the responsibility of the researcher / research community.

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

4.6. In case of conversion of objects to another (preferred or archival) file format (e.g. for preservation or access purposes), the [LTP Service] will also maintain the original file(s) <indicate what applies>:

- □ Files stored in the [LTP Service] from outdated formats are migrated to successor formats, and the archival metadata is updated accordingly.
- □ Files in outdated formats are preserved to maintain the chain of provenance.

4.7. Datasets or files that are archived, published and/or exposed in the [LTP Service] can only be deleted or made inaccessible by [LTP Institution]'s authorized staff for compelling reasons:

- □ If the content is unlawful.
- □ If there is a legally binding maximum preservation period for the content.
- □ If personal data appears to be preserved without permission of research subjects.
- □ If the content consists of personal data, and a research subject rightfully objects to the preservation of a digital object with an appeal to the GDPR (e.g. right to be forgotten; or revocation of informed consent).
- □ Other compelling reasons, to be decided by the director of the [LTP Institution].

4.8. In case a data object needs to be deleted, this is recorded in the metadata, which will continue to exist as a "tombstone" record.

5. Administration[13]

5.1. [LTP Institution] monitors the operations of its [LTP Service] and publishes periodic reports on its performance. A monitoring schedule is provided in Appendix 3.

5.2. [LPT Institution] develops a multi-annual strategy which includes the strategic direction and upgrading of the [LPT Service].

5.3. The planning and control cycle of [LTP Institution] is overseen by … <specify how or by whom the planning and control cycle is monitored>.

5.4. The strategy and functioning of [LTP Institution] is subject to periodic external evaluations, the reports of which are available publicly / on request / otherwise: … <specify what applies>.

5.5. A(n) (scientific) advisory board consisting of prominent members of the designated community advises [LPT Institution] on its strategic course and development.

5.6. [LTP Institution] maintains functions for providing customer support.

5.7. [LPT Institution] maintains a publicly available list and description of legal and statutory regulations which apply (see Appendix 5. for an overview of generally applicable European and national laws and regulations; select which specific regulations apply):

- □ General Terms and Conditions of Use
- □ Data Deposit Agreement
- □ Long-Term Preservation Agreement
- □ Data Processing Agreements
- □ User licenses
- □ Privacy policy
- □ Liability statement
- □ Ownership rights statement
- □ Other applicable regulations: … <specify>

[13] Note that several Administrative functions of the OAIS model are dealt with in other sections of this policy:

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. Preservation Planning

6.1. [LTP Institution] monitors the contents of its [LTP Service] and periodically publishes metrics on key performance indicators.

6.2. [LTP Institution] reviews its preferred and/or accepted formats list and guidelines periodically for obsolescence and completeness, updating and extending them as needed.

6.3. [LTP Institution] performs the following other monitoring functions for preservation planning <indicate which monitoring activities apply>:

- □ Monitoring security risks, in order to anticipate and control such risks.
- □ Technology watch, in order to recommend adaptations in its technology environment, in particular its archival system(s).
- □ Monitoring changes in service requirements of its Designated Communities.
- □ Other monitoring activities: … <specify>

6.4. Roles and responsibilities, confidentiality, limited liability <indicate what applies>:

- □ The Director of [LTP Institution] is responsible for maintaining this policy.
- □ The staff of [LTP Institution] implements this policy as appropriate to their roles and responsibilities.
- □ Preservation decisions about the [LPT Service] are made within the context of the [LPT Institution]'s mission and strategy, balancing the constraints of costs, scholarly value, user accessibility, and legal admissibility.
- □ [LTP Institution]'s staff, including temporary staff, trainees, (visiting) fellows and volunteers, are accountable for keeping confidentiality when processing digital assets stored in [LTP Service], in particular personal data, in any way whatsoever, by signing a confidentiality statement.
- □ This policy will be evaluated and, if necessary, will be revised every … years <specify the evaluation and revision cycle>, or as soon as security threats, changes in technology or legal and statutory context require to do so.
- □ [LTP Institution] is not liable for the contents, including errors, of research data ingested into its [LTP Service], nor for (possible errors in) the metadata and additional documentation associated with those datasets. [LTP Institution] is also not liable for incorrect inferences resulting from the analysis of those datasets.

6.5. Funding and contingency plan:

- □ [LTP Institution] declares it receives adequate funding to fulfil its mission, which includes the maintenance of the [LTP Service].
- □ In case of dissolution of the [LTP Institution], it has a continuity plan in place which guarantees that a successor organization will take over the care of the [LTP service][14].

[14] This is in conformance to the mandatory OAIS requirements and CTS requirement 3 "Continuity of access: The repository has a continuity plan to ensure ongoing access to and preservation of its holdings". For example, in the case of DANS: The NWO-KNAW *Samenwerkingsovereenkomst DANS* (Collaboration Agreement DANS) 2015 explicitly states that, in the case of discontinuity of DANS, NWO and KNAW will take over the responsibility for the digital assets archived at DANS and store these elsewhere "in the most responsible manner possible and under equivalent technical conditions" (article 10.6 of the Collaboration Agreement).

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

7. Access

As in the Ingest section (2), this section distinguishes two different situations, to which partly different articles apply:

A. [LTP Service] is provided by the [data service] itself. In this situation, the access takes place through access mechanisms of the [Data Service].

B. [LTP Service] is provided by an external organisation / separate [LTP Institution]. In this situation, the [data service] needs to indicate in the LTP Agreement whether it wants the datasets in [LTP Service] to be exposed (= to be findable and/or accessible) for end users via the [LTP Service] or not.
If for any reason the original [data service] is discontinued, the access function to the datasets of [data service], as ingested and preserved by [LTP Institution], will be taken over by the [LTP Service]. In this situation two extra articles apply (7.6 and 7.7).

7.1. [LTP Service] provides the following access functions <indicate what applies>:

- ☐ A user interface for browsing and searching its content
- ☐ An API using an open and universal protocol (e.g. OAI-PMH: Open Archives Initiative: Protocol for Metadata Harvesting)
- ☐ Other: … <specify>

7.2. Metadata in [LTP Service] are always openly accessible, i.e. without copy- or database-rights and without authentication or authorisation of users (either through the user interface or via the API).

7.3. Users need to register (create an account) and to login to the [LTP Service] if the license to view and download files requires them to do so (see Appendix 4).

7.4. Authentication by registration and logging in is not obligatory for viewing or downloading data with licences equivalent to full Open Access (CC0 Waiver).

7.5. In order to enable reuse, data released from [LTP Service] are always accompanied by <indicate what applies>:

- ☐ A clear conditions of use statement conforming to the applicable licence or CC0 waiver.
- ☐ A standard citation recommendation to encourage proper data citation.
- ☐ Other: … <specify>

Article 7.6 and 7.7 only apply in case [LTP Service] is provided by an external organisation / separate [LTP Institution] as described under situation B in the text box above.

7.6. By default, as long as the [data service] exists and its content is accessible for end users, the standard PID assigned by [data service] to a dataset archived in the [LTP Service] refers to the location in the original [data service].

7.7. In case the [data service] ceases to exist and becomes inaccessible for end users, the access to the datasets stored at the [LTP Service] will become findable and accessible via the search and download functions of [LTP Service], under similar access conditions and licenses as in the original [data service].

APPENDIX 2: Summary of the OAIS Reference Model

The EUDAT B2SHARE preservation policy follows the recommendations of the Open Archival Information System (OAIS) reference model. In spite of the fact that the report in which the OAIS model is described is called "recommended practice", it nevertheless contains a section of "mandatory responsibilities" that an organization wishing to operate as an OAIS Archive must fulfil:

Mandatory responsibilities

An OAIS Archive shall:

- Negotiate for and accept appropriate information from information Producers.
- Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.
- Determine, either by itself or in conjunction with other parties, which entities should become the Designated Community, that is, the communities that should be able to understand the information provided. Definition of the Designated Community includes a determination of their Knowledge Base.
- Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.
- Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Content Information with evidence supporting its Authenticity.

The reference model

The OAIS reference model is depicted in Figure 11. Our description of the six functional entities is based on this OAIS model, from which the definitions and summaries below have been derived[15].

[15] CCSDS, Recommendation for Space Data System Practices. Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2, Magenta Book [also known as "Purple Book"], CCSDS Secretariat (NASA), Washington, DC, June 2012.

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

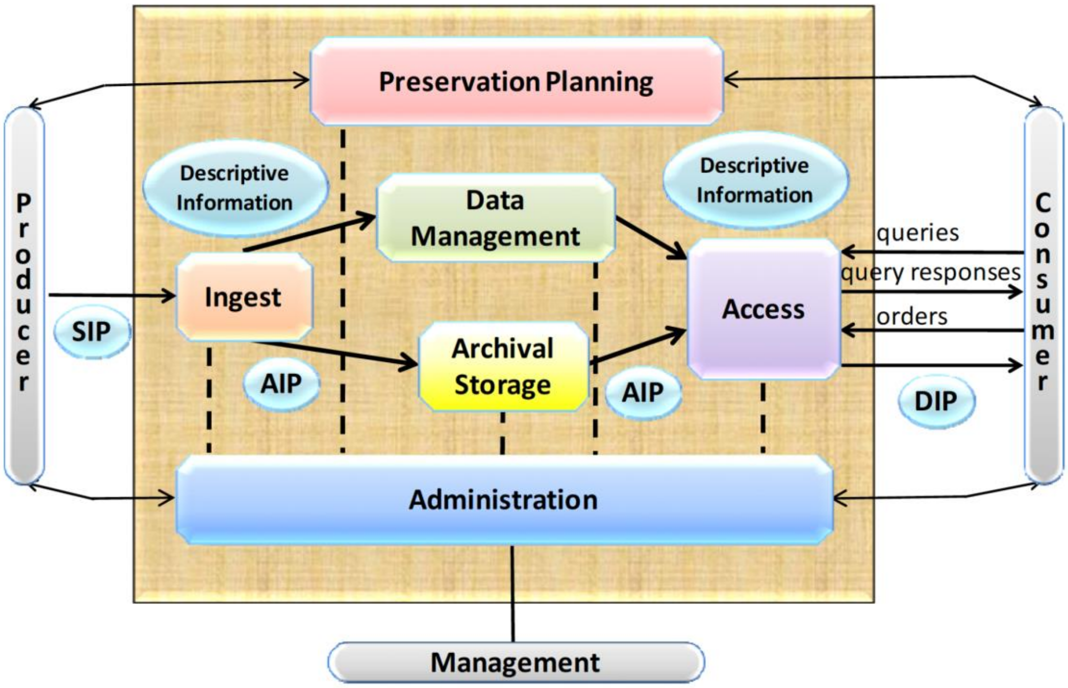


Figure 11. OAIS Functional Entities

Selected OAIS Definitions (slightly adapted for use in this document):

The definitions below are taken from the OAIS ”Purple Book”, with minor modifications or clarifications to fit the purposes of this document[16].

- **Archive**: An organization that intends and has accepted the responsibility to preserve information for access and use by a Designated Community. The people and systems working for the Archive may be part of a larger organization. If an Archive meets the responsibilities recommended by the OAIS reference model, it can be called an “OAIS Archive”. The term ‘Open’ in OAIS is used to imply that the recommendations and standards are developed in open forums, and it does not imply that access to the Archive is unrestricted. **Note:** in this document [LTP Service] is synonymous to “(OAIS) Archive”, whereas [LTP Institution] is the larger organisation to which the [LTP Service] belongs.
- **(Data) Producer**: The role played by those persons or client systems that provide the information to be preserved.
- **(Data) Consumer**: The role played by those persons, or client systems, who interact with archival services to find preserved information of interest and to access that information in detail.

[16] CCSDS, Recommendation for Space Data System Practices. Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2, Magenta Book [commonly referred to as the “Purple Book”], CCSDS Secretariat (NASA), Washington, DC, June 2012: https://public.ccsds.org/pubs/650x0m2.pdf; update October 2020 (draft, also called the “Pink Book”): https://public.ccsds.org/Lists/CCSDS%206500P21/650x0021.pdf.

- **Submission Information Package (SIP)**: An Information Package that is delivered by the Producer to the Archive for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.
- **Archival Information Package (AIP)**: An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an Archive.
- **Dissemination Information Package (DIP)**: An Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the Archive.
- **Preservation Description Information (PDI)**: The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, Context, and Access Rights Information.
- **Designated Community**: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time.
- **Authenticity**: The degree to which a person (or system) regards an object as what it is purported to be. Authenticity is judged on the basis of evidence.
- **Dissemination Information Package (DIP)**: An Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the Archive.
- **Long Term**: A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in an Archive. This period extends into the indefinite future.
- **Long Term Preservation**: The act of maintaining information, independently understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term.

OAIS Functional entities

The roles provided by each of the functional entities in Figure 4 are summarized as follows:

Ingest

The Ingest Functional Entity provides the services and functions to receive Submission Information Packages (SIPs) from Data Producers (or in this case: from the EUDAT B2SHARE repository) and to prepare the contents for storage and management within the Archive. Ingest functions include:

- receiving SIPs
- Performing quality assurance on SIPs
- Generating Archival Information Packages (AIP) which comply with the Archive's data formatting and documentation standards
- Extracting Descriptive Information from the AIPs for inclusion in the Archive metadatabase
- Coordinating updates to Archival Storage and Data Management.

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

Archival Storage

The Archival Storage Functional Entity provides the services and functions for the storage, maintenance and retrieval of AIPs. Archival Storage functions include:

- Receiving AIPs from Ingest and adding them to permanent storage
- Managing the storage hierarchy
- Refreshing the media on which Archive holdings are stored
- Performing routine and special error checking
- Providing disaster recovery capabilities
- Providing AIPs to Access to fulfill orders.

Data Management

The Data Management Functional Entity provides the services and functions for populating, maintaining, and accessing both Descriptive Information which identifies and documents Archive holdings and administrative data used to manage the Archive. Data Management functions include:

- Administering the Archive database functions (maintaining schema and view definitions, and
  referential integrity)

- Performing database updates (loading new descriptive information or Archive administrative data)
- Performing queries on the data management data to generate query responses
- Producing reports from these query responses.

Administration

The Administration Functional Entity provides the services and functions for the overall operation of the Archive system. Administration functions include:

- Soliciting and negotiating submission agreements with Producers
- Auditing submissions to ensure that they meet Archive standards
- Maintaining configuration management of system hardware and software.
- Monitoring and improving Archive operations
- Making inventories and reporting on the contents of the Archive
- Migrating/updating the contents of the Archive.
- Establishing and maintaining Archive standards and policies
- Providing customer support
- Activating stored requests.

Preservation Planning

The Preservation Planning Functional Entity provides the services and functions for monitoring the environment of the Archive, providing recommendations and preservation plans to ensure that the information stored in the Archive remains accessible to, and understandable by, the Designated Community over the Long Term, even if the original computing environment becomes obsolete. Preservation Planning functions include:

- Evaluating the contents of the Archive and periodically recommending archival information updates
- Recommending the migration of current Archive holdings
- Developing recommendations for Archive standards and policies
- Providing periodic risk analysis reports
- Monitoring changes in the technology environment
- Monitoring changes in the Designated Community’s service requirements and Knowledge Base
- Designing Information Package templates, providing design assistance and review to specialize these templates into SIPs and AIPs for specific submissions
- Developing detailed Migration plans, software prototypes and test plans to enable implementation of Administration migration goals.

Access

The Access Functional Entity provides the services and functions that support Consumers in determining the existence, description, location and availability of information stored in the Archive, and allowing Consumers to request and receive information products. Access functions include:

- Communicating with Consumers to receive requests
- Applying controls to limit access to specially protected information
- Coordinating the execution of requests to successful completion
- Generating responses (Dissemination Information Packages, query responses, reports) and delivering the responses to Consumers.

Common Services

In addition to the functional entities described above, according to the OAIS reference model there are various Common Services (not shown in the diagram) assumed to be available. They constitute the computing environment in which the Archive functions and have a supporting role. Think of the Operating System (providing basic functions such as naming and directory services) and the Computer Network (providing e.g. communication and file transfer functions), but also Security and Backup services are reckoned to be in this category.

APPENDIX 3: Recurring monitoring processes

This is an overview of processes in the [LTP Institution], which contribute to Preservation Planning, monitoring community, technology, legal or strategic developments, and risks.

Nr.	Process	Frequency	Responsible
1	Monitor the [LTP Service]'s designated communities for developments that may affect the [LTP Service], such as – requested – changes in technologies or file formats that communities use. This is done in contacts with the communities, e.g. during data acquisition, collaboration projects, membership of European Research Infrastructures, pilot studies with data producers, and training & consultancy, including workshops and conferences. Furthermore, the Business Intelligence Team (BIT) contributes to this monitoring activity, both supply-driven (from BIT to the [LTP Service]) and demand-driven (from the [LTP Service] to BIT).	Daily	[LTP Service] Team Lead + Business Intelligence Team
2	Check and maintain [LTP Institution]'s preferred formats list for obsolescence and completeness (given mission and scope of the [LTP Institution]). If not: ● Analyse and select alternative preferred formats; ● Migrate all relevant files from outdated formats to successor formats, conforming the [LTP Institution] updated preferred formats guidelines; ● Update the relevant internal & external documents	On a regular basis	Preservation Officer + "Preferred Formats" Team
3	Check the potential impact on the [LTP Service]] of – expected – legal and/or regulatory changes, including codes of conduct (e.g. with respect to personal data, database law, etc.)	Continuously	Legal Advisor
4	Monitor [LTP Institution]'s ICT systems and storage facilities ● Monitor archival system ● Monitor storage media ● Check backup and recovery procedures	Continuously	IT Support + external storage provider
5	Monitor potential external threats to the ICT systems ● Periodic security updates to all systems ● Keep security policy up-to-date	Continuously, plus periodic updates of the security policy	Security Officer
6	Monitor Preservation Plan: ● Is it still up-to-date or is there a reason to update? ● Consequences of revisions?	Biannually	[LTP Service] Team Lead
7	Update the [LTP Institution] multiannual strategy, including the [LTP-service], strategic goals and designated communities	Every four-five years	Director [LTP Institution]

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 4: Available Licenses for digital assets uploaded to the EUDAT B2SHARE service

https://github.com/ufal/public-license-selector/#available-licenses

All licenses are open licences. However, some of these licenses require to comply with conditions such as giving appropriate credit via citation, providing a link to the license, indicating if changes were made, or non-commercial use. In cases where these conditions apply, creating an account and logging into the [LTP Service] is necessary in order to guarantee compliance.

License name	URL
Affero General Public License 3 (AGPL-3.0)	http://opensource.org/licenses/AGPL-3.0
Apache License 2	http://www.apache.org/licenses/LICENSE-2.0
Artistic License 1.0	http://opensource.org/licenses/Artistic-Perl-1.0
Artistic License 2.0	http://opensource.org/licenses/Artistic-2.0
Common Development and Distribution License (CDDL-1.0)	http://opensource.org/licenses/CDDL-1.0
Creative Commons Attribution (CC-BY)	http://creativecommons.org/licenses/by/4.0/
Creative Commons Attribution-NoDerivs (CC-BY-ND)	http://creativecommons.org/licenses/by-nd/4.0/
Creative Commons Attribution-NonCommercial (CC-BY-NC)	http://creativecommons.org/licenses/by-nc/4.0/
Creative Commons Attribution-NonCommercial-NoDerivs (CC-BY-NC-ND)	http://creativecommons.org/licenses/by-nc-nd/4.0/
Creative Commons Attribution-NonCommercial-ShareAlike (CC-BY-NC-SA)	http://creativecommons.org/licenses/by-nc-sa/4.0/
Creative Commons Attribution-ShareAlike (CC-BY-SA)	http://creativecommons.org/licenses/by-sa/4.0/
Eclipse Public License 1.0 (EPL-1.0)	http://opensource.org/licenses/EPL-1.0
GNU General Public License 2 or later (GPL-2.0)	http://opensource.org/licenses/GPL-2.0
GNU General Public License 3 (GPL-3.0)	http://opensource.org/licenses/GPL-3.0
GNU Library or "Lesser" General Public License 2.1 or later (LGPL-2.1)	http://opensource.org/licenses/LGPL-2.1

GNU Library or "Lesser" General Public License 3.0 (LGPL-3.0)	http://opensource.org/licenses/LGPL-3.0
Mozilla Public License 2.0	http://opensource.org/licenses/MPL-2.0
Public Domain Dedication (CC Zero)	http://creativecommons.org/publicdomain/zero/1.0/
Public Domain Mark (PD)	http://creativecommons.org/publicdomain/mark/1.0/
The BSD 2-Clause "Simplified" or "FreeBSD" License	http://opensource.org/licenses/BSD-2-Clause
The BSD 3-Clause "New" or "Revised" License (BSD)	http://opensource.org/licenses/BSD-3-Clause
The MIT License (MIT)	http://opensource.org/licenses/mit-license.php

APPENDIX 5: Legal and Statutory Context and Requirements

The LTP policy needs to conform to national and international law, statutory regulations, and business requirements of the [LTP Institution]. The following legislations, regulations, codes of conduct and guidelines are relevant for the management and LTP of research data and related digital assets:

European legislation and regulations

- Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (PE/28/2019/REV/1 - in brief Open Data Directive, formerly Public Sector Information (PSI) Directive): https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L1024[17]
- Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (in brief: Database Directive): https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML.[18]
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), repealing Directive 95/46/EC (in brief: GDPR): https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

International codes of conduct

- Singapore Statement on Research Integrity (2010): http://wcrif.org/guidance/singapore-statement
- OECD Best Practices for Ensuring Scientific Integrity and Preventing Misconduct (2007): https://www.oecd.org/science/inno/40188303.pdf
- ALLEA European Code of Conduct for Research Integrity (2017): http://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf
- UNESCO Recommendation on Science and Scientific Researchers (2017): http://portal.unesco.org/en/ev.php-URL_ID=49455&URL_DO=DO_TOPIC&URL_SECTION=201.html
- Guidelines of the Committee on Publication Ethics (COPE): https://publicationethics.org/resources/guidelines

[17] The Open Data Directive will be complemented by the Data Governance Act (DGA), the legislative framework to facilitate data-sharing (COM/2020/767 final, officially proposed by the EC on 25 November 2020. See: https://data.europa.eu/en/highlights/data-governance-act-open-data-directive for background and

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L1024) for the draft text..

[18] The directive is being reviewed as part of a proposed Data Act, see: https://en.wikipedia.org/wiki/Data_Act_(European_Union).

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

Data archiving guidelines

- The ten LTP Principles (listed above in section 2.2): https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re
- The FAIR Data Principles: https://www.go-fair.org/fair-principles/
- The mandatory responsibilities of an Open Archival Information System (OAIS): http://www.oais.info/
- The requirements of
  - the CoreTrustSeal: https://www.coretrustseal.org/
  - and/or optionally nestorSeal DIN 31644: https://www.langzeitarchivierung.de
  - and/or ISO 16363 https://www.iso.org/standard/56510.html

National legislation and codes of conduct (example for The Netherlands)

- Uitvoeringswet Algemene Verordening Gegevensbescherming (UAVG); the Dutch law implementing the European GDPR (16 May 2018, valid from 1 July 2021. https://wetten.overheid.nl/BWBR0040940/2021-07-01
- Databankenwet (1999), the Dutch Database Act, implementing the European Database Directive: https://wetten.overheid.nl/BWBR0010591/2021-06-07
- Auteurswet; Dutch Copyright Act (1912/latest update 7/6/2021): https://wetten.overheid.nl/BWBR0001886/2021-06-07
- The Netherlands Code of Conduct for Research Integrity 2018 (https://doi.org/10.17026/dans-2cj-nvwu), replacing the Netherlands Code of Conduct for Academic Practice (2014 revision).[19]

[19] The 2014 revision of the Netherlands Code of Conduct for Scientific Practice prescribed a minimum retention period of ten years for raw research data (article 3.3.) The new Code of Conduct for Research Integrity 2018 no longer specifies a minimum retention period, but says that data, software codes, protocols, research material and corresponding metadata should be stored *permanently* (as far as possible; section 4.4., art. 12).

APPENDIX 6: Selected terms used

In the context of digital preservation, the terms principles, strategy, policy and plan(ning) are often used, sometimes interchangeably. However, it makes sense to make distinction between the terms:

- A ***principle*** is a basic proposition that serves as the foundation for a system of belief or behaviour or for a chain of reasoning. In this context, LTP principles are fundamental premises concerning the archiving of digital materials.
- A ***policy*** is defined as the set of guidelines, rules and procedures developed by an organisation to govern its actions (often in recurring situations). They define the limits (do's and don'ts) within which decisions must be made, and are to be widely communicated and available and accessible, both to the organisation's staff and its customers.
- A ***strategy*** is a high-level plan of action designed to achieve one or more of the organisation's objectives. A strategy fills the gap between "where we are" and "where we want to be", that is, "how are we going to get there?" It relates to how an organisation allocates and uses materials and human resources.
- A ***plan*** explains in more detail how a strategy will be executed; it is the operational side of the strategy.[20]

The focus of the task at hand here is the formulation of a model LTP Policy Template, i.e. a set of guidelines, rules and procedures, but it will be useful to base this policy on generally acceptable principles. In addition to these terms, also the term combination "**policy framework**" occurs. This is a somewhat vague term, as all of the terms mentioned after the bullets above can be part of such a "framework", and many more things as well[21].

A second terminological clarification concerns the distinction between digital preservation and long-term preservation (LTP), two terms that are also often used indiscriminately. In ***digital preservation*** *sec*, there is no assumption whatsoever about the time span, although in practice, it often assumes the long term. Although the term ***LTP*** does not explicitly state it is about digital objects, this is obvious in this context. The term itself does not give a specification about the length of "long term", which can be either indefinite or definite, and is usually longer than 5 or 10 years[22]. It is clear that the policies we are about to formulate concern the long term. To be exact, we are talking here about a ***long-term digital preservation*** policy.

For practical reasons, moreover, it is sometimes necessary to draw distinction between the ***institution*** providing an LTP service (or LTP Institution or LTP Service Provider) and the ***LTP service*** itself. To clarify the point in an example: DANS, an Institute of the Royal Netherlands

[20] Note that according to the OAIS reference model, the "preservation planning function" has a more limited definition (see Appendix 1).

[21] The Digital Preservation Policy Framework by Ohio State University Libraries (OSUL) provides an idea of the diversity of subjects to be covered by such a framework, see: https://library.osu.edu/documents/SDIWG/Digital_Preservation_Policy_Framework.pdf

[22] Jeff Rothenberg is often quoted for his expression: "digital information lasts forever—or five years, whichever comes first" in "Ensuring the Longevity of Digital Information", RAND Corporation, 22/01/1999, p. 2. https://www.clir.org/wp-content/uploads/sites/6/ensuring.pdf

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

Academy of Arts and Science (KNAW) is an LTP Service Provider; it provides several services, among which an LTP Service, which is called the “Data Vault”. For an LTP Service, one or more systems or components of systems may be used.

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 7: LTP comparison between B2SHARE and B2SAFE

In Table 9 below an assessment and comparison is made between B2SHARE and B2SAFE on compliance to the Long Term Preservation Policy.

Table 9. LTP comparison between B2SHARE and B2SAFE

Section and Article	Brief description	Data service					
		B2SHARE			B2SAFE		
		https://marketplace.eosc-portal.eu/services/b2share			https://marketplace.eosc-portal.eu/services/b2safe		
		Curation	Status	Explanation	Curation	Status	Explanation
I. Objectives, scope and delimitation of this policy		Enhanced			Bit preservation		
1.1	General aim of LTP policy	S		B2SHARE 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.	S		B2SAFE 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.
1.2	Specific aims	S		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.	S		B2SAFE offers an abstraction layer of large scale, heterogeneous data storages, guards against data loss in long-term archiving, and allows optimized access for users (e.g. from different regions),
1.3	Scope of LTP policy	S		The policy, when applied, is only applicable to the B2SHARE central service (https://b2share.eudat.eu/)	S		When included in a binding contract, the policy is only applicable to the

							B2SAFE instances involved within the contract.
1.4	Responsibility	S		EUDAT Ltd is the service organisation responsible for the B2SHARE service. CSC is the service provider partner hosting and operating the service on behalf of EUDAT.	S		Service is not a public service and is only offered on the basis of a contract including a SLA and DPA. EUDAT contracts are underpinned by a SDA and OLA with the service providers hosting and operating the service instances.
1.5	Remit and mission	S		EUDAT’s vision is Data is shared and preserved across borders and disciplines. Achieving this vision means enabling data stewardship within and between European research communities through a Collaborative Data Infrastructure (CDI), a common model and service infrastructure for managing data spanning all European research data centres and community data repositories.	S		EUDAT’s vision is Data is shared and preserved across borders and disciplines. Achieving this vision means enabling data stewardship within and between European research communities through a Collaborative Data Infrastructure (CDI), a common model and service infrastructure for managing data spanning all European research data centres and community data repositories.
1.6	Preservation duration	S	Partially	The EUDAT CDI partnership agreement guarantees service provision of the EUDAT services for a period of 10 years by its members.	S	Yes	Is in agreement with the customer and contract.
1.7	Designated community	S	Yes	On request of research communities community domains have been defined to support deposits of data records on basis community-defined metadata templates.	S	Yes	Is in agreement with the customer and contract.

2. Ingest								
2.1	Submission Information Package for Ingest	X	Yes	Supports metadata and data objects.	X	Optionally	Users are able to store AIP packages accompanying the data objects, but this is not mandatory.
2.2	Metadata index and Cataloguing	X	Optionally	Support 2 publication workflows: - Open publication workflow in which users are allowed to directly publish data records - Reviewed publication workflow in which users are only allowed to submit draft data records which are reviewed before publication		No	N/A within the specified curation level
2.3	Licenses	X	Optionally	Specifying a license is an optional field		No	
2.4	Persistent Identifiers	X	Yes	Assigns a DOI and Handle PID to the landing page of the data records and Handle PIDs to each of the ingested data objects		Yes	Assigns Handle PIDs to each of the ingested data objects
	Additional PIDs		Yes	For each creator the following Identifiers can be optionally specified: - Affiliation (e.g. RoR) - Name identifier (e.g. ORCID)		No	N/A within the specified curation level
2.5	Quality control	enhanced checks of metadata	enhanced checks of metadata		none	No	

2.5.1	Metadata supplied by depositor	X	Yes	Metadata needs to be specified according to community defined metadata templates.		No	
2.5.2	Ingest control	enhanced	Partially	Checksums are automatically generated during deposit, depositors are required to describe data with minimum metadata, including creators and contact personal data. No preferred data formats are specified and therefore no re-formatting is required.	none	Partially	Checksums are automatically calculated and verified.
2.5.3	Required metadata fields	Extended	Yes	Depositors need to specify Title, 1 or more Creators, description, if data is open access or not, metadata is always open, an embargo period and a publication date can be specified. Different community domains are defined, including community defined metadata extensions.	none	No	N/A within the specified curation level
2.5.4	Responsibility for metadata	X	Yes	Independent from the publication workflow, the depositor needs to provide the minimum required mandatory fields.		No	
2.5.5	Notification of deficient metadata	X	Optionally	Is supported via the reviewed publication workflow and can be configured per designated community.		No	
2.5.6	Data and metadata corrections	X	Yes	Major changes are supported via the versioning feature, minor changes are supported via the reviewed publication workflow and/or can be made by the depositor.		No	

When LTP is outsourced							
2.6	Submission Information Package for Ingest	X	N/A	At time of writing LTP is not outsourced	M	N/A	At time of writing LTP is not outsourced
2.7	Metadata index and Cataloguing	X	N/A		N/A	N/A	
2.8	Licenses	X	N/A		N/A	N/A	
2.9	Persistent Identifiers	X	N/A		N/A	N/A	
2.10	Additional PIDs	X	N/A		N/A	N/A	
2.11	Quality control	basic checks of metadata	N/A		N/A	N/A	
2.11.1	Metadata supplied by depositor	X	N/A		N/A	N/A	
2.11.2	Ingest control	basic	N/A		N/A	N/A	
2.11.3	Mapping of metadata	X	N/A		N/A	N/A	
2.11.4	Required metadata fields	basic	N/A		N/A	N/A	
2.11.5	Responsibility for metadata	X	N/A		N/A	N/A	
2.11.6	Notification of deficient metadata	X	N/A		N/A	N/A	
2.11.7	Data and metadata corrections	X	N/A		N/A	N/A	

3. Archival Storage							
3.1	Archival actions and chain of provenance	extended	Partially	Via version management changes to licenses, metadata and file format versions are documented.	none	No	N/A within the specified curation level
3.2	Archival storage of AIPs	X	Yes	For all data records in which all data is locally maintained, AIP information is stored within the storage facilities. Data objects which are externally hosted only PID references are maintained.	X	Yes	For each data object uploaded a checksum is calculated and a PID maintained.
3.3	No responsibility for data stored externally	X	Yes	Service provider does not assume responsibility for externally hosted data objects.	S	Yes	Via the replication policy data objects can be replicated across multiple locations, for each of the locations checksums are calculated and verified and PIDs are maintained and linked to each other.
3.4	Integrity and security measures	extended	Partially	Within the EUDAT CDI infrastructure, EUDAT Ltd has OLA agreements with the service providers running services on behalf of EUDAT, responsibility to perform security audits is delegated to the service providers running the services. Security incidents, including data loss, are managed through the EUDAT CSIRT (Computer Security Incident Response Team). Data alteration and/or corruption is considered as data loss.	checksums	Yes	Within the EUDAT CDI infrastructure, EUDAT Ltd has OLA agreements with the service providers running services on behalf of EUDAT, responsibility to perform security audits is delegated to the service providers running the services. Security incidents, including data loss, are managed through the EUDAT CSIRT (Computer Security Incident Response Team). Data alteration and/or corruption is considered as data loss. During the upload of the data, checksums are automatically calculated and verified.

3.5	Outsourcing storage to external provider	S	Yes	EUDAT Ltd has OLA agreements with the service providers running services on behalf of EUDAT. The OLA agreement has service level targets on the availability and quality of the service, service provisioning, incident handling, support and backups. The services are provided according to FitSM processes.	S	Yes	EUDAT Ltd has OLA agreements with the service providers running services on behalf of EUDAT. The OLA agreement has service level targets on the availability and quality of the service, service provisioning, incident handling, support and backups. The services are provided according to FitSM processes. Via the replication policy multiple copies across different locations can be created.
4. Data Management							
4.1	Metadata catalogue maintenance	X	Yes	A database is maintained with metadata which is searchable. The B2SHARE technology supports community metadata schemas on the basis of the EUDAT Core metadata schema with optional community extensions.		Partially	A database is maintained with basic data object information (file name, user, groups, permissions). In the format of triples alo user specified metadata can be provided which is searchable via command line tools. Users are not required to specify user descriptive metadata.
4.2	Archival metadata supporting version control	X	Yes	Service supports versioning via which change control is being managed	X	No	Versioning is not supported
4.3.1	Derived and dissemination formats	X	Yes	Data objects are maintained as-is, no derived formats are created		No	N/A within the specified curation level
4.3.2	Authenticity and integrity of AIPs	X	Yes	A user can only remove draft data records, published data records cannot be removed, only by the authorised staff. Published data records can only be updated via versioning. For each of the data objects checksums are generated and maintained to verify the integrity of the data objects.	X	Yes	During the upload of the data, checksums are automatically calculated and verified.

4.3.3	Documentation of chain of custody	X	Yes	Service supports versioning via which change control is being managed		No	N/A within the specified curation level
4.4	Preferred and accepted file formats	X	No	Users are free to upload data in any data format and are maintained as-is		No	
4.5	Obsolescence of file formats		No			No	
4.6.1	Migration of preferred formats		No			No	
4.6.2	Preserving outdated formats		No			No	
4.7	Deletion of datasets and files	X	Yes	Only authorised staff are able to delete published data records	X	No	Data owners are allowed to delete data objects
4.8	Tombstone records for deleted datasets	X	No	No tombstone records are created	X	No	No tombstone records are created
5. Administration							
5.1	Monitoring operations and reporting	S	Yes	The availability and reliability of the B2SHARE CDI service are actively monitored in the EUDAT central monitoring service.	S	Yes	The availability and reliability of B2SAFE instances are actively monitored in the EUDAT central monitoring service.
5.2	LTP Strategy and upgrading the [LTP Service]	S	Yes	A roadmap of the service and technology is maintained	S	Yes	A roadmap of the service and technology is maintained
5.3	Planning and control cycle	S	Yes		S	Yes	
5.4	Strategic and operational evaluations	S	Yes	EUDAT is a member organisation, the strategic functioning and direction is monitored by the EUDAT Council.	S	Yes	EUDAT is a member organisation, the strategic functioning and direction is monitored by the EUDAT Council.
5.5	Advisory board	S	Yes	EUDAT has a User Board with representatives from the scientific communities.	S	Yes	EUDAT has a User Board with representatives from the scientific communities.

5.6	Customer support	S	Yes	Customer and user support is provided via the EUDAT Helpdesk	S	Yes	Customer and user support is provided via the EUDAT Helpdesk
5.7	Legal and statutory regulations	S	Partially	Terms of Use, Acceptable use Policy and Data Privacy Statement are available.	S	Partially	Service is not a public service and is only offered on the basis of a contract including a SLA and DPA. EUDAT contracts are underpinned by a SDA and OLA with the service providers hosting and operating the service instances.

6. Preservation Planning							
6.1	Monitoring the content of the [LTP Service]	X	Yes	Periodically metrics are published on usage of the service.	X	No	Data records are maintained as-is
6.2	Reviewing and updating list of preferred formats	X	N/A	No preferred file format list is maintained		No	N/A within the specified curation level
6.3	Other monitoring for preservation planning						
6.3.1	Security risks	X	Yes	Security risks are regular and continuously monitored via the ISM (Information Security Management process and procedures.		Yes	Security risks are regular and continuously monitored via the ISM (Information Security Management process and procedures.
6.3.2	Technology watch	X	Yes	Technology watch is a continuous aspect within the service development process		Yes	Technology watch is a continuous aspect within the service development process
6.3.3	Service requirements	X	Yes	Via the issues in the software repository a feature request list is maintained		Yes	Via the issues in the software repository a feature request list is maintained
6.4	Roles and responsibilities, confidentiality, liability	X	Yes	The EUDAT secretariat will be responsible for maintaining the LTP policy, the implementation of the policy is delegated to the service provider running an instance of the service on behalf of EUDAT on the basis of the OLA agreement.		Yes	The EUDAT secretariat is responsible for maintaining the LTP policy, the implementation of the policy is delegated to the service provider running an instance of the service on behalf of EUDAT on the basis of an OLA agreement.
6.5.1	Funding adequate for sustaining [LTP Service]	X	Partially	The EUDAT CDI partnership agreement guarantees service provision of the EUDAT services for a period of 10 years by its members.		Partially	The EUDAT CDI partnership agreement guarantees service provision of the EUDAT services for a period of 10 years by its members.

6.5.2	Contingency plan	X	Yes	Yes, for the duration of the EUDAT CDI Partnership agreement. CSC is the service provider partner hosting and operating the service on behalf of the EUDAT. CSC is a level 1 partner of the EUDAT CDI.		Yes	Yes, for the duration of the EUDAT CDI Partnership agreement. If a more comprehensive contingency plan is required, this can be part of a contract.
7. Access							
7.1	Access functions (user interface, API)	S	Yes	A WUI and APIs are provided to access the service, including an OAI-PMH endpoint for harvesting.	S	Partially	Service is not accessible via a WUI, different APIs are provided to up/download and manage data.
7.2	Metadata openly accessible	S	Yes	Metadata is open accessible, data can be closed	S	No	Default data is only accessible to the owner.
7.3	User registration for accessing licensed datasets	S	No	Default data is open accessible. Users uploading a data record can close access to the data objects. If this is the case, users interested to download the data objects must request access.	S	Optionally	Optionally the owner can share data with registered users.
7.4	CC Waiver and Open Access without registration	S	Yes	Users do not need to register an account or to login to access Open Access data records and objects.	S	No	Default data is only accessible to the owner.
7.5	Conditions of use and citation recommendation	S	No	No explicit statements and/or citation recommendations are made.	S	No	No explicit statements and/or citation recommendations are made.
When LTP is outsourced							
7.6	Access through original data service via PID	S	No	At time of writing LTP is not outsourced	S	Yes	When data is replicated, PIDs to all copies of the data are interreferenced
7.7	Access in case original data service discontinued	S	No		S	Optionally	When data is replicated across multiple locations, access to the replicated copies can be provided.

APPENDIX 8: Example recursive curl_script for upload

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>
```

APPENDIX 9: Process and examples for using curl as a tool to publish data on B2SHARE

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

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

Object retrieval and publication

- 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″
      ```

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

- 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",
    "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”
    ```