


D5.1: Report on Service Building Status and Progress, year 1

Author(s)	Mark van de Sanden (SURFsara), Carl Johan Håkansson (SNIC), Claudio Cacciari (CINECA), Alison Packer (STFC), Emanuel Dima (EKUT), Benedikt von St. Vieth (JSC), Hannes Thiemann (DKRZ), Toni Cortes (BSC), Willem Elbers (CLARIN), Tobias Weigel (DKRZ), Roberto Mucci (CINECA), Yann Le Franc (eSDF), Christine Staiger (SURFsara)
Status	Final
Version	v1.0
Date	10/05/2016

Abstract: This document provides a comprehensive overview of the activities related to building data management services and of the architectural discussions about the EUDAT Common Data Infrastructure (CDI) that were conducted during the first year of the EUDAT2020 project. The service building activities concentrated on the consolidation and integration of the existing EUDAT services and on the uptake of new community requirements initiated from the community uptake plans and data pilots or as a result from Joint Result Activities (JRA). The discussions about the architecture of the EUDAT CDI focused on the definition of the data model used, the layered architecture and the support for metadata within the CDI. In addition, a service building roadmap was defined to facilitate communication between EUDAT and the communities and project enablers about new releases of the services and supported functionalities – the roadmap is updated every half year. Furthermore, a procedure for identifying the requirements for new or proposed services was established – this will make it easier to adapt the EUDAT CDI network to any new requirements in an agile way.

Document identifier: EUDAT2020-DEL-WP5-D5.1	
Deliverable lead	SURFsara
Related work package	WP5
Author(s)	Mark van de Sanden (SURFsara), Carl Johan Håkansson (SNIC), Claudio Cacciari (CINECA), Alison Packer (STFC), Emanuel Dima (EKUT), Benedikt von St. Vieth (JSC), Hannes Thiemann (DKRZ), Toni Cortes (BSC), Willem Elbers (CLARIN), Tobias Weigel (DKRZ), Roberto Mucci (CINECA), Yann Le Franc (eSDF), Christine Staiger (SURFsara)
Contributor(s)	Heinrich Widmann (DKRZ), Jens Jensen (STFC)
Due date	28/02/2016
Actual submission date	10/05/2016
Reviewed by	Marjan Grootveld (DANS), Dieter van Uytvanck (CLARIN)
Approved by	PMO
Dissemination level	PUBLIC
Website	www.eudat.eu
Call	H2020-EINFRA-2014-2
Project Number	654065
Start date of Project	01/03/2015
Duration	36 months
License	Creative Commons CC-BY 4.0
Keywords	CDI Layered Architecture, CDI Data Model, B2 services suite, Data Repository service, B2SHARE, Personal Cloud Storage, B2DROP, Data Discovery, B2FIND, Federated AAI, B2ACCESS, Data Management Policies, BSAFE, Persistent Identifiers, B2HANDLE, Data Staging, B2STAGE, Semantic Annotation, B2NOTE, Data Policy Manager, Data Type Registry, User Documentation, Training Materials, Service Building Requirements, Roadmap

Copyright notice: This work is licensed under the Creative Commons CC-BY 4.0 licence. To view a copy of this licence, visit <https://creativecommons.org/licenses/by/4.0>. 

Disclaimer: The content of the document herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the document is believed to be accurate, the author(s) or any other participant in the EUDAT Consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the EUDAT Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the EUDAT Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

TABLE OF CONTENTS

EXECUTIVE SUMMARY	6
1. INTRODUCTION	8
2. EUDAT B2 SERVICES SUITE	10
3. EUDAT CDI ARCHITECTURE.....	12
3.1. EUDAT Business Architecture	12
3.1.1. Using versus Joining.....	12
3.1.2. CDI Service Provider Levels.....	13
3.1.3. CDI Nodes	14
3.2. EUDAT Data Architecture	14
3.2.1. EUDAT Data Domains	14
3.2.2. EUDAT Data Model	15
3.2.3. PIDs in Digital Objects.....	17
3.3. EUDAT Application Architecture	17
3.3.1. CDI Network.....	18
3.3.2. CDI Centrally Provided Services.....	18
3.3.2.1. CDI Public Services.....	19
3.3.2.2. CDI Auxiliary B2 Services	19
3.3.2.3. Service Management Infrastructure	20
3.3.3. CDI Layered Network Architecture.....	21
3.3.4. CDI Node Layered Architecture	23
3.3.4.1. Interoperable Node Architecture	23
3.3.4.2. Integrated Node Architecture	25
3.3.5. Service Portfolio	27
3.4. Supporting Metadata within the CDI.....	28
3.4.1. Metadata Types	28
3.4.1.1. Descriptive Metadata	28
3.4.1.2. Administrative Metadata.....	29
3.4.1.3. Structural Metadata	29
3.4.1.4. System Metadata.....	29
3.4.2. Local Metadata Store	30
3.4.3. Data Flow	30
3.5. Outlook	31
4. DATA ACCESS AND RE-USE	33
4.1.1. Future Work in the Service Area	33
4.2. Data Repository – B2SHARE	33
4.2.1. Development Progress during the First Year.....	34
4.2.2. Outlook	35
4.3. Personal Cloud Storage – B2DROP	35
4.3.1. Development Progress during the First Year of EUDAT2020	36
4.3.2. Outlook	37
4.4. Data Discovery – B2FIND	37
4.4.1. Development Progress during the First Year of EUDAT2020	38
4.4.2. Outlook	39
4.5. Registry Services.....	39
4.5.1. Development Progress during the First Year of EUDAT2020	40
4.5.2. Deployment of CORDRA as a Testing Instance of the DTR in EUDAT's CDI.....	40
4.5.3. Outlook	40
4.6. Federated AAI – B2ACCESS.....	40
4.6.1. Architecture	41
4.6.2. Identity Provider Integration	42

4.6.3.	Service Provider Integration	42
4.6.4.	Outlook	43
5.	DATA PRESERVATION	44
5.1.	Data Management Policies	44
5.1.1.	B2SAFE	45
5.1.2.	Data Policy Manager	47
5.1.3.	Documentation	49
5.1.4.	Tests	49
5.1.5.	Outlook	49
5.2.	Persistent Identifiers (PIDs)	50
5.2.1.	B2Handle Development	50
5.2.2.	Coordination Activities	51
5.2.3.	Outlook	52
5.3.	Data Curation and Provenance	52
5.3.1.	Expected Outcomes	52
5.3.2.	Challenges and Opportunities	53
5.3.3.	Strategy	53
6.	DATA PROCESSING AND ANALYSIS	54
6.1.	Data Staging (B2STAGE)	54
6.1.1.	GridFTP iRODS-DSI	54
6.1.2.	Data Staging Script	55
6.1.3.	EUDAT Python Library	55
6.2.	Semantic Web Services	56
6.3.	Outlook	58
7.	USER DOCUMENTATION AND TRAINING MATERIALS	59
7.1.	User Documentation	60
7.2.	Training Materials	62
7.3.	Future Plans	62
8.	SERVICE BUILDING REQUIREMENTS AND ROADMAP	64
8.1.	Service Building Requirements Procedure	64
8.1.1.	Research Community Requirements	65
8.1.2.	Common Requirements	66
8.1.3.	Development Requirements	66
8.2.	Service Building Roadmap	66
9.	CONCLUSIONS	68
ANNEX A.	TOGAF ARCHITECTURE DEVELOPMENT METHOD	70
ANNEX B.	EUDAT CDI DATA MODEL TERMS AND DEFINITIONS	72
ANNEX C.	DESCRIPTIVE METADATA LEVELS	73
ANNEX D.	DPM TESTBED ENVIRONMENT	74
ANNEX E.	B2ACCESS ATTRIBUTES	75
ANNEX F.	SERVICE BUILDING REQUIREMENTS DESCRIPTION TEMPLATE	76
ANNEX G.	SERVICE BUILDING ROADMAP M13	78
ANNEX H.	GLOSSARY	80

LIST OF FIGURES

Figure 1: High level overview of the B2 service suite	10
Figure 2: Schematic overview of the CDI architecture reflecting Using and Joining.....	13
Figure 3: Relationship between EUDAT and ANDS data curation continuum	15
Figure 4: CDI data model (conceptual)	16
Figure 5: Logical representations of digital packages	17
Figure 6: PID relationship diagram of a DO uploaded as two separate digital entities	17
Figure 7: High level diagram of the CDI network with generic and thematic data centres	18
Figure 8: CDI Network layered architecture.....	23
Figure 9: Architectural diagram of an interoperable CDI node	24
Figure 10: CDI Integrated Node layered architecture	25
Figure 11: Examples of the different levels of descriptive metadata.....	29
Figure 12: Logical diagram of the local metadata store for the B2SAFE, B2SHARE and B2STAGE API services	30
Figure 13: Dataflow from a CLI/API	31
Figure 14: Dataflow from a browser	31
Figure 15: Dataflow from a GridFTP client	31
Figure 16: Uploading data in B2SHARE.....	34
Figure 17: Sharing a file in B2DROP via a link.....	36
Figure 18: Publishing queue view of B2DROP	37
Figure 19: B2FIND result page (right panel) of a combined faceted search (left navigation bar).....	38
Figure 20: Architectural overview of the B2ACCESS service	41
Figure 21: Technical overview of the B2ACCESS service	42
Figure 22: B2SAFE modules architectural overview	45
Figure 23: B2SAFE workflow: object upload and PID registration	46
Figure 24: B2SAFE workflow: object replication and PID registration	47
Figure 25: Data Policy Manager data flow	48
Figure 26: DMP test bed machines.....	49
Figure 27: The iterative evolution of the B2HANDLE service	51
Figure 28: Data transfer with iRODS-DSI	55
Figure 29: Data transfer with iRODS-DSI passing a PID as input	55
Figure 30: User documentation is a hub of EUDAT activity	60
Figure 31: The User Documentation web page, implementing the Engage-Deploy-Use model.	60
Figure 32: The user documentation workflow engages developers and operators.	62
Figure 33: Logical diagram to define service building requirements	65
Figure 34: TOGAF Architecture Development Method.....	70
Figure 35: TOGAF Architecture Domains.....	71
Figure 36: TOGAF in relation to other Frameworks	71

LIST OF TABLES

Table 1: EUDAT CDI data model terms and definitions.....	72
Table 2: Definition of descriptive metadata levels.....	73
Table 3: Overview of the testbed environment to test the Data Policy Manager	74
Table 4: Overview of the B2ACCESS attributes	75
Table 5: Service Building Roadmap M13	78

EXECUTIVE SUMMARY

The EUDAT2020 project is the continuation of the initial EUDAT project under the Horizon 2020 programme. This new phase of the EUDAT project started on the 1st of March 2015 and is based on the work of the initial phase of the project, which, in particular, produced the basis of the EUDAT Common Data Infrastructure (CDI) network and the B2 service suite offered by EUDAT. This deliverable provides an overview of the service building activities conducted within the first year of the current phase of the EUDAT project.

The main objectives of the service building activities within the current phase of EUDAT are to consolidate the development of the CDI common building blocks (with a special focus on the integration of those building blocks with each other into the CDI, to develop new services and tools to address the full data life cycle, and to define an architectural blueprint for the EUDAT CDI. In addition to developing the requisite data management services, EUDAT's service-building work package is responsible for providing technical user documentation and also contributes to the development of training material.

The various tasks involved with building services are undertaken by EUDAT's technically-oriented work packages (WPs) – namely Community Requirements and Engagement (WP4), Service Building (WP5), Operations (WP6), Cross e-Infrastructure Services (WP7), Data Life Cycle across Communities (WP8), and Technology Exploration (WP9). EUDAT has established a Technical Committee to organize the collaboration, knowledge transfers and information flow between these different work packages.

The work on defining the overall blueprint for the EUDAT CDI architecture is progressing: the basis of the CDI data model and the CDI layered architecture are still in the process of being defined, while the steps that are required for research communities and EUDAT centres to become *interoperable* and *integrated* CDI nodes have been finalised, both in the context of the current discussions on the *CDI collaboration agreement*, and in the context of *using and joining* the EUDAT CDI. Work on the topic of *how to support metadata* is ongoing, and this document provides details of the current status quo of those discussions.

EUDAT's service building activities to date have mainly focused on the consolidation of the existing B2 services and their integration. Here are the highlights of the developments for each of these services.

- **B2SHARE:** a new architectural design based on Invenio 3, support for role-based authorization, editable metadata, and integration with B2ACCESS and B2DROP
- **B2DROP:** branding of the B2DROP web interface, automated deployment, and integration with B2SHARE and B2ACCESS
- **B2FIND:** enhancing the ingesting of metadata, and improving and generalizing the semantic mapping
- **B2ACCESS:** first release and handover of the B2ACCESS service, integration with B2SHARE, integration with identity providers (IdPs) such as eduGAIN and social IdPs – like Facebook, Google, and Microsoft, providing extensive user documentation to ease the integration of the B2ACCESS service with the research community services and the other B2 services
- **B2SAFE:** first release of B2SAFE with iRODS v4 support, refactoring of the policies to support iRODS v4, beta release of the Data Policy Manager, assessing the integration of B2SAFE with B2ACCESS and the support for metadata
- **B2HANDLE:** branding of the B2HANDLE service, incorporating support for Handle v8, providing a unified B2HANDLE library to manage persistent identifiers (PIDs), initiating standardization of the EUDAT PID record and policies
- **B2STAGE:** providing support for data transfers via GridFTP based on PIDs, an EUDAT specific Python¹ library with support for data transfers and to search for data within B2FIND service and the definition of the HTTP Application Program Interface (API) to ease the up- and download of digital assets to and from the CDI.

In addition to these developments for the existing B2 services, new developments were initiated: a new service, **B2NOTE**, was taken up from the Joint Research Activity (JRA) activities from the first phase of the EUDAT project and the development of a **Data Type Registry** service came about from interest shown by EUDAT's research communities which lead to it being taken up from the RDA PID Information Type and Type Registries working groups.

¹ <https://www.python.org/>

In the second year of the EUDAT2020 project, EUDAT plans to introduce a number of important functionalities and new services. Special attention will be given to the further integration of the B2 services (i.e. B2ACCESS, B2SAFE, B2SHARE, B2DROP) with each other. The support for managing descriptive metadata within the CDI will be extended thanks to the introduction of local metadata store. Also, to extend the support for access to data within the CDI network, there will be an additional focus on developments in authorization.

To facilitate communication between EUDAT and communities and project enablers about new releases within the B2 services suite, a service building roadmap was defined and is updated every half year, Furthermore, a service building request procedure has been put in place to support requests for new functionalities and services.

1. INTRODUCTION

The strategic vision of the EUDAT project is to enable European researchers and practitioners from any research discipline to preserve, find, access, and process research data in a trusted environment. EUDAT is realising this vision by enriching the pan-European e-infrastructure landscape with a collaborative data infrastructure (CDI). The EUDAT CDI consists of a distributed network of data-related services which are provided by scientific data centres all over Europe and which are designed to enhance research collaborations using electronic data.

The current phase of the EUDAT project, which is formally known as EUDAT2020, is the continuation of the initial successful first phase of EUDAT, which finished on the 31st of March 2015. EUDAT2020 started on the 1st of March 2015. The strategy that this phase of the EUDAT project has followed during its first year, and that it will continue follow over the course of the next two years, aims at 1) bringing more instances of the EUDAT service into production, 2) integrating those instances with each other and with other e-infrastructures, and 3) providing high-quality data platforms that make it easier for researchers and research communities to collaborate with each other across Europe and also globally.

The purpose of EUDAT's Service Building work package (WP 5) is to consolidate and further develop the technical architecture of the EUDAT CDI. Furthermore, this work package is concerned with building services that enable European research communities, research organisations, and individual researchers to properly manage their research data properly throughout the whole life cycle of the data – from its creation and then preservation to re-using it.

The Service Building work package plays a central role in the current phase of the EUDAT project – it translates the requirements of European research communities and researchers into a viable CDI network, which includes technical solutions that support the community data management requirements, and provides high-quality services that are then brought into production by the EUDAT operations team (WP6). In order to fulfil their mission, the service building work package collaborates closely with the EUDAT stakeholders so that all the developments within the project are steered by the needs of the stakeholders. The Service building work package will continue to develop the EUDAT CDI architecture and the core building blocks of the CDI (such as the B2 suite of services). In addition, WP5 will take up the proposals that have come out of EUDAT's previous research activities (such as the ideas of implementing services for Semantic Annotations and a Generic Execution Workflow) and will initiate the development of new services in the areas of data registries, data curation, data provenance and workspaces.

The activities that are being undertaken by the service building work package are divided into three *service areas*: 1) data access and re-use, 2) data preservation and 3) data processing and analysis. The work in each area is led by a dedicated *service area manager*. Within each of these three service areas, the developments are categorised according to the service they are related to and the work on each of those services is in turn coordinated by a designated development coordinator. The *service area manager* steers the overall developments within service area and acts as the liaison with the other work packages for each of the services within his/her own service area.

As part of providing high quality services for handling research data, the Services Building work package also generates easy-to-use documentation for people who use the EUDAT services. Since having good documentation makes it much easier for research communities, researchers and the centres that are nodes in the EUDAT CDI to take up and/or use the CDI services, providing good quality user documentation is a vital part of the service delivery process. In this phase of the EUDAT project, the activities related to user documentation were relocated from the operations work package (which had taken care of documentation in the first phase of the project) to the Service Building work package. The aim of that change was to improve the process of producing documentation by locating the team that coordinates and edits the user documentation alongside the pool of developers who develop the services and hence provide the raw material on which the documentation is based. Additionally, in this phase of the project, the User Documentation team is also responsible for producing training material for the core EUDAT services. To this end, the User Documentation team works closely together with the Communication, Training and Outreach work package (WP3), which is responsible for publishing documentation on the EUDAT website and is also instrumental in determining the strategic directions when it comes to planning training material that meets

the needs of EUDAT's users. In particular the User Documentation team is concerned with setting up sandboxes (that is, virtual environments) that are used to help people learn how to use the EUDAT services, and also works in close collaboration with the developers to produce hands-on material for training sessions on using the EUDAT services.

EUDAT is a user-driven project in which European research communities and researchers collaborate to define the functionalities that are then provided through the EUDAT CDI network. The process of coming to a consensus on the appropriate functionality involves a lot of discussions between the different groups involved in the project. EUDAT has therefore established a Technical Committee (TC) to provide a platform for the necessary technical discussions and to facilitate the process of making the requisite decisions in the current phase of the project. The TC steers the technical discussions on a higher level, namely on the level of the CDI architecture and service definitions. It provides an environment where joint requirements are gathered and then prioritised, it makes decisions on approving and rejecting requirements, and it prioritizes and plans the releases of the functionalities which surface from the requirements. To communicate about functional releases within the B2 suite services, within the project and with the EUDAT research communities, the TC maintains a service building roadmap. The TC consists of the work packages leaders from the work packages on Community Requirements and Engagement (WP4), Service Building (WP5), Operations (WP6), Cross e-Infrastructure Services (WP7), Data Life Cycle across Communities (WP8), and Technology Exploration (WP9), along with the services area managers from the three service areas. The TC is coordinated by the leader of the Service Building work package.

EUDAT is a user-driven project and therefore the research communities participate in and contribute to the development of the B2 services in almost all the service building activities. For example, in the developments of the B2ACCESS service, there is active participating from the CLARIN² and ENES³ research communities which helps by bringing in views and expertise on authentication and authorisation from a research community perspective. In case of the B2ACCESS service, the development is also coordinated by a colleague from the CLARIN community. The situation is similar for the developmental work on the B2SHARE, B2FIND and B2HANDLE services which is also coordinated by research community partners (namely CLARIN/EKUT⁴ and ENES/DKRZ⁵). When it comes to the development of the B2SAFE service, there is active participating from the EPOS⁶ community generally, with CLARIN, ENES and EPOS all contributing to the developments relating to workflows and workspaces. It is very important that the research communities participate in the actual development process in this way as it ensures that the resulting services match the needs and expectations of EUDAT's research community partners.

The Service Strategy, Policy and Sustainability work package (WP2) is responsible for orchestrating the service management process within EUDAT and maintaining what are known as the *service portfolio*⁷ and the *service catalogue*⁸. As the Service Building work package is responsible for the development of new services, it is also responsible for the technical definition of the services that are ultimately brought into production by EUDAT. Hence, to keep the service portfolio up to date, the Service Strategy, Policy and Sustainability work package and the Service Building work package need to exchange information with each other about the services on a regular basis, preferably via a tool that automates the gathering of the relevant information.

Having briefly introduced the work of the Service Building work package, the rest of this document goes into more details about WP5's activities: chapter 2 gives a short overview of the B2 services suite, chapter 3 provides an overview of the current status of discussions about the architecture of the EUDAT CDI and handling metadata, chapters 4 through to 6 provide an overview of the service building activities conducted within the first year of this phase of the EUDAT project in relation to the three service areas – Data Access and Re-use (chapter 4), Data Preservation (chapter 5) and Data Processing and Analysis (chapter 6). Chapter 7 provides an overview of the activities conducted by the User Documentation and Training Material team and Chapter 8 describes the processes involved in determining the requirements for building services and also outlines the service building roadmap.

² CLARIN is the Common Language Resources and Technology Infrastructure (<http://clarin.eu/>)

³ IS-ENES is the Infrastructure for the European Network of Earth System Modelling (<https://is.enes.org/>)

⁴ <https://www.uni-tuebingen.de/>

⁵ <https://www.dkrz.de/>

⁶ EPOS is the European Research Infrastructure on solid earth (<https://www.epos-ip.org/>)

⁷ The concept of the *service portfolio* is explained in deliverable D2.1 Service Portfolio processes definition and SLA template

⁸ The concept of the *service catalogue* is explained in deliverable D2.1 Service Portfolio processes definition and SLA template

2. EUDAT B2 SERVICES SUITE

Instead of using a top-down approach for gathering requirements and building data handling services, the first phase of the EUDAT project based its development process on the following basic assumptions. (1) All of the services offered by EUDAT should be layered, that is, they should be modular not only with respect to their functional components but also with respect to the underlying technology. This was to enable EUDAT to adopt new software/hardware requirements and technology swiftly, by easily replacing out-dated building blocks. (2) All the EUDAT services should be designed in such a way that existing research community data organizations would only need to make moderate adaptations where applicable (3) The requirements of the research communities have the highest priority and consequently the elements of an overall architecture and service concepts need to emerge in a bottom-up fashion.

EUDAT followed the advice of the research communities and initially established a small number of services and used them to understand what the implications of “typical” common services are. In this way the B2 services suite was developed. The suite consists of five B2 services (B2DROP, B2SHARE, B2SAFE, B2STAGE and B2FIND) with two auxiliary supporting services for federated authentication and authorization (B2ACCESS) and persistent identifiers (B2HANDLE) – the relations between these services are shown in Figure 1.

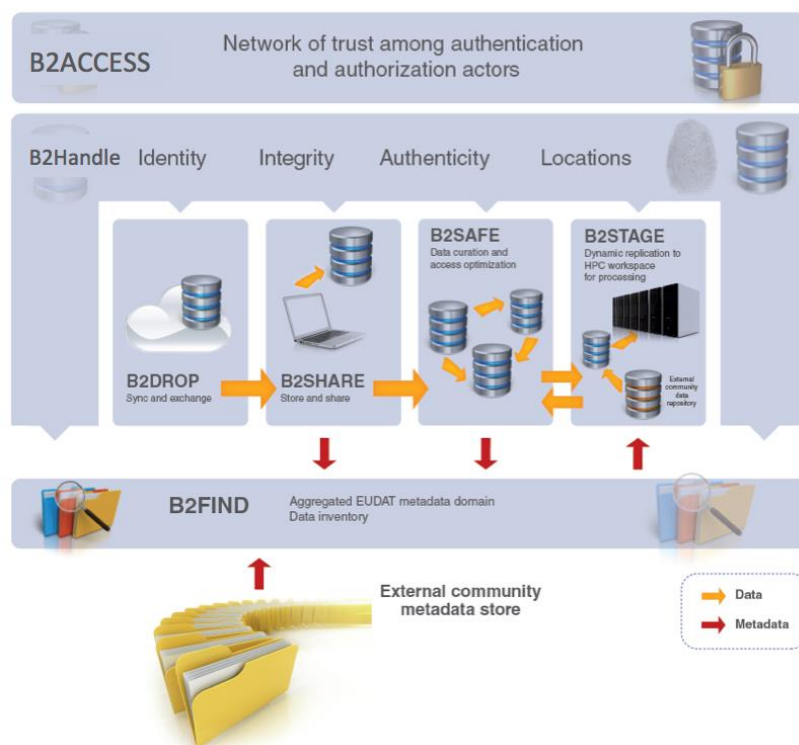


Figure 1: High level overview of the B2 service suite

EUDAT’s B2 suite services provide a range of valuable data handling services to assist researchers and research data managers in the following ways.

- **B2DROP** is a cloud storage service for long-tail data where individual researchers, small research groups and citizen scientists can store and exchange research data.
- **B2HARE** is an easy-to-use data repository service where individual researchers, research groups and citizen scientists can preserve, publish and share their research data.
- **B2SAFE** is a robust, safe and highly available service which makes it possible for community and departmental repositories to implement data management policies on their research data across multiple administrative domains in a trustworthy manner.
- **B2STAGE** is a reliable, efficient and easy-to-use service for transferring research data sets between EUDAT storage resources and high performance computing (HPC) workspaces.
- **B2FIND** is a central metadata catalogue where metadata relating to data stored within the EUDAT CDI network, and also metadata relating to data from research communities or from external

metadata providers, is gathered and made available in such a way that the metadata can be queried. This enables researchers to find collections of multi-disciplinary research data and thus facilitates the re-use of data and cross-disciplinary research.

- **B2HANDLE** provides tools for managing persistent identifiers within the EUDAT CDI network, that is, it enables people to register data, and thus makes it possible to refer to or cite research data for decades into the future.
- **B2ACCESS** adds authentication and authorization functionality to the EUDAT CDI network and thereby links the EUDAT CDI, research communities and external identity provider domains.

The initial EUDAT services were developed using a bottom-up approach and largely independently of each other. This has meant that, although those services provide the basic pillars that are needed for data management, it is not always so easy to move data smoothly between them. For example, the B2STAGE APIs and tools were developed for moving data to and from B2SAFE, while the B2SAFE service provided the registration of persistent identifiers in B2HANDLE for the data. B2SHARE was developed with descriptive metadata being harvested by B2FIND in mind. However at the time of writing, data and metadata in B2SAFE is only accessible to the person who puts it there. To enable researchers to easily publish data from B2SAFE and to make it findable via B2FIND, further integration and hence development is needed.

The two main objectives of EUDAT's service building activities are to produce an architectural blueprint of the EUDAT CDI and to integrate the B2 services more fully with each other. The number of B2 services that are in production and the relevant usage statistics will be reported in the D6.1 deliverable, along with a first year status and progress report on the operations within the current phase of the EUDAT project.

3. EUDAT CDI ARCHITECTURE

The EUDAT Consortium is a collaboration which focuses on the development and realization of the EUDAT CDI and its supporting ecosystem of expertise in the practice of research data management. The consortium includes operations service providers, software development service providers, data service providers from research communities, and expert service providers.

From this point of view, EUDAT can be considered to be an enterprise organization that is developing and operating a complex distributed IT infrastructure for managing research data (namely the EUDAT CDI). In practice, different methodologies⁹ are used to define an enterprise architecture, with the TOGAF¹⁰ methodology being widely used in Europe.

In the context of EUDAT, the main reasons for using the TOGAF framework (and adapting it to the needs of EUDAT) are:

- to agree on a common terminology for defining components and actors within the EUDAT CDI infrastructure,
- to come to a common understanding of the baseline and target CDI infrastructure,
- to have a common method for reaching agreement on definitions, standards and guidelines to be used within the CDI infrastructure, and
- to have a common method for agreeing on definitions, standards and guidelines to be used for developing services.

The following sections of this chapter describe the EUDAT CDI architecture according to the TOGAF architecture domains for business (3.1), data (3.2) and applications (3.3). The technology architecture domain is not addressed in this chapter but instead in the sections on the different service areas (which are in chapters 4, 5 and 6). In addition, Annex A provides a short introduction to TOGAF for anyone who is not familiar with it.

Supporting metadata within the EUDAT CDI network was a major topic of discussion in first year of this phase of the EUDAT project and it has greatly influenced both the architecture of the CDI and EUDAT's underlying data model. At the start of this phase of the project, the only support that was provided for descriptive metadata was via the B2SHARE and B2FIND services. However requirements from the research communities showed that there was a definite need to extend the support of descriptive metadata throughout the EUDAT CDI network and across the B2 services. The initial results from the discussions on this topic and an explanation of how it influences the CDI architecture are given in the section on Supporting metadata within the CDI (3.4) and are further addressed in the business, data and application architecture sections.

3.1. EUDAT Business Architecture

Although the business aspects of the EUDAT collaboration are mostly addressed by the Service Strategy, Policy and Sustainability work package (WP2), aspects of decisions made on the business and strategic level also influence the architectural definitions within EUDAT. The purpose of this section is not to provide a full overview of the business directions and discussions taking place within this phase of the EUDAT project, but rather to give a comprehensive overview of the various aspects that influence the architecture of the CDI at the data and application level. These aspects concern how a research community or user can make use of the CDI network and services, and also how potential partners can join the CDI collaboration and become a CDI service provider (for example, as a CDI node). We also describe the different types of CDI nodes that make up the EUDAT CDI network.

3.1.1. Using versus Joining

EUDAT distinguishes between two types of customers: those who are users of the EUDAT CDI and those who join the CDI collaboration agreement and become part of the CDI network as a CDI node. Both generic and thematic data centres may become part of the CDI under the collaboration agreement. The difference between *using* and *joining* the EUDAT CDI can be summarized as follows.

⁹ https://en.wikipedia.org/wiki/Enterprise_architecture_framework

¹⁰ <http://www.opengroup.org/subjectareas/enterprise/togaf>

- Any researcher, or research group or collaboration may **use** EUDAT services – this allows them to use services that are supplied by EUDAT and deployed at specific sites (which could be CDI generic or thematic nodes) according to the terms and conditions of the service specified by EUDAT and the service provider.
- Generic and thematic service providers may **join** the EUDAT CDI by signing the CDI collaboration agreement, and thus become a part of the CDI network and provide CDI services to a generic or thematic user community. Partners can join the CDI in two different ways, each of which corresponds to different levels of integration and different levels of responsibility. Both options are open to any generic or thematic site hosting a data repository and providing Information and Communication Technology (ICT) services to facilitate scientific research and are described in the next subsection.

3.1.2. CDI Service Provider Levels

- Interoperable nodes** must have a data repository in which they preserve or curate¹¹ data from a single research community or where they host data from several research communities or experiments (*storage* and *access* services). Interoperable nodes must identify the data that is hosted in their repository via some form of persistent identifier (*persistent identification*), and it must be possible to harvest and discover the associated metadata using EUDAT's B2FIND service (*metadata*).
- Integrated nodes** fulfil the requirements for being interoperable (as above) and also integrate their local data infrastructure with the EUDAT CDI's data management services, provide a common data access layer, integrate the enabled data infrastructure with the common authentication and authorization infrastructure (AAI), and connect their services to the common EUDAT CDI service management infrastructure and operate the integrated services according to the service management framework.

Figure 2 is a schematic diagram that illustrates the difference between the concepts of *Using EUDAT* as opposed to *Joining EUDAT* and shows the required levels of integration of partners joining as *interoperable* or as *integrated* partners. The diagram also shows the layered nature of the EUDAT CDI architecture that is explained in section 3.3.3. In addition, the figure illustrates the notion of providing services locally to the user community designated for the particular service provider in contrast to the idea of providing services centrally as a public service, CDI auxiliary service or as part of the service management infrastructure. The services that are provided centrally are described in more detail in section 3.3.2.

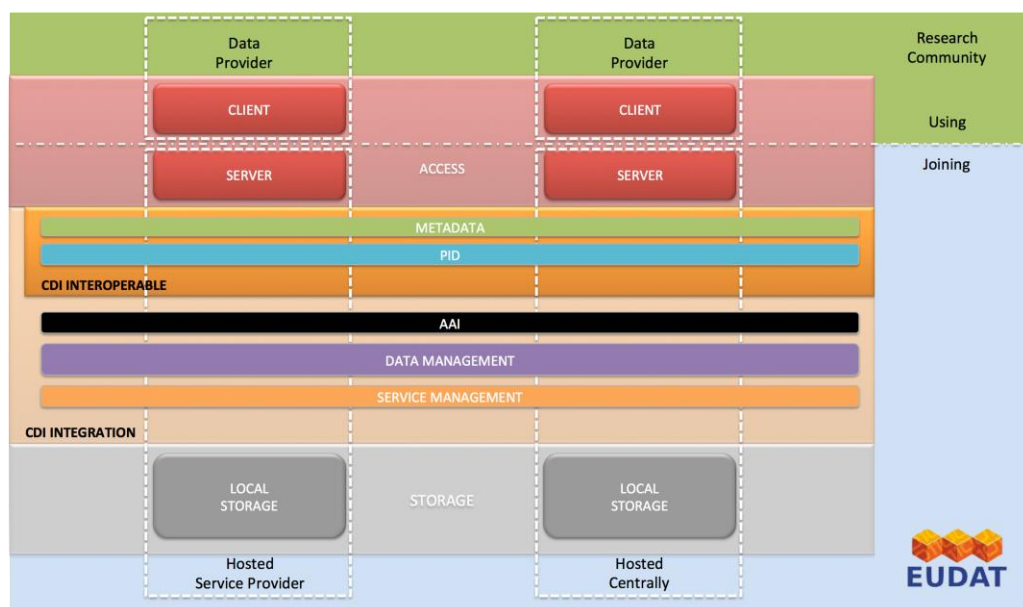


Figure 2: Schematic overview of the CDI architecture reflecting Using and Joining

¹¹ https://en.wikipedia.org/wiki/Digital_curation

3.1.3. CDI Nodes

The EUDAT CDI is realised through active and continuous collaboration between *service providers* and *research communities* contributing to the development and operation of an interoperable layer of common data services. A service provider is a partner that has signed the EUDAT CDI *collaboration agreement* and joined the CDI network providing services to its own designated user community. Thus the service provider can be either a *generic* or *thematic* node in the EUDAT CDI. The CDI nodes form the core of the CDI, contributing hardware and software to create the underlying service infrastructure.

- **Generic nodes** have regional, organizational or national mandates to support research, usually from different disciplines.
- **Thematic nodes** are discipline-specific organisations mandated to support a well-defined research community or group of customers and users.

3.2. EUDAT Data Architecture

In the context of TOGAF, the data architecture of an enterprise describes the structure and interaction of the major types and sources of data for that enterprise, along with its logical and physical data assets, and its data management resources. The following sections describe the current status of the discussions on the EUDAT Data Architecture: section 3.2.1 defines the EUDAT data continuum and section 3.2.2 describes a data model that is able to support the research community data models and that consequently must be supported throughout the CDI network.

3.2.1. EUDAT Data Domains

The EUDAT CDI model regards the CDI services as vehicles to support the research data lifecycle, and, in particular, the transition of research data between different domains – from *private* to *shared* and *published* domains. These domains have been described by the Australian National Data Service (ANDS) *data curation continuum* model¹² and can be regarded as a pyramid, which is shown in Figure 3. Between these domains or stages lie so-called curation boundaries – these boundaries indicate where curation decisions are needed as data transits from a private domain to a collaborative environment, and ultimately into the public domain through publication.

EUDAT primarily focuses on the *shared data domain* and therefore designed the CDI data model, as well as the service and policy standards, in such a way as to provide the best means for sharing and managing research data within that domain. EUDAT also provides some services and functionality for the *private* and *published* data domains to support both the management of data within specific levels of the data continuum and the transition of data between the different levels. For example, within the *published* data domain, EUDAT provides services for facilitating the transition from actively changing research objects to static data-of-record, which is to be linked inside publications and accessed for the purposes of data discovery.

The EUDAT data continuum consists of two levels or domains. These two data domains target services for storing information within the private and shared data domains respectively. In EUDAT terms, these data domains are called the registered data domain (that is, where data is shared) and the workspace data domain (for private data). These EUDAT data domains are defined as follows.

- In the **registered data domain** digital objects (for a definition of these, please refer to Section 3.2.2) are stored and managed in such a way that data carrying associated descriptive metadata is discoverable and can be referred to or retrieved using persistent identifiers. The registered data domain allows data managers to apply data management policies.
- The **workspace data domain** is used to store and manage temporary working copies of data, transient bitstreams of digital entities or objects that are not yet subject to data management policies. This workspace is designed to make it possible to update, modify and delete data as part of the normal research processes of individual researchers or larger collaborative groups. Within the workspace data domain, the individual researcher is mainly responsible for managing his/her data.

Figure 3 illustrates the relationship between the EUDAT registered and workspace data domains and the levels in the ANDS data curation continuum model.

¹² Australian National Data Service website at <http://ands.org.au/guides/curation.continuum> , accessed November 2015.

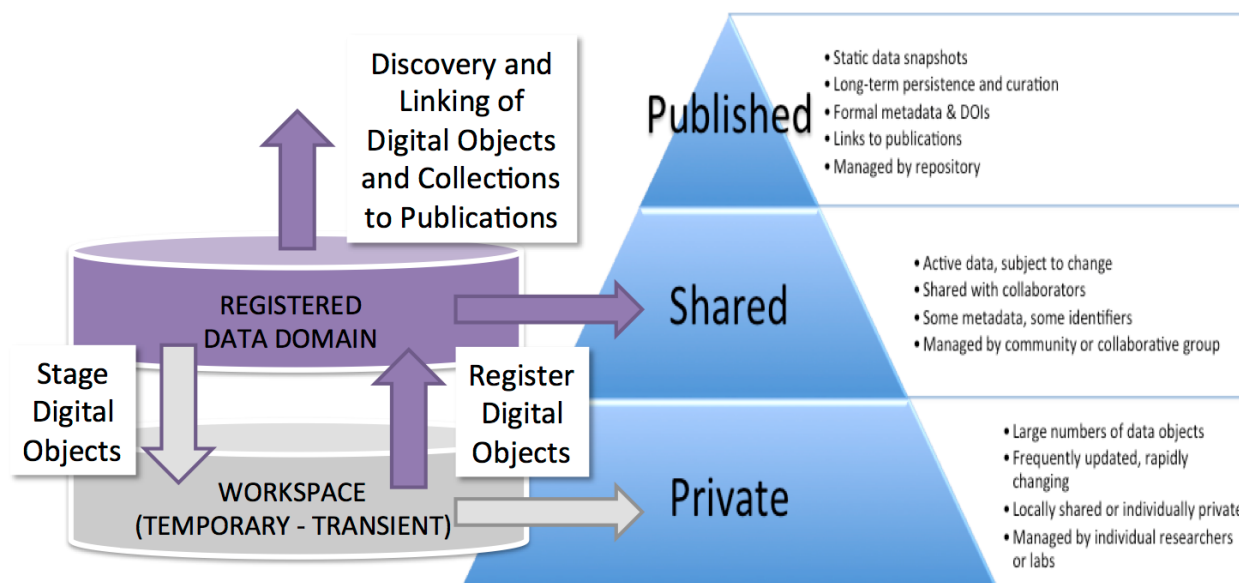


Figure 3: Relationship between EUDAT and ANDS data curation continuum

3.2.2. EUDAT Data Model

The data model proposed by EUDAT is based on that defined by the RDA-DFT¹³ working group. The RDA-DFT data model distinguishes between levels and types of objects (such as bitstreams, digital entities, digital objects, metadata and persistent identifiers) which are described in more detail in the core terms definition document¹⁴ provided by the RDA-DFT working group. However, the purpose of this section is to provide the reader with a high-level overview of the EUDAT CDI data model. A more complete description of the EUDAT model, which will deal with access rights, supported formats and so forth, is under preparation. In the meantime, a schematic diagram of the data model is shown in Figure 4. At the time of writing, the proposed data model is not yet fully supported within the CDI network and B2 services. The B2 services are in the process of being adapted to support the proposed data model.

The lowest form of data handled by EUDAT is the *bitstream*. Bitstreams are the way data enters the EUDAT CDI. Each bitstream is the representation of either a *digital entity* or a *digital object*, the difference being that the latter has *descriptive metadata* and a *persistent identifier* associated with it, while the former represents a 'working copy' of data which has no PID and need not have associated metadata. In terms of the ANDS Data Curation Continuum Model, digital entities belong to the lowest level in the pyramid, while digital objects can belong to the shared or published sections, depending on the level of maturity of the digital object. *Digital collections* are used to aggregate digital entities and digital objects, and are themselves digital objects with associated metadata and persistent identifiers.

Metadata within the EUDAT CDI is also considered as a digital object with its own unique persistent identifier. Thus, it is possible for a digital object to consist only of metadata with its PID. Metadata itself comprises at least two separate logical entities: user supplied metadata (descriptive metadata) and state information (system metadata) created and applied by the CDI to support the underlying data management functions. The latter type of metadata includes at least the size and checksum of the data, and is associated with a digital entity, digital object and digital collection. In the case of a digital collection, the metadata includes a means of identifying child objects within the collection. All the metadata is made available through a *metadata repository*, which is a searchable interface for finding registered objects, maintained within the CDI.

¹³ <https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html>

¹⁴ <https://b2share.eudat.eu/record/247/files/DFT3%20-%20snapshot%20of%20core%20terms.pdf>

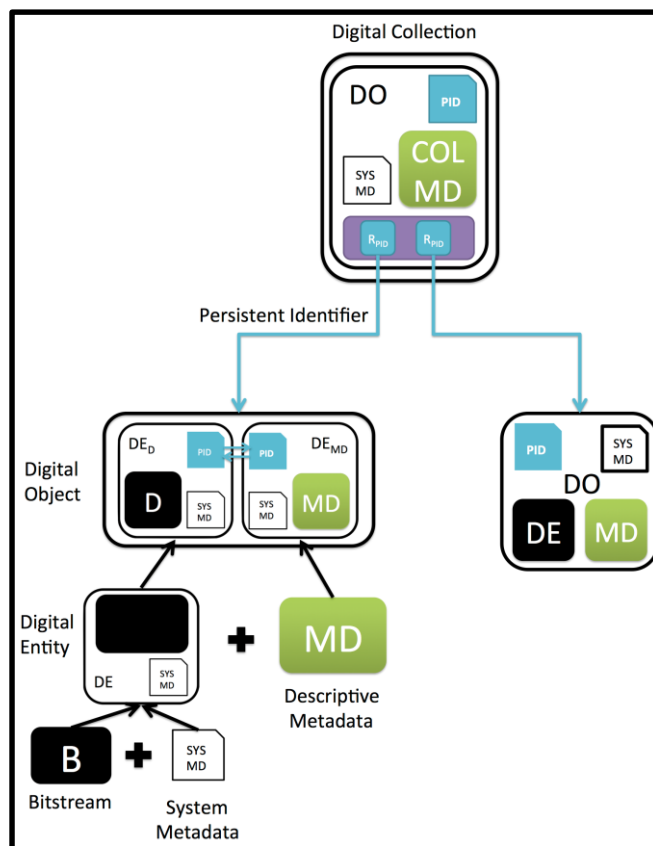


Figure 4: CDI data model (conceptual)

The EUDAT data model also accounts for what is known as a digital package – this is a digital object or a collection of digital objects or entities packaged into a single bitstream. Within a digital package, the structure, content and relationship between the individual objects, entities or bitstreams must be described. This can either be done via a structured bitstream (using, for example, HDF¹⁵ or NetCDF¹⁶ formats) from which metadata, object and relationship information can be extracted, or via constructed formats (such as zip or tar) which include a manifest (for example, METS¹⁷ and BagIt¹⁸) describing the structure and contents, along with the relationship between individual objects or entities in the package. The advantage of digital packages is that they make it possible to ingest digital objects and collections as a single bitstream. This can become useful and more efficient when uploading a collection consisting of many small objects. In the context of the Open Archival Information System¹⁹ (OAIS) model, digital packages can be used as submission information packages (SIP). Figure 5 shows logical representations of digital objects as bitstreams and digital packages, and of a digital collection as a digital package. The use of digital packages strongly depends on the data model adopted by the research community and/or data owner. Because of this, the CDI data model must be sufficiently flexible to support these kind of use cases.

¹⁵ <https://www.hdfgroup.org/>

¹⁶ <http://www.unidata.ucar.edu/software/netcdf/>

¹⁷ <http://www.loc.gov/standards/mets/>

¹⁸ <https://tools.ietf.org/html/draft-kunze-bagit-13>

¹⁹ <https://www.iso.org/obp/ui/#iso:std:iso:14721:ed-2:v1:en>

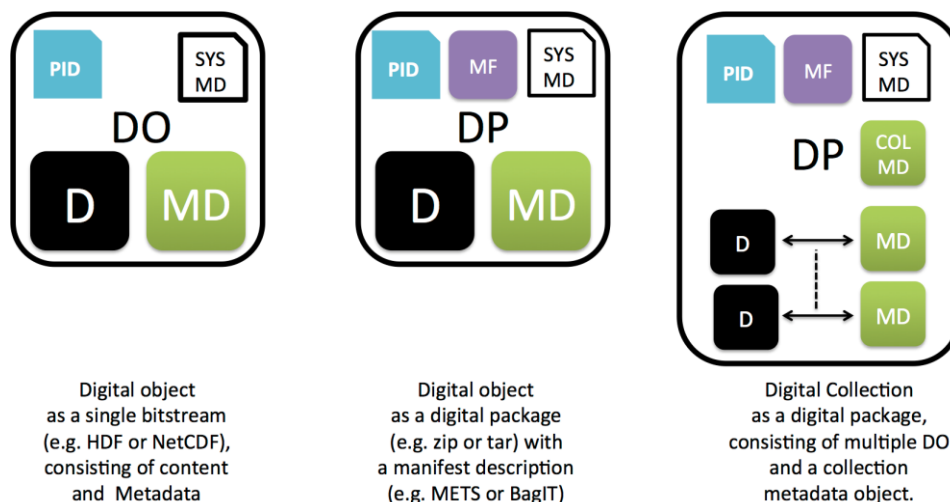


Figure 5: Logical representations of digital packages

In documentation also the term digital asset is used. A digital asset is not a specific kind of digital entity, object, collection, or package, but is used to depict all the aforementioned concepts with a single word. Annex B provides a table with the descriptions of the data types that are defined within the EUDAT CDI data model.

3.2.3. PIDs in Digital Objects

In the EUDAT data model a digital object is defined as a bitstream having descriptive metadata and that is referenced with a PID. As described in section 3.2.2, a digital object can be represented in different ways: (a) as a single digital object, (b) as several digital objects containing the raw data and the descriptive metadata or (c) as a package. When a digital object is uploaded to B2SAFE as a single object (a), a single PID will be registered. When a digital object (b) is uploaded as separate content and metadata entities, PIDs are generated for each digital entity. The relationship between the digital entities is maintained within the PIDs by cross-references. Figure 6 shows a logical diagram of the PIDs generated for a digital object (DO) uploaded as separate entities. The PID (that is, PID_D) generated for the digital entity containing the content is considered as the primary PID of the DO (namely PID_{DO}) and should be used to reference the DO. The PID record structure will support separate fields to cross-reference between the two PID records and to identify if a PID record is referencing a digital entity containing metadata or content.

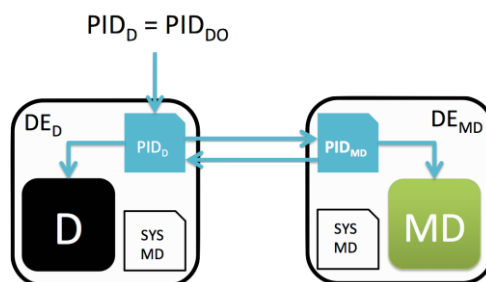


Figure 6: PID relationship diagram of a DO uploaded as two separate digital entities

When a digital package (c) is uploaded as a single object, how the digital package is stored depends on the relevant policy. The package could be uploaded as a single entity (a) or as separate entities (b). When a package is stored as separate entities (b), each entity is registered with a separate PID.

In the current practice of the B2SHARE service, a PID is assigned to the landing page of a data collection of one or more files that is described with metadata. In contrast, in the B2SAFE service a PID is registered for each individual digital entity.

3.3. EUDAT Application Architecture

According to the TOGAF definition, an Application Architecture describes the structure and interaction of an application as groups of capabilities that provide key business functions and manage data assets. When it

comes to the EUDAT Application Architecture, the EUDAT services are provided in the context of the EUDAT CDI network, in which some services are provided to specific research communities (which could be generic or thematic), some are provided as public services to individual researchers, citizen scientists or small research groups, and a number of auxiliary services are needed to run and operate the CDI as a whole. This section describes the CDI network (3.3.1), and explains which services are centrally provided (3.3.2), either as public or auxiliary services or within the service management infrastructure (SMI). This section also includes a description of the CDI layered architecture (3.3.3) and the node-layered architecture (3.3.4) which takes into account whether partners join the EUDAT CDI as interoperable or integrated nodes.

3.3.1. CDI Network

The EUDAT CDI includes a defined data model and a set of technical standards and policies adopted by European research data centres and research community data repositories to create a single European e-infrastructure of interoperable data services. The scope of the EUDAT CDI covers data management functions and policies for upload and retrieval, identification and description, movement, and replication of data along with maintaining data integrity.

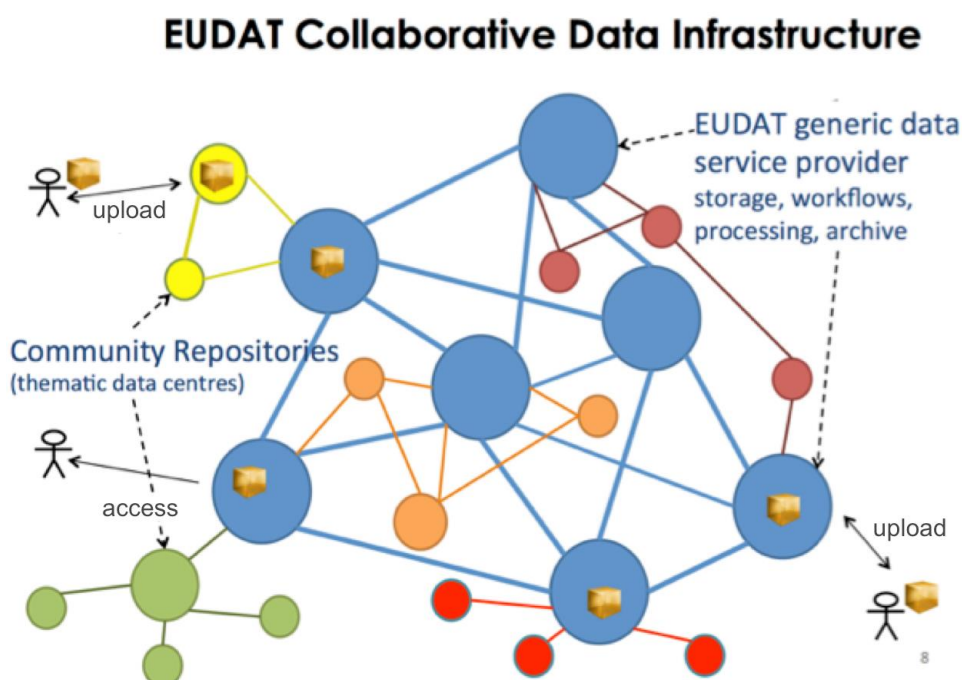


Figure 7: High level diagram of the CDI network with generic and thematic data centres

The EUDAT CDI consists of a set of services that are provided via operational service providers (that is, the CDI nodes). The CDI distinguishes between services which are provided locally by generic or thematic nodes for their own user domain, and some centrally provided services that are needed to operate the CDI or to provide central public services for researchers and citizen scientists that are not with specific research communities or institutions.

3.3.2. CDI Centrally Provided Services

The data management services within the EUDAT CDI network are provided by generic and thematic service providers. These services can be provided for four different purposes:

- to serve the local (generic or thematic) user community of the CDI node,
- as a public service to offer storage for long-tail data to citizen scientists or researchers who are not with a particular research community or centre,
- as an auxiliary service (within the EUDAT B2 service suit) that provides a central function within the CDI network, and
- as part of the EUDAT Service Management Infrastructure which operates the CDI network and supports the organization.

The next three sections discuss different types of services – the public services (3.3.2.1), auxiliary B2 services (3.3.2.2), and the services that fall within the Service Management Infrastructure (3.3.2.3).

3.3.2.1. CDI Public Services

In terms of the business architecture of the EUDAT CDI, the majority of users belong to organized research communities, or come from national or regional research institutions or other types of formal organizations. Such users are commonly served via generic or thematic nodes in the CDI. In this manner, the majority of the European researchers will have access to EUDAT CDI services. However, to serve the citizen scientists and researchers who are not affiliated with specific organizations, EUDAT will also provide a number of public services. These public services comprise a cloud storage service for managing and exchange personal data (B2DROP), a data repository service for sharing and publishing data sets (B2SHARE), and a metadata catalogue for data discovery (B2FIND).

- **B2DROP (b2drop.eudat.eu)** is a personal cloud storage service that lets users manage private research data. The B2DROP service provides synchronization functionality for each up-/download and synchronization of data across multiple devices. Users must register with the B2DROP service and will be given a default storage quota of 20GB. Via the B2DROP interface, users can easily define with whom they want to share and exchange data. Data stored in the B2DROP service falls within the *workspace* data domain.
- **B2SHARE (b2share.eudat.eu)** is a data repository service that enables users to easily describe, upload, share and publish large datasets. The B2SHARE service provides a Web User Interface (WUI) and an API interface for uploading and describing datasets. To make it easier to use the B2SHARE service, there are multiple metadata templates (specific to various research communities) that are provided to simplify the process of describing datasets. Each dataset that is uploaded to B2SHARE will be assigned a persistent identifier (PID) that is globally resolvable. This PID can be used to refer to the dataset. Users must register with the B2SHARE service, but there are no specific limiting quotas to the amount of data that a user can upload as the service has a *fair use* policy.
- **B2FIND (b2find.eudat.eu)** is the central metadata catalogue of the EUDAT CDI. This service makes metadata harvested from within the CDI, and from external metadata providers browse and searchable. The B2FIND service brings metadata from different research domains together to facilitate and support cross-disciplinary research. At the moment, the B2FIND service harvests metadata from B2SHARE and 14 external metadata service providers²⁰. All the metadata that is harvested is harmonized and mapped into 10 facets to make it easy to select and browse through the results. In addition, B2FIND provides a full text search and an API. Users do not need to register in order to use the B2FIND service.

The B2DROP and B2SHARE service stacks are also provided as technologies to generic and thematic CDI nodes that want to run a local instance of the services for their own user communities.

3.3.2.2. CDI Auxiliary B2 Services

The EUDAT CDI network is a distributed e-infrastructure, which is operated by a number of CDI nodes. To implement some functionalities (such as AAI, Policy Management, Data Types, and PIDs) within the CDI network, a number of auxiliary services are required. Of these auxiliary services, one or more instances of these services are centrally provided and operated by CDI nodes. It is important to be aware that it is not necessary for each of the integrated partners to run and operate each of these auxiliary services.

- **B2ACCESS (b2access.eudat.eu)** provides the central place for the registration and identification of users within the EUDAT CDI network. B2ACCESS functions as a bridge between globally provided identity domains (such as eduGAIN or social identities), community identity domains (for example the CLARIN or ENES IdP federation) and the CDI identity domain. This allows users to authenticate (that is, to identify themselves) via the trusted Identity provider for their research community or institute. Social identities are also enabled via the B2ACCESS service, but these identities have a lower level of assurance, compared with identities provided by research communities or eduGAIN institutes. As well as serving as a bridge between identity domains, the B2ACCESS service provides a

²⁰ The current list of metadata service providers: <http://b2find.eudat.eu/group>

bridge between different methods of authentication and authorisation to a common method used within the EUDAT CDI network and services.

- The **B2SAFE Data Policy Manager (DPM)** is currently under development. Within the CDI network the policy-based management function forms one of the backbones of the EUDAT CDI network and is provided by the B2SAFE service, which connects the CDI nodes. Organizational data managers can manage policies within B2SAFE centrally using the DPM service, which is a centrally operated portal. At the moment, B2SAFE supports policies for data replication, PID registration and data integrity checks. In the future more policies will be supported via the B2SAFE service and the DPM portal. The B2SAFE service also allows partners to implement site-specific policies, which are managed locally and not via the DPM portal. Although the B2SAFE-DPM portal is under development, the first release of the B2SAFE-DPM is planned for the next release period.
- **B2HANDLE (multiple instances)** is a service for registering persistent identifiers. For scalability and reliability reasons the B2HANDLE service is centrally provided by six of the CDI nodes. Persistent identifiers make digital collections and objects citable and findable in the long-term in such a way as to be independent of the current or future physical location of the specific digital asset. Therefore the use of persistent identifiers is essential for data management in a federated and evolving environment. Consequently the use of persistent identifiers is one of the core principles in the EUDAT CDI. For each digital object and digital collection stored in the CDI, a persistent identifier is registered. This PID can then be used for citing the item in publications and/or to find it while processing. If a digital object consists of separate digital entities (for content and metadata), each of those digital entities is assigned a PID and the relationship between the entities is maintained within the PIDs. The EUDAT CDI data management layer can be used to replicate digital collections and objects to other nodes within the CDI. This can be done for the purpose of having multiple copies for long-term preservation of the data or to optimize access to the data. In this context, PIDs are also used for location management and integrity checking.
- **B2HOST (which has multiple instances)** allows research communities to deploy and operate their own applications and data-oriented services on systems close to the relevant data storage location. The reasons for running such close to where the data is located can be that the volume of the data is too large to be efficiently transferred on demand to third party data processing and analysis facilities, or because licensing restrictions prevent even the smallest volume of data from being copied to a third party site which provides the compute facilities.
- The **Data Type Registry (or DTR, which is under development)** provides a central place where EUDAT and research communities can define and register data types. Although data types can be used in many different ways, they are commonly used by the research communities to define and build up descriptive metadata records. For each registered data type, a PID is generated. To register data types, the EUDAT CDI network is providing a central DTR service. In addition to being used to catalogue data types, the DTR can also be used in automatic workflows to interpret digital objects and validate scientific content to defined types, since the data types are globally resolvable via the PIDs.
- **B2NOTE (which is under development)** is a semantic annotation service which lets users make and maintain annotations for digital objects. The annotations are commonly made or related to the descriptive metadata associated with the digital objects. Because the semantic annotation service is closely related to descriptive metadata, EUDAT plans to integrate the service with the metadata catalogue B2FIND and with the data repository service B2SHARE service and technology.

3.3.2.3. Service Management Infrastructure

The EUDAT service management Infrastructure is considered to be one of the backbones of the CDI network. It provides functions for operating and managing services within the distributed CDI network, as well as providing services for accounting, site and service registries, helpdesk facilities and for monitoring availability and reliability. The following services come under this umbrella.

- The **Site and Service registry (creg.eudat.eu, CREG)** is, as a general rule, used to register service and service components in the site and service registry in order to preserve a global overview of the services provided within the EUDAT CDI network. This is mandatory for partners who join the CDI collaboration agreement (CA) as integrated partners. For partners that join the EUDAT CA as

interoperable partners, this is optional, but still regarded as a best practice. To manage registered services and endpoints, service managers must register themselves with the site and service registry.

- The **Availability and Reliability monitoring (cmon.eudat.eu, CMON)** service actively monitors the services within the CDI network to ensure that a high-quality level of service is provided by the CDI partners. To monitor the services, the principle of functional testing is being adopted. This means that the services are actively tested with regards to the functions that each service must provide; for this, service-specific tests are being designed. These tests are initiated by the central monitoring system from the perspective of an external user. To build up a map of sites and services that should be monitored, the monitoring system queries the Site and Service registry for the endpoints that are registered. Therefore it is essential for new services that enter the CDI to be registered within the Site and Service registry. If new types of services are being built, specific functional tests must be designed and developed to monitor those services. These functional tests must be added to the central monitoring system. In the near future, it is planned to have the CMON service replaced by a new central monitoring service called ARGON.
- The **Accounting service (rct.eudat.eu, RCT)** provides an overview of the number and volume of the digital assets managed within the CDI network. To gather this information, services that manage digital assets are requested to publish statistical figures (about the number of assets and the volume that is managed) in the central accounting system. Within the service delivery and resource provisioning process, a so-called data project is defined for each of the service and resource requests from a research community. The statistical figures are then reported on a project basis; in this way it is easy to provide information from the data project, research community and site-based views. As with the site and service registry, it is mandatory for integrated partners, and optional for interoperable partners, to publish these numbers. In the near future, it is planned to have the RCT service replaced by the Data Project Coordination Portal (DPCP).
- The **Data Project Coordination Portal (dp.eudat.eu, DPCP)** is a tool for managing and coordinating the realisation of data projects throughout their lifecycle – from the definition of a data management plan (project proposal) to its implementation. The DPCP will be the central service within the CDI network through which principal investigators (PIs) from research communities can request a specific data project, and which resource providers (CDI nodes) can use to offer service components, as well as storage and compute capacity. Project enablers will also be able to use the DPCP to obtain and manage information about service configurations. To manage the resources that are available or that are being used, the DPCP will gather statistical information about the number of assets and the volume being managed at the CDI nodes in a similar way to the RCT. In this respect, the DPCP will supersede the Accounting service (rct.eudat.eu).
- The **Helpdesk (helpdesk.eudat.eu, HELPDESK)** service is a service that provides an interface for research communities and users to interact with the support organization of the EUDAT CDI if they have issues or questions that need to be resolved. To guide users by providing relevant information and directing them to right branch within the supporting organization, the helpdesk system provides channels for each of the services and sites.

3.3.3. CDI Layered Network Architecture

As was explained in the section about the EUDAT CDI network, this network is built up from CDI nodes (such as service providers) that provide public or auxiliary central services and/or make services available to their own local or thematic user communities.

Within the CDI layered network architecture, seven distinct layers are recognized, of which five are defined as CDI integration layers. Figure 8 shows a high level diagram of the CDI network layered architecture. The different layers are as follows, and are explained in further detail in the subsequent sections.

- **Access:** The access layer provides the tools and interfaces for interacting with the EUDAT CDI services (such as the building blocks) within the CDI network. The tools consist of client tools that should be installed on the client side, along with API and Web-based User Interfaces (for example, CDI Gateways) run at CDI nodes.
- **Metadata:** The metadata layer is the layer that concerns the management and discoverability of metadata. It makes it possible to harvest metadata that is stored locally at the CDI nodes, and provides a central catalogue (B2FIND) that lets people discover digital collections and objects, both

within the EUDAT CDI and from collaborating partners. The metadata layer also provides services for registering data types and for making annotations of and on metadata.

- **Persistent Identifier:** The persistent identifier layer provides a building block (namely B2HANDLE) for registering, managing and discovering persistent identifiers within the CDI network. Because PIDs are intrinsic to the registered data domain, which is the primary focus of the CDI network, the usage of PIDs is an integral facet of data management within the CDI network, and hence many services interact with the PID layer.
- **AAI:** The AAI layer provides a means for authentication and authorization within the EUDAT CDI network. It consists of a building block (B2ACCESS) that is integrated with the external National Research and Education Network (NREN), and with research community and social identity federations, in order to provide single identities for users within the CDI network. The AAI layer also converts users' credentials to credentials that are accepted by the CDI services when needed. In addition, the layer comprises building blocks that make it possible to enforce access rules on digital objects independently of their location within the CDI network.
- **Data Management:** The data management layer is the layer where data management policies are implemented, for example, for the registration of PIDs, replication of digital objects, or integrity checks. This layer also maintains the relationships between content and metadata, and between the logical layer and the storage layer where the bitstreams are physically stored. Therefore the data management layer comprises a building block (B2SAFE) that implements a policy engine and which is integrated with all the layers of the CDI architecture.
- **Service Management:** The service management layer does not provide logical functions for the management of digital assets within the EUDAT CDI network, but rather it provides services that are essential for the operation of the CDI services and for the organization of the CDI network. This layer provides building blocks, for example, for the registration of sites and services (CREG), availability monitoring (CMON), resources accounting (RCT) and helpdesk services (HELPDESK). Because the service management framework provides central services for operating the CDI network as a whole, in principle, all the CDI services interact with the services provided by the service management layer in one way or another.
- **Storage:** The storage layer is concerned with providing the physical storage resources within the EUDAT CDI network and is also involved in the management of the physical bitstreams of the digital assets within the network. Therefore this layer provides building blocks as physical storage resources that are located at and provided via the CDI nodes.

As is apparent from these short descriptions, the layered service architecture recognizes different abstract building blocks.

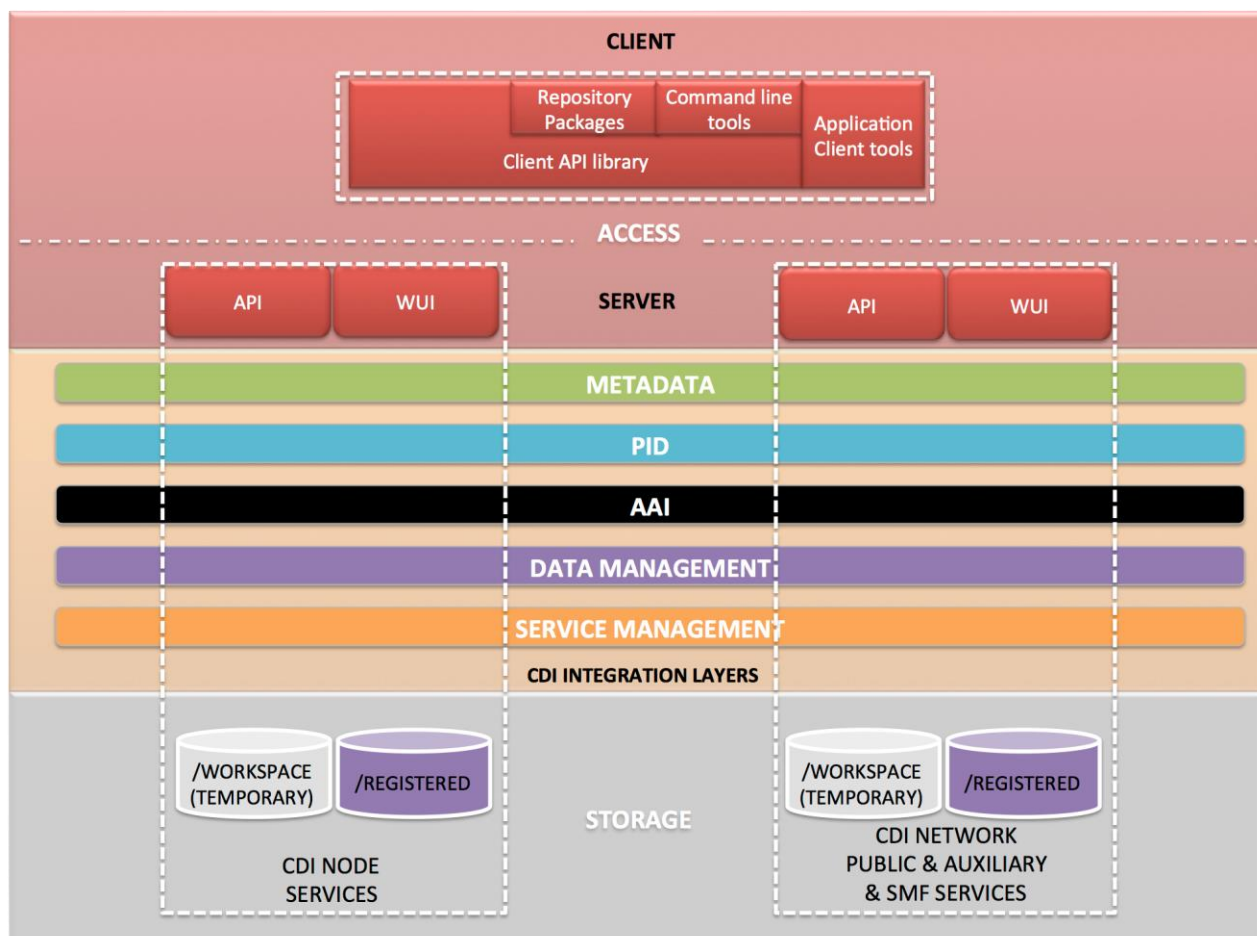


Figure 8: CDI Network layered architecture

In the next two sections the CDI architecture is explained in the context of the interoperable and integrated CDI nodes.

3.3.4. CDI Node Layered Architecture

As described in the section about the CDI Service Provider Levels (3.1.2), two different levels of collaboration with EUDAT (namely interoperable and integrated) are distinguished – each of these involves different levels of responsibilities and demands a different level of integration with the CDI network. In the next two sections, the differences between the interoperable and integrated levels will be explained from an architectural point of view.

3.3.4.1. Interoperable Node Architecture

A minimum requirement for interoperable CDI nodes is to have a data repository in which digital assets are preserved and/or curated. The digital collections and objects in the repository need to be registered and it must be possible to identify them using persistent identifiers²¹. Interoperable partners must make the metadata that is hosted in their data repository service harvestable and discoverable through EUDAT's B2FIND service. It must always be possible to resolve the persistent identifiers to arrive at the landing page or API service associated with the digital objects or collections. Any users with sufficient access rights must be able to access the digital assets.

The architecture of an interoperable CDI node is shown in Figure 9.

²¹ Note that the persistent identifier does not need to be a European Persistent Identifier Consortium (EPIC) handle, as used within EUDAT, but could be a PURL, DOI or URN:NBN for example.

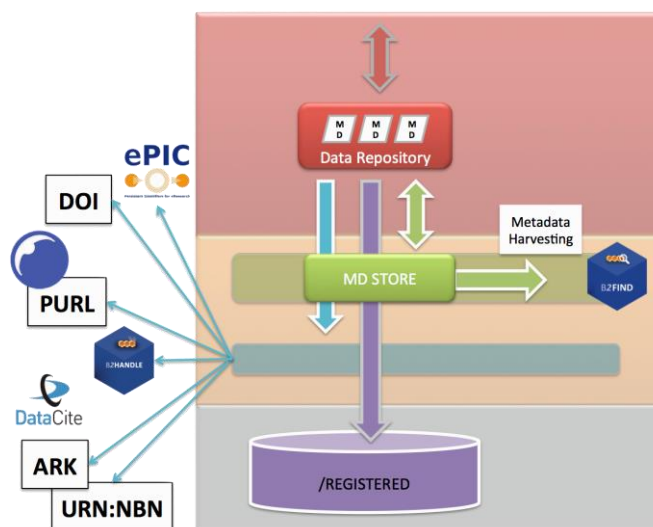


Figure 9: Architectural diagram of an interoperable CDI node

- **Integration with the Access Layer**
 - When metadata descriptions are harvested, it must be possible to resolve the persistent identifiers that are used in them to a service that is accessible via the internet. Therefore interoperable CDI nodes must have a data repository service that either provides a WUI in which PIDs resolve to a landing page, or an API service that the PIDs resolve to and where the digital assets can be accessed from a programmable interface. This does not mean that all data must be openly accessible, although EUDAT does support the Open Access principle; access rules can be applied.
- **Integration with the Metadata Layer**
 - Interoperable CDI nodes must make it possible for their metadata to be harvested by the B2FIND service. For this to happen, the data repository service must be extended with an OAI-PMH endpoint. If that is not possible, other means of harvesting can be discussed.
- **Integration with the Persistent Identifier Layer**
 - It is mandatory for interoperable CDI nodes to use persistent identifiers to make it possible to discover digital assets and to refer to them. EUDAT is aware that different PID systems (such as EPIC, DOI²², ARK²³, PURL²⁴ and URN:NBN²⁵) are available and in use by various research communities. EUDAT just requires that the PIDs that are used must be globally resolvable via the internet. And hence, for interoperable partners who join the CDI, EUDAT does not prescribe a specific PID system, although PIDs based on the Handle system are generally preferred. If new and prospective partners do not yet use PIDs, EUDAT can assist those partners to implement PIDs within their local data repository services, using the B2HANDLE PID service provided by EUDAT.
- **Integration with the Storage Layer**
 - Because it is mandatory for interoperable CDI nodes to maintain descriptive metadata and persistent identifiers for digital assets, all such digital assets are considered to be digital objects that are part of the registered data domain.

It should be noted that an interoperable CDI node does not need to commit to anything or integrate in any way beyond the minimum requirements that have been specified in order to join the EUDAT CDI. It is optional for an interoperable CDI node to integrate with other services in the CDI network (in order to add additional functionality). It is optional for interoperable nodes to integrate with the CDI AAI infrastructure (B2ACCESS), to implement the policy-based Data Management service (B2SAFE) or integrate with the CDI Service Management Infrastructure to register the CDI node services in the EUDAT site and service registry, or to make use of the CDI monitoring or accounting and helpdesk services. However, the more EUDAT CDI

²² <https://www.doi.org/>

²³ <https://confluence.ucop.edu/display/Curation/ARK>

²⁴ <https://purl.org/docs/index.html>

²⁵ <http://www.ietf.org/rfc/rfc3188.txt>

functionality that is adopted, the easier it is for the interoperable node to become an *integrated* node later on.

3.3.4.2. Integrated Node Architecture

From the business architecture point of view, integrated partners are expected to integrate their data services that provide resources for either the registered and/or workspace data domain with the different EUDAT services at the different integration layers within the CDI network. This means that, in addition to fulfilling the requirements of an interoperable partner (that is, making use of persistent identifiers, having a local metadata store and making metadata harvestable), integrated partners need to enable their local data infrastructure and resources via EUDAT's policy-based data management service (B2SAFE), comply with the EUDAT common policy framework, implement a suitable service for persistent identifiers (such as B2HANDLE or an equivalent), and integrate their infrastructure with the CDI AAI infrastructure (so as to make it possible to access identities registered with the EUDAT CDI) and with the common EUDAT Service Management Infrastructure for site and service registry, monitoring, accounting and providing helpdesk services. Integrated CDI nodes must also apply the EUDAT CDI gateways on top of the integrated infrastructure to provide a common method of access for the user communities. Figure 10 shows a high level diagram of the layered service architecture for an integrated CDI node.

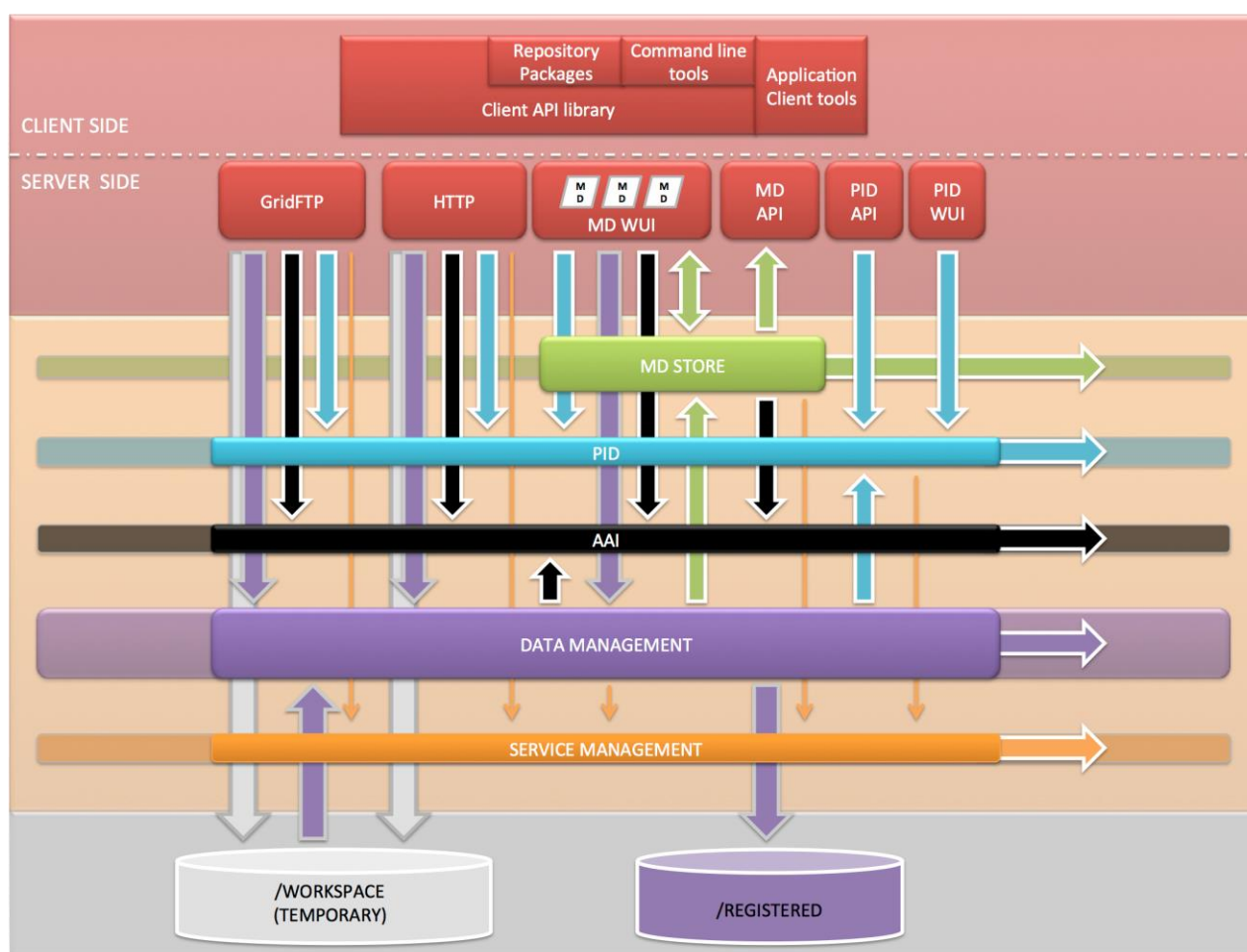


Figure 10: CDI Integrated Node layered architecture

As mentioned earlier, integrated partners are expected to be integrated at each level of the CDI layered architecture. Because these integrations have already been described extensively in the previous sections, they are only summarised briefly here from the point of view of an integrated node.

- Integration with the Access Layer
 - An integrated CDI node must implement the EUDAT CDI common API services on top of all the local data infrastructures that are made accessible to the CDI network. This means that the common HTTP API (B2STAGE) should be implemented on top of the local data infrastructure. This allows the research communities and users of the EUDAT CDI network

to easily integrate and/or access data maintained with the CDI network via a common set of client tools. Implementing the GridFTP API service is optional for integrated service providers.

- To provide a human friendly interface for accessing and finding digital objects and collections at the level of a CDI node, a WUI interface is provided on top of the integrated infrastructure. This gives users access to data repository functionality at the level of a CDI node.
- Integrated service providers must implement the EUDAT CDI persistent identifiers. They can make use of one of the centrally provided services (such as B2HANDLE or an equivalent) to do this. Therefore it is not mandatory for integrated service providers to install and configure a local B2HANDLE service. Although it is not mandatory to host a local B2HANDLE service, many integrated partners nevertheless implement the EUDAT PID service to register and manage local persistent identifiers. If a local B2HANDLE service is provided, the PID service provides a RESTful²⁶ API (for example, Handle and/or EPIC) and a WUI service.
- Integration with the Metadata Layer
 - As with the interoperable CDI nodes, integrated nodes need to store (a copy) of the descriptive metadata of digital objects and collections in a local metadata store. This local metadata store needs to be set up so that it can be harvested centrally by the B2FIND service. This metadata is then made accessible in a user-friendly manner via a WUI interface.
- Integration with the Persistent Identifier Layer
 - The use of persistent identifiers is one of the core principles in the EUDAT CDI network. For each of the digital objects or collections within the CDI, a persistent identifier is registered. These PIDs are also used for location management and integrity checking within the CDI data management layer. Integrated CDI nodes must acquire a globally resolvable prefix (for example, a Handle or an equivalent) that is registered and administrated in the CDI B2HANDLE service. It is important to be aware that this prefix is used for registering digital assets that are managed within the B2SAFE service and that the prefix is essential for the correct functioning of the current common CDI policies. As mentioned previously, it is not a mandatory requirement for an integrated CDI node to run a B2HANDLE service locally, but many CDI nodes do. Some examples of possible reasons for running a local B2HANDLE service are for reasons of performance, reliability or security, or because the CDI node wants to provide a PID service to their local user community.
- Integration with AAI Layer
 - Integrated CDI nodes do not have to install the AAI service (B2ACCESS) themselves as this is centrally operated within the EUDAT CDI. However integrated partners should integrate those of their local data and CDI services that are enabled via the CDI network with the CDI AAI. In this way, centrally registered users within the CDI can have access to the services provided by the partners via a common identity. Using service domains configured in the B2ACCESS service, the integrated node maintains full control over who has access to the local service.
- Integration with Data Management Layer
 - Integration with the data management layer means that CDI nodes install and configure the data management service (B2SAFE) locally at the CDI node. B2SAFE must be integrated with the local storage infrastructure, and a storage area for the registered data domain needs to be provided. It is not necessary for integrated partners to run the DPM portal, as this portal is provided centrally. Integrated CDI nodes must apply the common CDI policy framework within the B2SAFE service. At the time of writing, the common CDI policy framework consisted of policies for registration with PIDs, data integrity verification via checksums and the replication of digital objects. In terms of the implementation of the replication policy, integrated CDI nodes need to establish network links with the other integrated CDI nodes to make it possible to transfer data between the nodes.

²⁶ https://en.wikipedia.org/wiki/Representational_state_transfer

- **Integration with Storage Layer**
 - Registered data domain: Integrated CDI nodes are required to make storage resources for the registered data domain available via the data management service. This involves implementing the B2SAFE service on top of the local data infrastructure. The storage resources that are made available must include resources for the preservation of bitstreams (which can be high latency) and low latency storage for bitstreams where fast access is needed (for example, for metadata or user requests) or for data that is best maintained on low latency storage (such as small sized bitstreams).
 - Workspace data domain: Integrated CDI nodes may choose whether or not they will provide resources for the workspace data domain. However, if they do decide to provide storage resources for the workspace data domain, the storage resources must be made available via one of the CDI Gateway APIs.
- **Integration with the Service Management Layer**
 - Site and service registry: For each integrated CDI node, a site entry is created within the site and service registry. A CDI node describes basic information (such as contact information, or responsible people and roles) in the site entry. It is mandatory to register all the service endpoints that are enabled within the CDI network.
 - Availability and reliability monitoring: To build up trust with EUDAT's users and partners, the quality of the service provisioning is actively monitored within the CDI network. Integrated CDI nodes are required to make it possible for each service and service endpoint that is integrated into the CDI network to be overseen by the central monitoring service. Where necessary, monitoring plugins and/or special administrative accounts for monitoring should be provided.
 - Accounting: To monitor resource allocation, the CDI has a central accounting system. Integrated partners are required to publish accounting figures (such as the volume and number of objects) in the central accounting system for each storage resource made available within the CDI network.

In addition to integrating the local data infrastructure with the CDI network, integrated partners are also required to sign an EUDAT Operational Level Agreement (OLA), to provide on-site support staff to address any site-specific issues in a timely manner (as defined in the OLA), to agree to software updates released by EUDAT and to maintain hardware security patches to an acceptable level.

3.3.5. Service Portfolio

The previous sections about the CDI application architecture have identified many building blocks of the CDI (such as types of services). These building blocks, and the integration between these building blocks, were described in general terms in the previous sections from an architectural point of view.

In this phase of the EUDAT project, a service management approach is being adopted on FitSM²⁷ to provide sustainability of the EUDAT CDI and its services. The Service Portfolio and Service Level Management processes are described in deliverable D2.1 EUDAT Service Portfolio Processes Definition and SLA Template Set. As part of the Service Portfolio Management, a Service Portfolio is being developed. The Service Portfolio includes meta-information about services, such as their value proposition, target customer base, descriptions, technical specifications, cost and price, risks to the provider, service level packages offered and so forth. From a management point of view, services and the evolution of services can be managed throughout the whole lifetime of a service via the Service Portfolio.

From an architectural point of view, when a building block is identified and a decision is made to start development, the building block is registered in the Service Portfolio.

²⁷ <http://fitsm.itemo.org/>

3.4. Supporting Metadata within the CDI

Metadata is structured information that describes some data; therefore metadata is often called data about data. To research communities, metadata is a vital part of communication to facilitate research. To enable long-term preservation of data to be reliable, metadata is essential, and it is also vital for understanding the output from research. Metadata often provides information about who generated the data and why the data was generated. It usually describes the content of the data in a simple way that saves people from having to read and interpret the whole content in order to find out what it is about. This makes it much easier to discover and understand the data that results from research. Administrative metadata makes it easier to perform data management tasks or to determine who has access to data, and how data can be used and reused. Technical information (such as structural or system metadata) provides information about how to interpret and/or validate the content of data.

Within the domain of registered data and in the context of the data model that EUDAT has defined, metadata is essential and intrinsic to a digital object. Since EUDAT's main focus is the registered data domain, it is absolutely essential for the EUDAT CDI network and CDI services to support the management of metadata.

EUDAT provides common building blocks for data management which can be adopted and adapted to the needs of the research communities. To heighten the overall support for metadata that is provided within the CDI network, the support of metadata has been one of the main topics of discussion in the architectural discussions held by the EUDAT Technical Committee and amongst the related service development teams (namely B2SHARE, B2SAFE, and B2STAGE). As a result from these discussions is that the support for descriptive metadata within the CDI network is put on the development roadmap.

3.4.1. Metadata Types

Metadata can be classified²⁸ according to various standards and terminologies, and it is out of the scope of this document to provide a complete coverage of such classifications. The EUDAT data model recognizes different levels of metadata according to the following definitions.

- **Descriptive metadata** is metadata that is required for discovering data, and for searching for and retrieving data. This type of metadata is provided by end users.
- **Administrative metadata** is metadata that is used for data management purposes, and includes, for example, details about intellectual property rights, access, provenance and so forth.
- **Structural metadata** is metadata that describes the internal structure of the data.
- **System metadata** is metadata that describes some basic attributes of the data that are related to the physical implementation of the technology used for storing the data (such as the name of the file, the internal path in the system, the size of the file, or a checksum).

These different levels of metadata play different roles and are used in different ways in the workings of the EUDAT CDI network and services.

3.4.1.1. Descriptive Metadata

Descriptive metadata is metadata that is provided by the end users to describe any kind of information. The relevant information could be digital content (such as digital objects and collections), but it could also be physical items (for example, museum pieces or DNA samples) or logical entities (like institutions and organizations) or people. The descriptive metadata is defined and described by the end user who chooses the terms and vocabulary familiar to his/her research community. To make it easier to produce this kind of metadata, many research communities have already defined metadata concepts and standardized vocabularies.

EUDAT's research communities are heterogeneous and understanding metadata schemas requires domain-specific knowledge about the data. Hence, it is beyond EUDAT's mandate to develop tools to support all metadata concepts defined by the research communities in a semantic way. To extend the available support for descriptive metadata, multiple levels of descriptive metadata were identified for which support must be provided. The levels of descriptive metadata range from metadata templates supported by EUDAT, through metadata objects that are defined by research communities and are interpretable or that are identified but

²⁸ <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>

cannot be interpreted, all the way to completely unidentified metadata objects. Figure 11 shows some examples of the different levels of descriptive metadata – more extensive descriptions can be found in the table in Annex C.

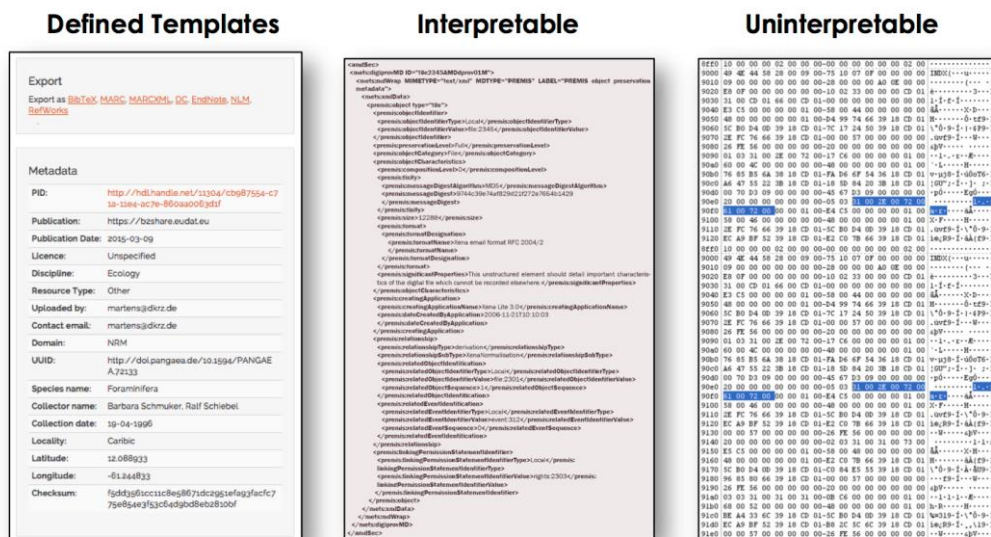


Figure 11: Examples of the different levels of descriptive metadata

3.4.1.2. Administrative Metadata

Administrative metadata is metadata required for long-term data management; this metadata provides additional information about intellectual property rights, access, provenance and so forth. In the realm of the EUDAT CDI network, administrative metadata is essential to internal workings of the data management policies. For example, to replicate data, the locations of the physical data are needed and hence need to be part of the administrative metadata that the replication policy can employ. Similarly the integrity check policy employs the checksums to verify the integrity between bitstreams across different physical locations. Because administrative metadata is an intrinsic and essential element for performing data management, administrative metadata is associated with each individual digital asset stored within the CDI network.

3.4.1.3. Structural Metadata

As already explained in section 3.2.2, the EUDAT Data model accounts for different types of digital assets (digital entities, objects, collection and packages). These digital assets can be constructed or aggregated in many different ways. For example, a digital package can contain a single digital object in which two digital entities (such as content and descriptive metadata) are aggregated into a single package, or it can contain a whole collection with many digital entities. To understand how a digital package is constructed and to understand the meaning of individual entities, structural metadata must be provided. When a bitstream is a complex digital asset, structural metadata describes the internal structure. To describe structural metadata different standards (for example, METS and BagIt) have been emerging. The data entity in which structured metadata is described is commonly identified as a manifest file. At the time of writing, the need for having and handling structured metadata was identified. Therefore, a manifest was added to the EUDAT data model. The standard that will be used is still under discussion. Because structural metadata describes the internal structure of a digital asset, structural metadata is associated with a digital package, as shown in Figure 5.

3.4.1.4. System Metadata

System metadata is metadata which describes basic attributes of the data related to the physical implementation of the storage technology (such as a file name, system internal path, file size, or a checksum). This metadata is automatically generated and maintained by the system or technologies in which the data (for example, bitstreams) are stored. When bitstreams are stored on a POSIX²⁹ compliant file system, information which is commonly provided via the UNIX/Linux *stat()*³⁰ system call is considered to be system

²⁹ <http://pubs.opengroup.org/onlinepubs/969919799/>

³⁰ [https://en.wikipedia.org/wiki/Stat_\(system_call\)](https://en.wikipedia.org/wiki/Stat_(system_call))

metadata. In the case of an iRODS system (such as B2SAFE), information returned by the *ils*³¹ command can be considered as system metadata. System metadata is associated with each individual bitstream stored within the CDI network. Within the EUDAT data model, a bitstream with associated system metadata is called a digital entity.

3.4.2. Local Metadata Store

In addition to identifying the different types of metadata (see above), metadata must be maintained, made easily-accessible, harvestable and discoverable. To this end, the concept of the local metadata store, or local B2SHARE service, was developed – it functions as an extension of the B2SAFE service and will be hosted at each integrated CDI node. When a WUI (web user interface) is applied on top of the local metadata, the local metadata store will function as a data repository service at a CDI node, in a similar manner as with the current B2SHARE service.

Metadata is automatically extracted and ingested into the local metadata store using policies for metadata extraction that are applied to uploaded digital objects. In this way the local metadata storage of a CDI node will be automatically filled. Because the local metadata store is automatically being harvested, digital objects that are uploaded to it will automatically be made discoverable via the central catalogue (namely B2FIND).

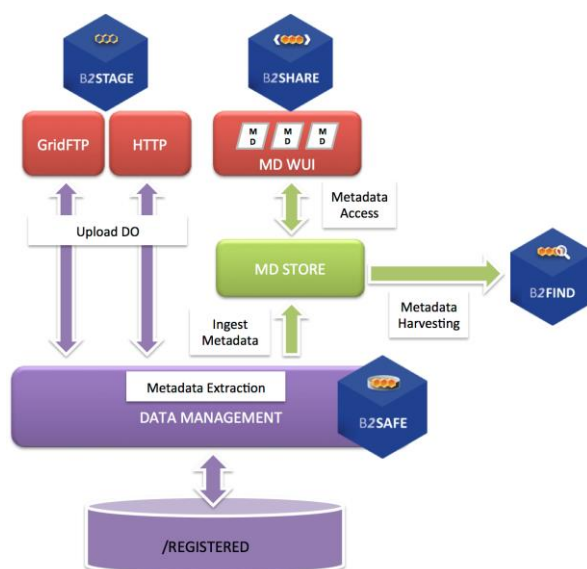


Figure 12: Logical diagram of the local metadata store for the B2SAFE, B2SHARE and B2STAGE API services

Since it is not always known beforehand which metadata schema is used to describe the digital objects that are uploaded, the local metadata store, as well as the WUI and API interfaces, need to support different metadata schemas in a flexible way. To implement the local metadata store, version 2 of B2SHARE (which is based on Invenio v3) and graph database technology (as a possible uptake from the graph-based database technology which is being investigated by EUDAT task T9.2) are being evaluated at the time of writing.

3.4.3. Data Flow

The main use case for the EUDAT CDI is to provide a data infrastructure where research communities and users can store, maintain and retrieve digital objects, and for which the CDI nodes provides different API services (namely CDI Gateways). Within the CDI Gateways, three types of interfaces – HTTP RESTful API, Web User Interface and GridFTP – are distinguished, and each types of API provides different capabilities for interacting with the data.

The API services do not all provide the same level of functionality. The HTTP RESTful API and the WUI provide a rich range of functions for up- and downloading and managing digital assets, whereas the GridFTP API service just provides basic functionalities for up- and downloading digital assets. Also the context in which the API services are accessed plays a role. The HTTP RESTful API and GridFTP API services are best-suited for automated access within workflows or with command line client tools, whereas the WUI is intended for

³¹ <https://docs.irods.org/master/icommands/user/#ils>

human interaction from web browsers. Depending on the context, the data flow and interaction between the client and the CDI Gateways will differ.

The interactions from the different client contexts and CDI Gateways are explained in the following diagrams.

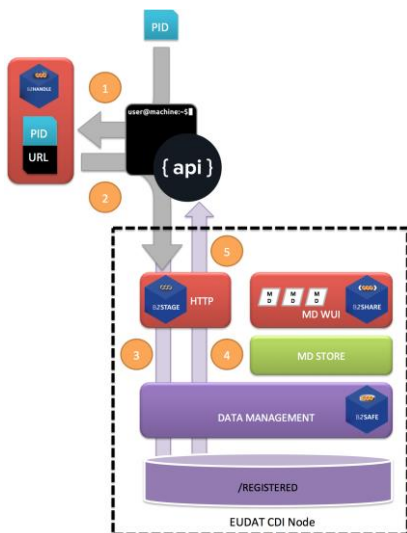


Figure 13: Dataflow from a CLI/API

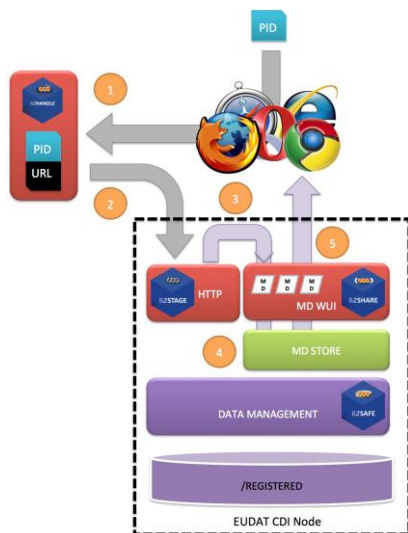


Figure 14: Dataflow from a browser

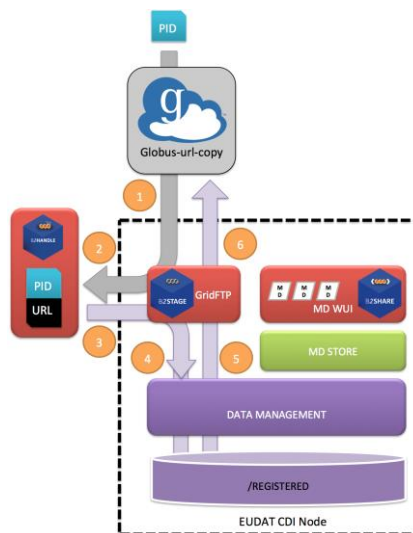


Figure 15: Dataflow from a GridFTP client

1. PID is provided as input to the HTTP client, for example, curl
2. The client contacts the PID resolution system, which resolves the PID to a URL and redirects the client to the URL endpoint (HTTP API)
3. The HTTP API service understands the URL and gets the DO from B2SAFE
4. B2SAFE returns the DO bitstream to the HTTP API service
5. HTTP API service returns the DO bitstream to the client

1. PID is provided as input to a browser
2. The browser contacts the PID resolution system, which resolves the PID to a URL and redirects the browser to the URL endpoint (HTTP API)
3. HTTP API service understands access from the browser, via content negotiation, and redirects access to the B2SHARE WUI
4. B2SHARE WUI retrieves descriptive metadata record of the DO from the local metadata store
5. B2SHARE WUI returns a rendered HTML page with the DO metadata content to the browser

1. PID is provided as input to the GridFTP client, for example, globus-url-copy
2. The client contacts the GridFTP API service
3. The GridFTP API service contacts the PID resolution system, which resolves the PID to a URL
4. The GridFTP service understands the URL and gets the DO from B2SAFE
5. B2SAFE returns the DO bitstream to the GridFTP API service
6. GridFTP API service returns the DO bitstream to the client

Depending on the context in which a digital object is accessed, the CDI Gateways try to predict (for example by content negotiation) the expected behaviour and return a DO in the expected format. In the case of access from a CLI/API, either via HTTP or GridFTP-based, the digital entity with the content of the DO is returned. When a DO is accessed via a browser, the browser will be automatically redirected to the WUI, to the landing page of the DO.

3.5. Outlook

From the start of this phase of the EUDAT project, the discussions to define the architecture of the EUDAT CDI, the data model and the interaction and integration between the services were intensified via the Technical Committee. During the first year of this part of the project, these discussions provided a good basis for the data model and the basic CDI architecture, and also provided strategic directions for the development

of the local metadata store and the functions to be supported via the HTTP RESTful API service. The discussions on the CDI architecture and related aspects are far from finished and will continue throughout the next years of this phase of the EUDAT project. These discussions will focus on further defining the CDI architecture, the data model and data flows, and on the definitions of the interfaces between the individual services and the integration between the existing and new services.

4. DATA ACCESS AND RE-USE

The purpose of the Data Access and Re-use service area is to maintain and improve the services that are available for discovering, accessing, sharing, and re-using research data in the EUDAT CDI by consolidating and further developing the existing services for collaboration (B2DROP), public sharing and publishing of data (B2SHARE) and data discovery (B2FIND). This service area does also provide a Federated Authentication and Authorisation Infrastructure (B2ACCESS) and will develop new services in the area of registries (Registry Services).

The main objectives for the service area during the first twelve months of this phase of the EUDAT project were to bring the Federated AAI service into production and to begin the process of integrating the respective services in this service area and in other areas with each other.

At the start of the project, B2FIND was already harvesting information from B2SHARE and thus some integration was already in place. During this first twelve-month period, we brought B2ACCESS into production and integrated it with B2SHARE. Subsequently, the integration of other services with B2ACCESS started, and authentication with B2ACCESS is being prepared for B2DROP, B2SAFE and for several of the supportive services within the CDI.

The integration between B2DROP and B2SHARE was developed and will be taken into production during the second quarter of 2016. The integration of B2SAFE and B2STAGE with B2ACCESS and B2SHARE is planned for next year.

In the course of the first year of this phase, the service area also participated in the process of designing the common infrastructure's usage of persistent identifiers, a task that is not concluded yet, and in the design and definition of a common EUDAT HTTP REST interface for connecting the EUDAT service instances with each other and with the external world.

Common achievements for the entire service area during 2015 include project planning and organisational preparations for EUDAT2020. All the teams introduced their services and the current states at the EUDAT2020 kick-off event in March 2015. To foster the integration between services, the B2SHARE team hosted a hackathon session in October which was attended by other service area members from the work package, and dedicated meetings were held for B2FIND and B2ACCESS.

4.1.1. Future Work in the Service Area

At the beginning of the next twelve-month period of EUDAT2020, we are looking forward to the release of B2SHARE 2. In this new version we focused on a new architecture that will make B2SHARE much easier to deploy so that it can be taken into production at more sites. This release will also include the integration between B2SHARE and B2DROP.

The integration between the services will continue. B2ACCESS will support more and more services for Federated AAI. A metadata store for B2SAFE is in the planning phase, and with that it will be possible to also integrate B2SHARE and B2FIND with B2SAFE. The common CDI HTTP REST API will be developed during the second year of this phase of the project, and the services will be integrated with the common API. The Registry Services task group is developing a data type registry which will be taken into production and utilised by other EUDAT services within the course of the project.

Further details are given in each of the following service sections.

4.2. Data Repository – B2SHARE

B2SHARE is EUDAT's service for storing, sharing and publishing so-called "long-tail" data, that is, small and medium scale research data in various formats, which is usually not covered by institutional data preservation policies. Figure 16 shows a screenshot of the upload page of the B2SHARE service.

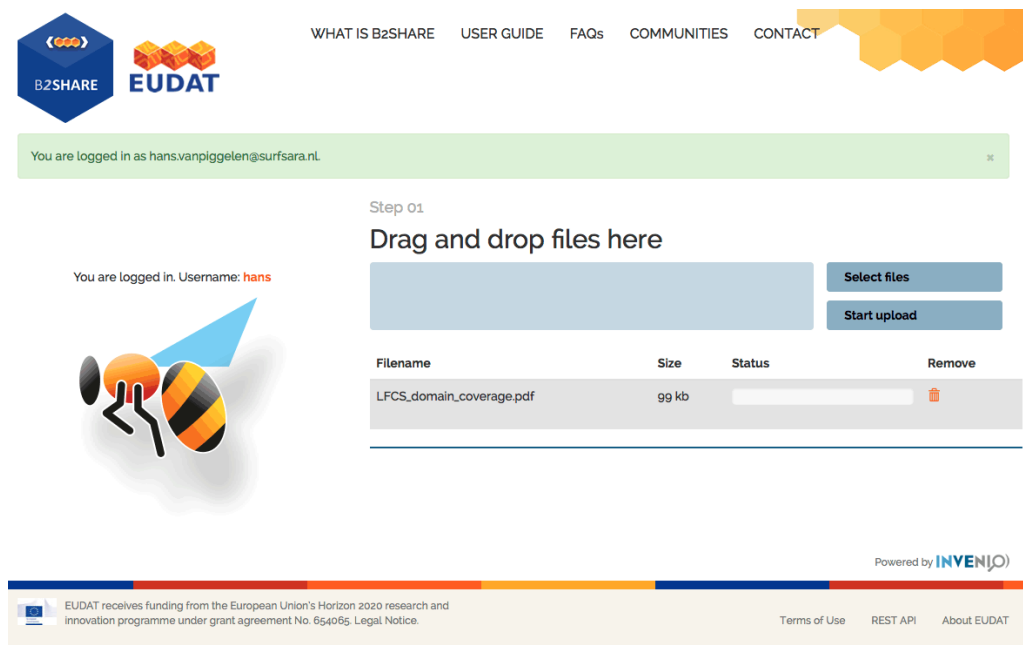


Figure 16: Uploading data in B2SHARE

B2SHARE is developed in Python using the Flask framework and CERN's Invenio software as a back-end platform. The B2SHARE development team works in close collaboration with CERN's Invenio team.

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

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

To understand how the service operates, we will have a closer look at the data model. In B2SHARE all data is handled at collection level – this means that, even if the user uploads a single data file, it will be put into a collection (which is called a “deposit” in B2SHARE terminology). This general model influences the assignment of PIDs and metadata. It is worth mentioning that, in the current version of B2SHARE, users only create metadata at collection level. Moreover, PIDs are also assigned on collection level, in other words, the service assigns one PID for a set of data and its associated metadata.

4.2.1. Development Progress during the First Year

B2SHARE development during the first year of the current phase of the EUDAT project focused on data management facilities and on creating a new architectural design for the B2SHARE software platform. Integration with the other services has also been prioritized.

In the first half of the year, work on designing the new architecture for the B2SHARE software platform and developing a new user interface module progressed. The team gathered for a two-day planning event at CERN from the 8th to the 9th of July, in order to further elaborate the design of the architecture and also to initiate a closer collaboration with CERN's Invenio team, which is providing the primary base platform for B2SHARE. Development of the new architecture has continued during the year in parallel with maintaining and further developing the production version.

Many of the new features that have been implemented during the year primarily target the area of data-management, like embargo functionality, role-based access control and editable metadata. Adaptations to support the RDA community have also been a large part of the team's activities. The B2SHARE REST API has been further enriched and team members have taken part in the work on the common EUDAT CDI REST API.

Integration with the B2ACCESS service has also been completed during the autumn and, as a result, new B2SHARE users are currently directed to B2ACCESS for registration. Existing B2SHARE users are still allowed to log in with credentials that have previously been created in B2SHARE, but this behaviour will be removed in the next major version.

A pilot integration with B2DROP has also been developed and demonstrated. This functionality will be released with the release of B2SHARE version 2.

4.2.2. Outlook

The release of B2SHARE 2.0 is planned for May 2016. Version 2.0 of B2SHARE will have the new modular architecture and improved internal functionality based on Invenio 3. Its release will also include the integration with B2DROP, support for Digital Object Identifiers (DOIs), using Puppet³² and Docker³³ for deployment, and support for metadata management by research community data managers. A pilot for using B2SHARE as a CDI Metadata Store for data in other services is also planned for the near future.

The slightly longer term perspective also includes support for authorisation management, versioning and cloud storage services, as well as support for other storage back-ends like cloud storage solutions and object stores. Integration with B2SAFE, B2NOTE and the Data Type Registry is also in the plans. The integration with B2SAFE poses a particular challenge since both services use different data models and handle metadata and PIDs at different levels.

4.3. Personal Cloud Storage – B2DROP

EUDAT's B2DROP provides researchers with a common service for synchronizing and exchanging active research data within a small group of researchers and with fine-grained access control mechanisms.

B2DROP is a user-friendly and trustworthy storage environment which allows users to synchronize their active data across different devices and to easily share this data with peers.

The service is intended for the long-tail and still mutable data objects which can change and are still subject to active research, such as drafts of papers. Therefore B2DROP offers versioning of all ingested files but does not attach persistent identifiers to them. Since B2DROP works on mutable data, no PIDs or metadata are attached to the data files explicitly; the service is considered to be part of the workspace data domain of EUDAT.

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

The B2DROP architecture and software is based on ownCloud³⁴.

³² [https://en.wikipedia.org/wiki/Puppet_\(software\)](https://en.wikipedia.org/wiki/Puppet_(software))

³³ [https://en.wikipedia.org/wiki/Docker_\(software\)](https://en.wikipedia.org/wiki/Docker_(software))

³⁴ <https://owncloud.org/>

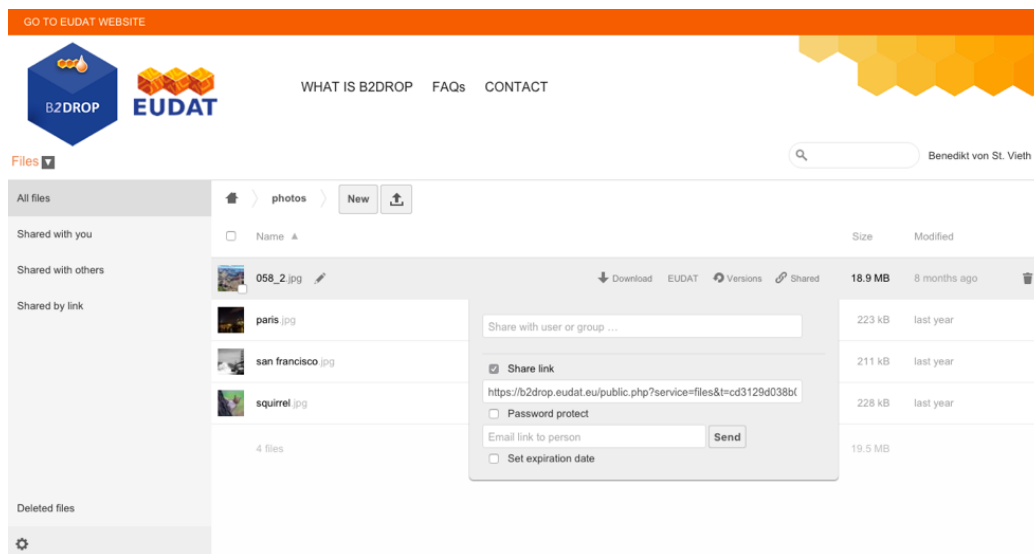


Figure 17: Sharing a file in B2DROP via a link

4.3.1. Development Progress during the First Year of EUDAT2020

During the first year of this phase of the EUDAT project, the B2DROP team has mainly worked on the B2DROP web interface, the automated deployment of the service and integrating B2DROP with B2SHARE and B2ACCESS.

With new versions of ownCloud being released, modifications of the B2DROP web interface have been required to provide the common EUDAT look and feel. OwnCloud is considered to be a stable piece of software. However, due to the intensive ongoing development of ownCloud in recent times (which was reflected in many release versions of the software) there were many changes and consequently it was often necessary to adopt EUDAT’s web interface for B2DROP to those changes in ownCloud and to further develop the underlying software for the web interface.

A second task was the automated deployment of B2DROP. In the past, experience with other services has shown that manual, error-prone, and time-consuming installations of software hinder EUDAT-wide deployment of services. In general there is clearly the need to make it easy to deploy services and update services to the latest version promptly when a new release is published. That is why the B2DROP developers have put significant effort into developing a configuration management (puppet) module for B2DROP and a container image (Docker). Both solutions are used for production and development systems. The dDocker image has been used to deploy a B2DROP instance at CSC in Finland.

Aside from work on user interface adoption and deployment descriptions, the B2DROP team developed the integration for B2DROP with the B2SHARE service. A first version of the integration with B2SHARE, the so-called “B2SHARE bridge”, was implemented as an ownCloud module for ownCloud version 8. This bridge allows a B2DROP user to select a file for publishing. After this selection, B2DROP will do a third party transfer to the publishing service B2SHARE. Since this upload is happening in the background, a view was implemented that shows the published files, their current upload state, and the URL they received in the publishing service. The upload is done via HTTP, and the B2SHARE bridge was implemented in such a way that the publishing backend could be changed to other HTTP services, for example, an implementation for OpenStack Swift is available.

GO TO EUDAT WEBSITE

B2DROP EUDAT

WHAT IS B2DROP FAQs CONTACT

B2SHARE bvonstvieth

Queued Jobs (3)

Transfer ID	Filename	Request Date
1	/DSC_6390.JPG	Tue, 12 Jan 2016 07:39:17
2	/DSC_8909.jpg	Tue, 12 Jan 2016 07:39:20
3	/DSC_8909.jpg	Tue, 12 Jan 2016 07:39:38

Publishing History (3)

Transfer ID	Filename	Status	Publish URL	Publish Date	Last Update
3	/DSC_8909.jpg	new	URL	Tue, 12 Jan 2016 07:39:38	Tue, 12 Jan 2016 07:39:38
2	/DSC_8909.jpg	new	URL	Tue, 12 Jan 2016 07:39:20	Tue, 12 Jan 2016 07:39:20
1	/DSC_6390.JPG	new	URL	Tue, 12 Jan 2016 07:39:17	Tue, 12 Jan 2016 07:39:17

Figure 18: Publishing queue view of B2DROP

A version of the B2SHARE bridge was deployed on development servers. The final version will be made available to the production system when B2SHARE 2.0 is released.

4.3.2. Outlook

The integration of B2DROP with B2ACCESS using Shibboleth is currently under development, as is support for Puppet and Docker for deployment purposes. The integration of B2DROP with B2SHARE is finished but waiting to be brought in production with the release of B2SHARE 2.0.

4.4. Data Discovery – B2FIND

The EUDAT metadata service B2FIND provides a comprehensive joint metadata catalogue and a powerful discovery portal. Metadata is stored through EUDAT services, such as B2SHARE, and harvested—from community-owned repositories encompassing a wide scope of research disciplines.

The B2FIND portal and API provide users with advanced search functionalities and make it possible to access the data resources that are associated with the metadata that can be found in the EUDAT metadata catalogue.

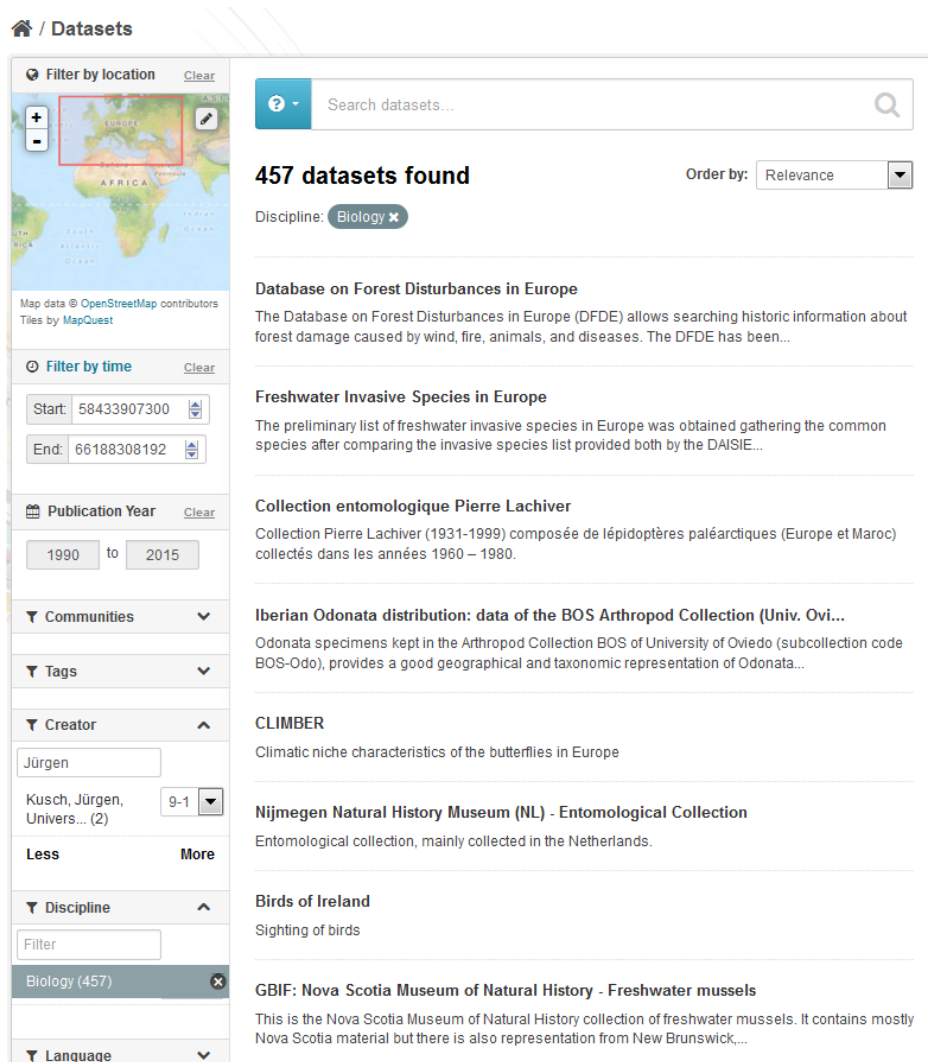


Figure 19: B2FIND result page (right panel) of a combined faceted search (left navigation bar)

Metadata that is made searchable in B2FIND is harvested from metadata providers using the standard OAI-PMH interface or repository-specific APIs. The research communities themselves decide which metadata is made available to the EUDAT services. EUDAT, in cooperation with the data and metadata experts from the research communities, has established a mapping of community metadata elements to B2FIND – this needs to be finally agreed on by the communities. A sophisticated framework ensures that metadata providers are harvested regularly to display complete and up to date information. To make it easier for users to browse and find interesting datasets the framework also provides the translation from the community metadata schema to standard facets in the B2FIND metadata catalogue.

B2FIND is based on the CKAN³⁵ data management software, which is a widely used open-source piece of software from the Open Knowledge Foundation that is used for publishing, sharing and finding data.

4.4.1. Development Progress during the First Year of EUDAT2020

Much of the development during the first year of this phase of the EUDAT project was focused on the requirements for the continued uptake of B2FIND by the research communities – in fact, this was the main area of objectives for B2FIND during the first part of the year.

One major area of emphasis has been the enhancement of the B2FIND ingestion software stack. The following are some significant aspects of this:

- a) the redesign from the old Java implementation of the XPath (a syntax used to define rules on XML files, such as which fields to extract) pre-selection functionalities into a new and more flexible Python module, and

³⁵ <http://ckan.org/>

- b) the extension of the B2FIND stack by a basic validation and verification mode which makes it possible to generate , for example, simple statistics about facet coverage.

Work on the improvement of semantic mappings was continued – this included further stabilization of the B2FIND schema, the generalization of mapping methods and adapting to the requirements of new and old research communities.

Another central topic was the improvement of the GUI. Selected facets now use an auto-complete functionality and different sorting techniques. The facet language is now based on the ISO639-3 and CLDR libraries. Preliminary ideas on how GUI templates can be introduced into the GUI have been generated.

Much effort went into enhancing and evolving different methods for accessing B2FIND. One example of this involves scripts utilizing the B2FIND-API. One of the new ideas is to make it possible to access B2FIND using the SRU (Search/Retrieval via URL) protocol – some preliminary concepts for this have been developed.

It is crucial to document the specific needs of the research communities with regards to metadata mapping, and this has been done accurately with relevant input from B2FIND being given to other subtasks (like documentation) or work packages, in particular WP6. This will make it easier for research communities that are new to EUDAT to make their metadata searchable and discoverable via the B2FIND service.

Development within the different branches of B2FIND (ingest, mapping, GUI, and API) continued. Functionality which would make it possible for metadata providers to initiate additional harvesting cycles is under consideration.

Other achievements of the B2FIND team include improvements to the user interface, to the faceted search facilities and to B2FIND's search filtering and sorting. Several bugs were also fixed. The process of integrating B2FIND with the research community infrastructures is ongoing: metadata from more research communities was integrated and ingested; some communities are in the process of integrating their metadata with EUDAT and several other research communities have shown interest in integrating their metadata with EUDAT. With the uptake of further communities and the resultant increase in the sheer amount and diversity of the ingested metadata, the consolidation of the data stock and the consistency checking have become increasingly important tasks that requires additional effort. New documentation about "B2FIND usage" and "B2FIND integration" were also produced.

4.4.2. Outlook

Work on improvements in handling the granularity of metadata and data were commenced. An SRU protocol interface and the integration of semantic annotations is currently being prepared. The integration of research community metadata into EUDAT will continue, and so will the work on improving the experience, performance and scalability of B2FIND. In terms of a slightly longer perspective, there are plans for a prototype of the SRU interface, extended harvesting methods, improved search functionalities and improved semantic mapping.

4.5. Registry Services

The so-called "Registry services" constitute a new development task in this phase of EUDAT and the first year's efforts have therefore mostly been focused on planning and gathering requirements. EUDAT is planning to provide a generic and public Data Type Registry. Parsing and interpreting large amounts of data automatically requires that data, or parts of the data, are strongly formatted and adhere to standards. To properly define and communicate these formats and standards, a public Data Type Registry can be queried, which then provides an application with the appropriate definitions and thus facilitates the automatic analysis of data. EUDAT is planning to use the Data Type Registry to define the B2HANDLE PID records and the B2SHARE metadata templates and related data types.

The CNRI³⁶ software package Cordra³⁷ was evaluated for this purpose in WP8 and the Registries task team commenced building the Data Type Registry service.

³⁶ <https://www.cnri.reston.va.us/>

³⁷ <https://www.cordra.org/>

4.5.1. Development Progress during the First Year of EUDAT2020

During the sixth month of this phase of the EUDAT project, WP8 delivered an assessment showing that Cordra could serve as a perfect base technology for implementing a Data Type Registry (DTR) as a core service of EUDAT. The Registry Services part of the WP5 team made use of those WP8 results, and started to deploy a testing instance of a DTR to be used both by EUDAT core services and by research communities that have shown interest to this service early on.

4.5.2. Deployment of CORDRA as a Testing Instance of the DTR in EUDAT's CDI

This Virtual Machine is administratively managed by the Operations Department as part of BSC's infrastructure, which ensures that backups of its contents are performed periodically. Additionally, in order to guarantee that the contents of the DTR service Database are consistent, the database itself is replicated and compressed via rsync to another in-house machine. This should suffice to safeguard all the data that is provided by the research communities testing the server until it is officially deployed into the EUDAT CDI.

The following tasks were completed in order to set up the testing instance of the Cordra software for this Virtual Machine.

- c) *Cordra Installation*: The version of Cordra that was chosen for this installation was v1.0.5, which was the most recent version available at the time of writing this document. That version also includes the full source and development files so that it can be adapted to fit EUDAT's objectives, if needed.
- d) *Basic Configuration*: Given that the expiration of handles from CNRI after 60 days could have become a problem for the testing instance, the installation that was deployed was configured to use the ePIC service already hosted in production by BSC, which uses the handle prefix "11101".
- e) *Configuration as a Data Type Registry*: Cordra can be used by default to hold different types of Digital Objects, which means that it needs to be configured to generate and manage the Data Types that are needed by this phase of the EUDAT project. In order to do that, a new JSON schema that describes these Data Types was added through Cordra's administration Web Interface.

4.5.3. Outlook

The development plans for the EUDAT Data Type Registry services during the next period include continued work on the service pilot, and collaboration with pilot communities on further adaptations, followed by branding the service as one of the B2-services, integrating B2SHARE and B2HANDLE data types into the DTR, and integrating the DTR with B2ACCESS.

4.6. Federated AAI – B2ACCESS

The main goal of EUDAT's B2ACCESS service is to provide a unified authentication and authorization service for the EUDAT CDI while making sure that it is as simple as possible for Identity Providers (IDPs) and Service Providers (SPs) to be integrated into the EUDAT CDI services. The B2ACCESS service provides a bridge between external IDPs³⁸ and the SPs within the EUDAT CDI³⁹, as shown in Figure 20

³⁸ This includes organisational IDPs, social IDPs and the EUDAT IDP.

³⁹ Including but not limited to the services shown in the diagram.

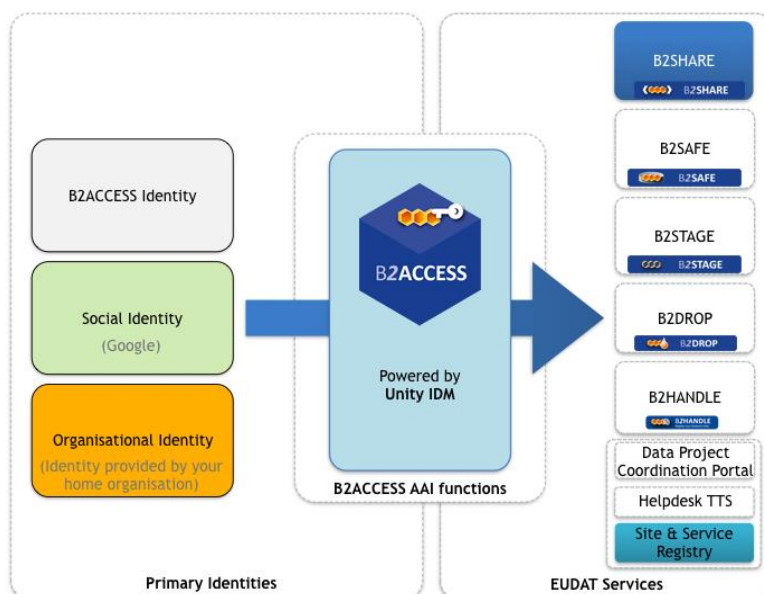


Figure 20: Architectural overview of the B2ACCESS service

Users can use their social identity or organisational identity to log in to EUDAT services through B2ACCESS. Users without access to any of these external IDPs can create an account and use that account to access the EUDAT services. The external user identity, together with the externally provided user attributes, is mapped onto an EUDAT identity and augmented with required information that is not provided by the IDP but which is necessary in the EUDAT domain. This includes a level of assurance (loa) attribute indicating the level of trust in the method of authentication that is used. Social IDPs, for example, are assigned a low(er) level of trust than institutional IDPs.

The user's EUDAT identity, together with the information about that user which is needed by EUDAT, is made available to the backend service providers, that is, the B2 services.

Based on the level of assurance for the method of authentication that was used and on the user's attributes, B2ACCESS can create short-lived certificates (proxies) for some of the services, such as B2SAFE, on behalf of the user. This bridge is needed since several protocols which the EUDAT services depend on to transfer data cannot handle user (SAML) attributes, hence it is necessary to translate user identities and attributes to proxies.

The B2ACCESS service also provides group management services, managed on a service level, in order to reduce the administrative overhead for the main B2ACCESS administrator. This means that a trusted administrator from a particular research community can assign attributes and privileges to members of his or her own community. Group membership is encoded as an attribute and should be used by backend service providers to make authorization decisions. The B2ACCESS service does not store authorization rules directly, but instead provides input for the parameters in such rules by means of attributes. The full list of attributes is included in Annex E: B2ACCESS Attributes.

The first release of B2ACCESS was handed over and deployed in production⁴⁰ in October 2015.

4.6.1. Architecture

The core component of the B2ACCESS service is Unity IDM⁴¹, an open source cloud identity and federation management service, developed by ICM⁴², PL-GRID⁴³ and Unicore⁴⁴. An EUDAT Certification Authority (CA), based on the implementation from the Contrail project, is integrated into the B2ACCESS service in order to issue short-lived certificates. The OAuth authorization server, which is capable of issuing security tokens, is part of the Unity IDM (identity management) service stack.

⁴⁰ <https://b2access.eudat.eu/>

⁴¹ <http://unity-idm.eu/>

⁴² University of Warsaw, <http://www.icm.edu.pl/web/guest>

⁴³ <http://www.plgrid.pl/>

⁴⁴ <http://www.unicore.eu/>

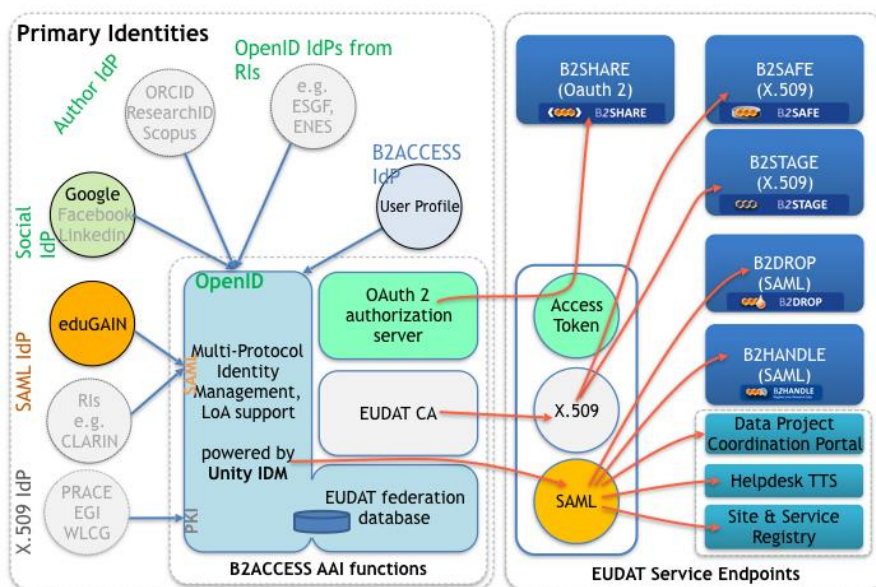


Figure 21: Technical overview of the B2ACCESS service

4.6.2. Identity Provider Integration

B2ACCESS supports multiple technologies to integrate research community IdPs: (1) SAML⁴⁵, (2) OAuth2⁴⁶ and OpenID Connect⁴⁷, (3) X.509⁴⁸ certificate, (4) ldap, and (5) username/password. Currently, eduGAIN⁴⁹ is in the course of being integrated into EUDAT using the SAML approach. Google accounts are supported via OAuth/OpenID Connect. It should be noted that users can register an account directly in B2ACCESS if they do not have access to an eduGAIN IdP or do not own a Google account, thus so-called “homeless” users can be supported. We are currently working on enabling more social IdPs such as, but not limited to, Facebook and GitHub. Certificate-based authentication will also be enabled soon, as this will provide a very high level of assurance, which is needed for some backend service providers. The integration of new IdPs is tested in the B2ACCESS staging test environment, and then configured in production by the B2ACCESS system administrator.

4.6.3. Service Provider Integration

B2ACCESS supports multiple technologies to integrate backend services: (1) SAML, (2) OAuth2 and (3) X.509. Currently, B2SHARE is integrated using the OAuth2 approach and the data project coordination portal is integrated using the SAML approach. Integration with the B2DROP web-interface (using SAML) and B2SAFE (via X.509 certificates) is underway. An in-depth technical description of all three integration approaches will be available online soon. Integration with B2ACCESS allows users to obtain access to all federated backend services. More specifically, upon successful authentication through B2ACCESS, the user is redirected to the backend service. In the case of SAML, all user attributes are encoded in the SAML response; in the X.509 scenario, the attributes are encoded in the certificate and in the OAuth2 scenario, an extra call can be made to the B2ACCESS service to retrieve the user attributes. The backend service can then use the attribute values to make an authorization decision. See Annex E for an overview of all attributes released by the B2ACCESS service.

⁴⁵ SAML is the Security Assertion Markup Language, an OASIS standard. The profile used here is [SAML Web SSO](#).

⁴⁶ OAuth2 is a protocol originally designed to delegate rights to services on the web. The protocol has been extended to provide authentication. OAuth2 is an IETF standard defined in [RFC6749](#).

⁴⁷ OpenID Connect (OIDC) is an authentication layer on top of OAuth2. The core OIDC functionality is defined in the [OpenID Connect Core 1.0](#) specification.

⁴⁸ In cryptography, X.509 is an [ITU-T](#) standard for a public key infrastructure (PKI) and Privilege Management Infrastructure (PMI). X.509 specifies, amongst other things, standard formats for public key certificates, certificate revocation lists, attribute certificates, and a certification path validation algorithm.

⁴⁹ eduGAIN is an international interederation service interconnecting research and education identity federations: <http://services.geant.net/eduGAIN/>

4.6.4. Outlook

For the coming project year, we will focus on integrating more EUDAT services. Integration work has already started for B2DROP and B2SAFE, and these will be followed by B2STAGE and the HTTP API. Another major task, which has already been started and will continue during the coming year, is designing and implementing a solution for distributed authorization. We are currently working on a XACML⁵⁰-based solution. Identity provider integration with PRACE and EGI will begin in the next period, and installation packages for easy deployment and distributed set-up will be provided. Other ongoing tasks are evaluating the end user workflow and making improvements where possible, as well as responding to bug reports and feature requests from WP4 and WP6.

⁵⁰ The eXtensible Access Control Markup Language (XACML) is an OASIS open standard that defines a declarative access control policy language implemented in XML and a processing model describing how to evaluate access requests according to the rules defined in policies.

5. DATA PRESERVATION

The Data Preservation service area is concerned with providing tools for long-term preservation of data and for policy-driven data management.

This service area is grouped in three subtasks: 1) Data Management Policies, which include the topics of data replication across the CDI/EUDAT network (B2SAFE) and defining policies (Data policy manager) which the data is subject to, 2) Persistent Identifiers for tracking replicas of data in the EUDAT domain and 3) Data Curation and Provenance which play an important role in keeping data usable in the longer term. The work performed within the first two subtasks is described in the following paragraphs. Work on the third subtask has not yet started. A short overview about the future activities relating to it is given in section 5.3.

The Data Preservation service area sets out from the already existing B2SAFE service that was developed in the first phase of EUDAT. This service makes it possible to replicate data objects across multiple, geographically separated locations and focuses on bit-stream preservation, which is the basis of data preservation.

To date in the second phase of the EUDAT project, the service has been enhanced to make it easy to adapt to different usage scenarios and to provide better maintenance by improving the architectural modularity. Moreover, the plan is to integrate B2SAFE with other EUDAT services to enable the data to be moved between these services according to the data life cycle. In accordance with the requests from EUDAT's research communities, the B2SAFE service will be extended with the capability to manage metadata.

During the past year, the activities in these tasks were focused on giving more control to the users, thereby realising the community-driven principle of the EUDAT project. Therefore, effort was put into developing new features and interfaces to allow users to define data management policies and to improve the readability of the metadata.

Two major updates of the underlying systems, iRODS and the Handle system, influenced the development activities of this service area substantially.

5.1. Data Management Policies

A Data Management Policy in the context of the EUDAT B2 services is a way to formalize and structure a set of abstract processes and “concrete” pieces of information into a single description that is represented using a specific language. The objective of using data management policies is to simplify the management of such processes on the one hand, and, on the other hand, it makes it possible to re-use similar policies in similar contexts, derive new policies from existing ones, and define a complex hierarchy of policies which would be impossible to deal with otherwise.

The main objective of the Data Management Policies subtask is to provide services and tools to support long-term preservation of data. Apart from long-term preservation, EUDAT's research communities requested EUDAT to implement services for data distribution based on policies, not only for preservation purposes, but also to make it easier to move data to computational resources and to improve access to data.

In contrast to B2STAGE – a service which also moves data between servers, but in an ad-hoc way and on demand – here data will be replicated automatically according to certain patterns (for example, on a regular basis) and according to policies.

The Data Management Policies subtask involves the B2SAFE service and the Data Policy manager. While the B2SAFE service takes care of the execution of policies (such as for PID registration, data replication, and integrity checks), the Data Policy manager makes it possible to define and distribute policies across the EUDAT CDI network.

During the past year, the existing architecture has been consolidated and improved, taking the input from the research communities into account. The following issues were tackled.

- *Compatibility with iRODS software*⁵¹: iRODS is the underlying technology of the B2SAFE service. During the past year iRODS itself was subject to heavy code refactoring, moving from version 3 to

⁵¹ <http://irods.org>

version 4 at the beginning of 2015. The reasons for these changes were to include security bug-fixes, to add new features and to simplify the installation process. However, these changes had a huge impact on the replication mechanism implemented by B2SAFE. B2SAFE itself is a set of iRODS rules which implement data policies. However, these rules were not compatible with the new iRODS version 4. Hence, the B2SAFE rules also had to be partially rewritten.

- *Performance improvements for scalability:* The B2SAFE service was designed to replicate massive amounts of data and associate each object (file) with a persistent identifier (PID). These identifiers can be resolved to the objects themselves. Additionally, they provide the means to link data and their replicas and are thus help to track data and replicas across the EUDAT domain. PIDs are stored in a dedicated registry and kept as long as required by policies. However, the creation of a PID for each object or replica forces the B2SAFE service to interact with another, often remote, service (B2HANDLE) and, by this, decreases the performance in terms of running time. In some cases where whole collections of data were being replicated, the amount of time that was required to register PIDs was too long. To improve this performance hiccup, the replication process was reorganized as described in Section 5.1.1.
- *Usability:* To encourage and simplify the usage of B2SAFE and the DPM within EUDAT, as well as by research communities, work was done to make the installation, maintenance and deployment easier.
- *Community metadata:* While the initial version (2.3) of B2SAFE was already able to manage the metadata created by the EUDAT services, the service did not provide the means to identify and manage metadata from the research communities. Work has been started on developing features which enable B2SAFE to handle such metadata.

The following sections give further details of the current developments.

5.1.1. B2SAFE

The B2SAFE service, at the core, exploits the iRODS rule engine in order to perform a set of actions to implement specific behaviour defined in data management policies. The actions are defined by a set of iRODS rules which can either be executed on regular basis or be triggered by actions like data ingest. The rules interact with external software components which deliver functionalities such as PID registration. Several Python scripts facilitate the interaction. The whole B2SAFE module, including the Python scripts, can be conceptualized as modules as shown in the architectural overview in Figure 22.

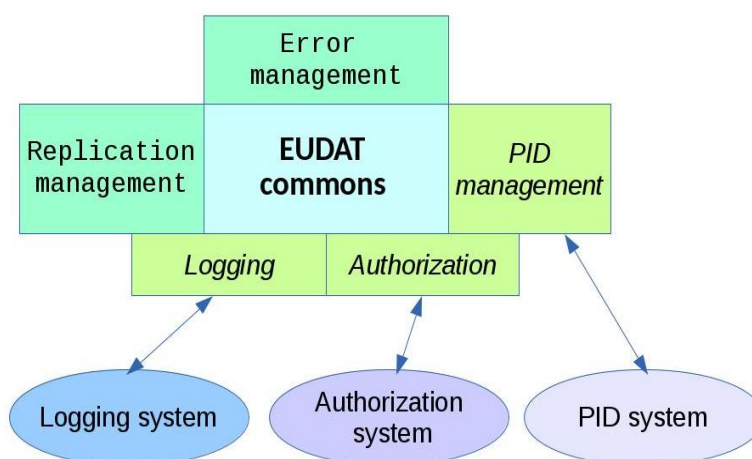


Figure 22: B2SAFE modules architectural overview

The evolution of the B2SAFE service has been driven by the aforementioned requirements and scheduled according to a six-month development cycle which is reflected by the releases⁵².

- V. 3.0.0 beta was released in July 2015 – the beta release was a preview for EUDAT partners to evaluate the last changes to the service and to check the compatibility with its own environment.

⁵² <https://github.com/EUDAT-B2SAFE/B2SAFE-core/releases>

- V. 3.0.0 stable was released in October 2015 when it was handed over to the Operations team, ready for production environments.
- V. 3.1.0 beta was released in January 2016.
- V. 3.1.0 stable is expected to be released by April 2016.

More precisely, the developments concerning *Compatibility* had impact on the following modules.

- Replication: B2SAFE functions relying on obsolete iRODS 3 functions were updated with the corresponding new functions in iRODS v.4.x. The logic of some low level mechanisms was updated to benefit from the new iRODS rule language features.
- PID management: The planned adoption of the new PID system, the Handle System v.8.x⁵³ (see section 5.2) made it necessary to also update the client employed by B2SAFE to create and update PIDs. Although the new system has not been put into production yet, tests to verify the interoperability between it and B2SAFE have already been executed.

The improvements on *Performance* involved the same two modules.

- Replication: The object replication and the object registration (in a PID registry) were two synchronous operations. To improve the performance of the replication process for large data collections, these two actions were separated.
- PID management: Some calls to the client to communicate with the PID system were merged in order to optimize the interaction with it.

To improve the *Usability* of B2SAFE, the main usage patterns were selected, consolidated and documented as follows.

1. Object upload and synchronous PID registration

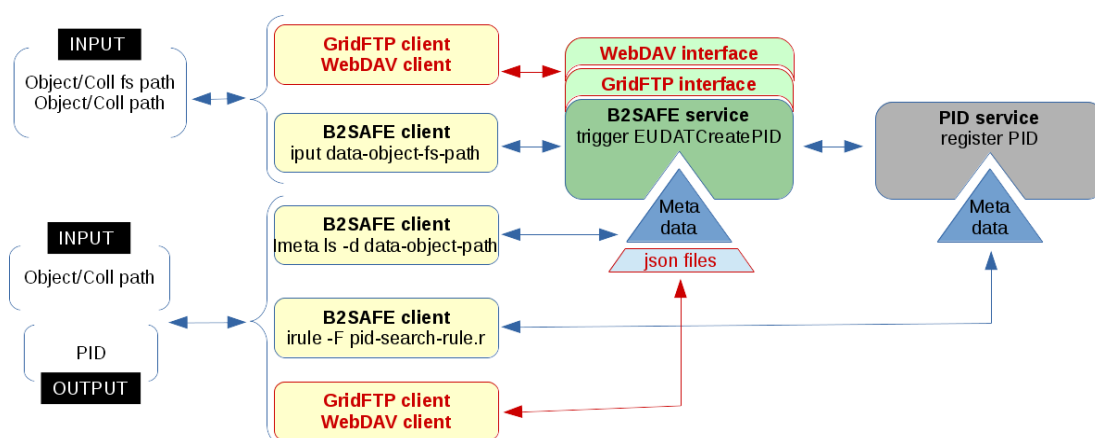


Figure 23: B2SAFE workflow: object upload and PID registration

Figure 23 shows a set of clients (the first two yellow rounded rectangular shapes from the top), which upload data to the B2SAFE service. The service is configured to register the object (that is, to create a new PID for it) as soon as it is stored (in other words, it is done in a synchronous way). The arrows indicate the flow in terms of uploaded data, metadata and messages. All three clients are retrieving the metadata (system metadata) associated with the uploaded data by querying the iRODS central (iCAT) database (top), the PID registry (green box in the middle), and the iRODS service where the metadata is stored as json files (bottom).

2. Object/collection replication and synchronous PID registration

⁵³ https://www.handle.net/download_hnr.html

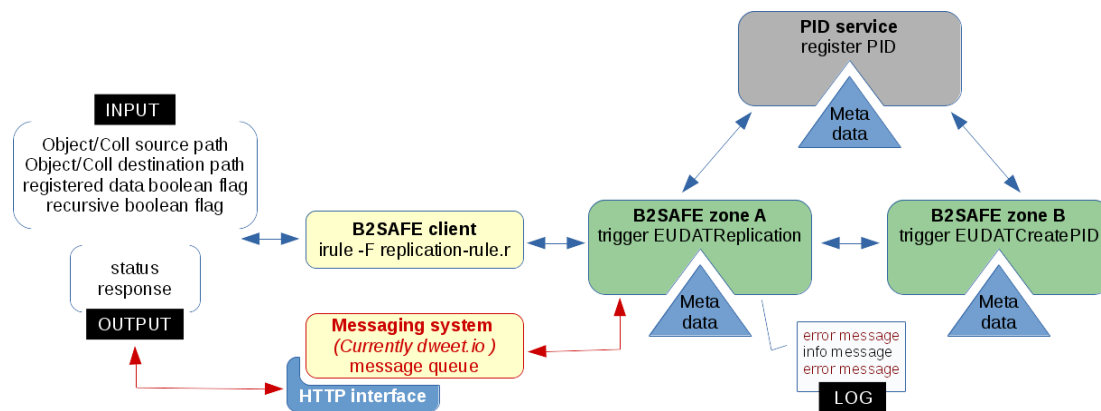


Figure 24: B2SAFE workflow: object replication and PID registration

Figure 24 represents the triggering of a replication operation through a client. It shows two B2SAFE service instances placed in two different (iRODS) zones, where zone A is the source and zone B is the target of the replication. The source data set is already registered with PIDs, while the replica data set will receive PIDs as soon as the transfer is completed (that is, in a synchronous way). Moreover, the status of the operation is reported in two different ways: by logging to a file, local to the B2SAFE instance and accessible only to the system administrators, and by sending messages to the queue of a publicly accessible messaging system.

3. Object/Collection replication and asynchronous PID registration
The workflow is identical to the one describing the synchronous PID registration with one exception. The PID registration of the replica can be started any time after the completion of the data transfer.
4. Recovery of failed replications
5. Data integrity check
6. Update of the location pointed to by the PID

Only the details of the first two patterns are provided – this is to highlight the roles of the new components (rounded squares outlined in red). In fact the B2SAFE v. 3.0.0 stable release includes two feature previews which are designed to fill the following gaps in its usability.

- Metadata is provided in different formats: Usually metadata is stored in the iRODS database (iCAT). Additionally we also provide metadata as json files, so that clients based on protocols like GridFTP and webDAV are able to retrieve it, and thus get access to, for example, the PID for data stored in B2SAFE.
- Messaging system: In the previous version of B2SAFE, communication about the status of data replication took place via the system administrators. To automate communication and make it independent from human interaction, the results of the replication are provided via a messaging system that also makes it possible to retrieve the information in an asynchronous way.

Finally, the work to implement the *community metadata management* has been planned and the architectural design has been documented.

To consolidate the B2SAFE service, several components and scripts were tested and updated according to the latest specifications. The scripts to synchronize the user accounts with the B2ACCESS service were updated to reflect the latest changes in the B2ACCESS service. In a similar way the repository package component⁵⁴, which includes the connector to interface with the archiving system DSPACE, was tested and updated in the latest version of B2SAFE. Moreover, a new monitoring probe⁵⁵ compatible with iRODS 4.x was made available to the WP6 monitoring team.

5.1.2. Data Policy Manager

The Data Policy Manager (DPM) was developed as prototype during the first phase of the EUDAT project. Therefore, the main objective of the work during the first year of the second phase of the project was to develop it into a stable and production-ready tool. The DPM consists of two components: 1) a central service

⁵⁴ <https://github.com/EUDAT-B2SAFE/B2SAFE-repository-package>

⁵⁵ <https://github.com/EUDAT-B2SAFE/B2SAFE-core/tree/master/monitoring/irods>

with a web interface and 2) distributed agents. Figure 25 shows how the two components are related to each other. The web interface allows users to describe high-level data management policies, abstracting from the specific tool and programming language that are used to enforce the policy. The agents then translate the policies into sets of iRODS rules which in turn can be executed by the local rule engine. During this development period the following happened.

- The DPM central service was re-factored in order to simplify the deployment and configuration processed.
- The DPM distributed agent code was developed further in order to manage time-based policy-scheduling. This means that a policy can be scheduled locally at each CDI node relying on a system scheduler (the Unix cron tool⁵⁶) which is not part of the B2SAFE service, but which is assumed to be available in the local environment where the service is running.
- User identities at the EUDAT-level can now be mapped to the local user identity in B2SAFE, thus allowing the execution of policies to impersonate the policy author. This is done to ensure that the policies are executed with the same privileges as those of the policy author.

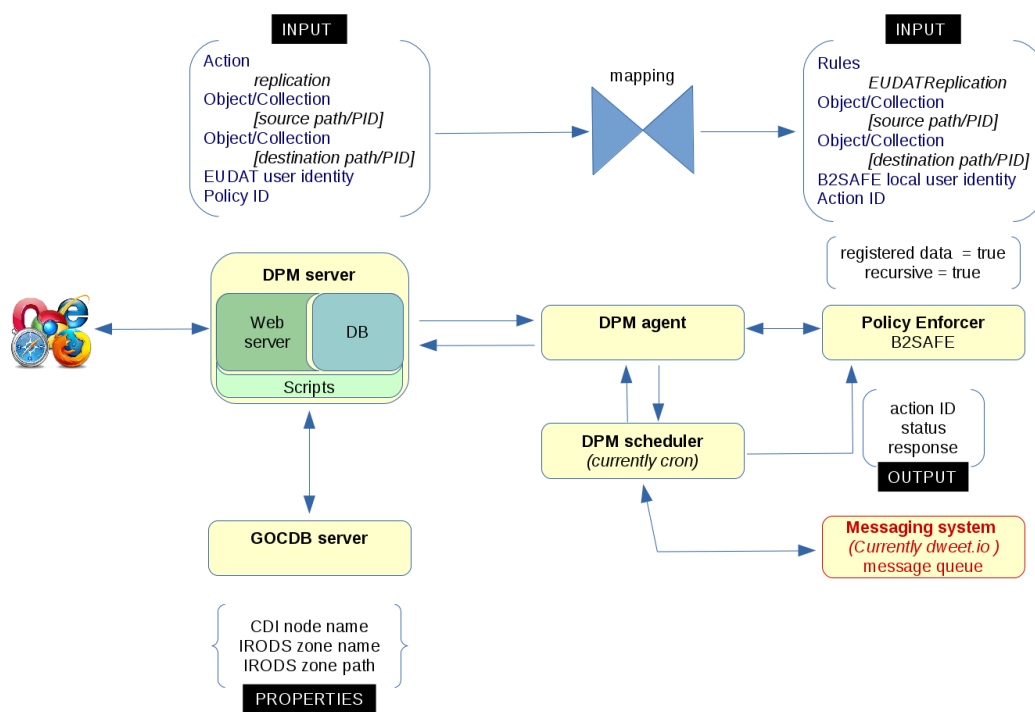


Figure 25: Data Policy Manager data flow

Figure 25 shows the data flow of the data exchanged between the DPM and the other components and services and gives some details about the contents of the data that is exchanged.

- Within the DPM server, an authorized research community data manager defines a policy by filling in a policy template. To do this, the community data manager must log in to the DPM service. The community data manager will be logged in on his/her profile page where he/she can define new policies or update existing ones. Policies are defined in a technology-independent abstract way.
- A research community data manager can only define policies on resources granted to that research community or data project. For this, the DPM server fetches resource property information from the service and site registry service, that is, the GOCDB.
- To distribute the policies across the B2SAFE instances, a DPM agent runs at each CDI node running a B2SAFE instance which, on a regular basis, pulls policy information from the DPM server. Because the policies managed in the DPM server are defined in an abstract way, the DPM translates the policies into technology specific policies and loads the policies into the B2SAFE service.

⁵⁶ <https://en.wikipedia.org/wiki/Cron>

- When the policies are loaded into the B2SAFE service, B2SAFE will execute the policies according to the defined event trigger (for example, run once or periodic)
- The B2SAFE service returns audit information from the execution of the policies to the DPM server. In this way, a community data manager can monitor the execution of the policies which run in the distributed CDI from a central place.

In November 2015, a beta version 1.0⁵⁷ of the Data Policy Manager was released.

- The DPM central server was connected to the GOCDB⁵⁸. The information about specific EUDAT CDI nodes was retrieved from the GOCDB and made available to the DPM user in a transparent way.
- The Federated-Identity approach was implemented through the integration between the web interface and the B2ACCESS service via the Shibboleth protocol.

5.1.3. Documentation

In parallel with the software development, the DPM task supported the WP5 Documentation team in writing the relevant documentation, which was published on the EUDAT project web site⁵⁹. The documentation that has been published so far is related to the software that has already been handed over to the operation team ready for production, that is, the B2SAFE service. Since the Data Policy Manager is still in a beta version, producing the documentation for this part has been postponed till the final release.

5.1.4. Tests

Another activity, which was carried out in parallel with the software development and in order to support it, is the creation and maintenance of a test bed that is shared among all the developers. The primary aim with the test bed was to make it possible to perform functional testing and, secondly, to be able to run some preliminary non-functional testing. Annex D summarizes the main properties of the test bed. The test bed consisted of three B2SAFE instances via which the different policies supported in the DPM could be tested. Figure 26 shows how the B2SAFE instances are connected to each other.

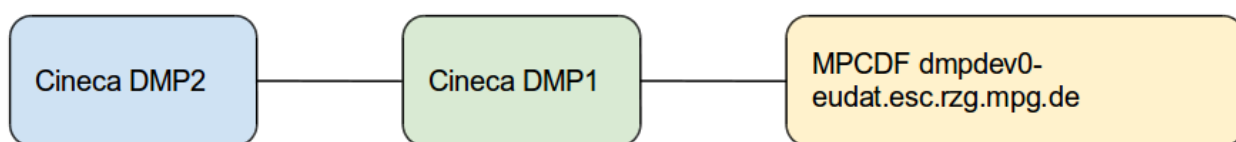


Figure 26: DMP test bed machines

Annex D gives a complete overview of the configuration of the B2SAFE instances of the DPM test bed environment.

5.1.5. Outlook

During the next year of this phase of EUDAT, both B2SAFE and the Data Policy Manager will be developed further and improved, which will include integrating the two tools better. Improving the management of community metadata in B2SAFE is also prioritized and the implementation of the proposed solution is planned to be released for production before summer 2017. Furthermore, the aim is to provide WP6 with a solid and user-friendly version of the DPM.

There are also lower priority plans to integrate the DPM with the data project coordination portal developed by WP6 in order to provide a first level of policy validation against the real storage resources available in the CDI. Moreover, we will continue the integration of B2SAFE with other services, like B2SHARE. In addition, a more advanced authorization mechanism should be implemented according to the guidelines provided by the B2ACCESS team.

⁵⁷ <https://github.com/EUDAT-B2SAFE/B2SAFE-DPM/releases>

⁵⁸ <https://registry-eudat.esc.rzg.mpg.de>

⁵⁹ <http://eudat.eu/services/userdoc/configure-b2safe>

5.2. Persistent Identifiers (PIDs)

The PIDs subtask deals with software related to persistent identifiers and the development of policies relating to PIDs. The PIDs subtask is closely related to the developments in ePIC⁶⁰ (the e-Science Persistent Identifier Consortium) and the Handle System.

EPIC is a consortium of European institutions that aim to provide PID services to their respective clients. GWDG⁶¹ is a major member of ePIC and acts as a prefix administrator (MPA) for the DONA Foundation⁶², which oversees the administration of the global Handle System. This means that ePIC members have signed contracts with GWDG and paid annual fees to gain the right to issue Handle namespaces (prefixes) which depend on the policies of the individual institutions. EUDAT relies on these processes to secure Handle prefixes, while maintaining its own infrastructure for operating Handle servers and mirroring PIDs. In the past, EUDAT has also relied on ePIC to provide a software component that makes it easier to access Handle services, however, this is undergoing a change and EUDAT is moving towards a software suite that is fully maintained by EUDAT as part of the B2HANDLE service. All of these activities aim to address risks that were uncovered towards the end of the first phase of EUDAT and during the Helsinki kick-off for the second phase of the project.

The work on the PID subtask started at the kick-off meeting in Helsinki in March 2015 with a dedicated side event that aimed to clarify the goals of this subtask for the whole of this second phase of the EUDAT project, and also to gather more information on the current status of the PID services from the first phase of the project and provide a forum for interaction between EUDAT and ePIC. EPIC was represented at this meeting as well and the future plans for ePIC development were presented by GWDG. Critical points were raised by EUDAT concerning the relationship between EUDAT and ePIC, particularly regarding the sustainability and quality of the ePIC software (which the EUDAT services were highly dependent on at that point). An external factor driving these concerns was the announcement of the release of Handle System v8 by CNRI, as features of that system would replace a significant part of what the ePIC software provided, therefore making it at least partially obsolete. Another concern was the sustainability of the organizational form of ePIC, which was set up as a consortium of partners but lacked the framework of being a dedicated legal entity – this was perceived as a weakness given that EUDAT expressed a need to have dedicated EUDAT PID policies to foster trust in EUDAT services and address long-term concerns from the user communities. The concerns raised and discussed during the meeting strengthened EUDAT's resolve to initiate activities that address these sustainability issues in both the mid- and long-term perspectives.

In consequence, the members of the PIDs subtask assembled a more detailed software development plan and initiated the development of the necessary software, taking into account the concrete points raised at the kick-off meeting. A major goal of the PIDs subtask activities throughout the first year of this phase of EUDAT was to decrease the dependency of EUDAT services on external software provided by ePIC through dedicated development, which resulted in the B2Handle library. EUDAT continues to rely on ePIC for access to Handle System prefixes and collaboration with DONA and other partners, but has reduced the technical and operational dependencies on ePIC.

5.2.1. B2Handle Development

The goals with the development of the B2Handle library are 1) to decrease EUDAT's dependency on the ePIC API and 2) to unify various scripts (EPIC clients) that were developed by different tasks and research communities during the first phase of EUDAT.

The main development activity during the first year of this phase of the project was to design and develop a Python library (called the "B2Handle library") that unifies the various existing scripts, which were developed during the first phase of EUDAT, into one coherent manageable software artefact that is designed according to best practices and development standards. The library provides a Python-level interface to Handle Server services. Python was chosen as the implementation platform because all the scripts interacting with Handle services existing at the beginning of this phase of the project were written in Python.

⁶⁰ <http://www.pidconsortium.eu/>

⁶¹ <http://www.gwdg.de/>

⁶² <http://www.dona.net/>

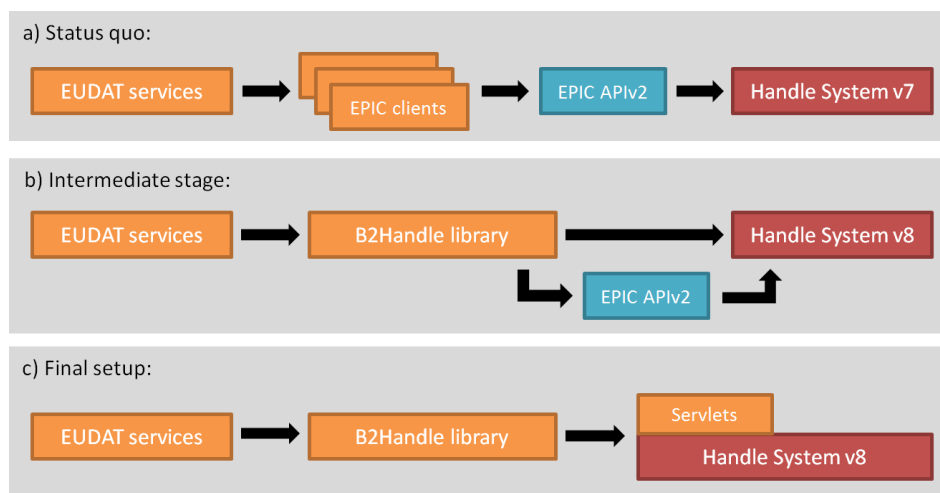


Figure 27: The iterative evolution of the B2HANDLE service

Figure 27 illustrates the iterative evolution of the EUDAT PID services to their future final state. All the EUDAT components are coloured in orange. From top to bottom, the development stages are as follows.

- Top row: In the current setup, the EUDAT services rely fully on the EPIC APIv2 for all interaction with the Handle servers. Multiple client scripts exist within EUDAT that connect to the EPIC API.
- Middle row: In an intermediate stage, the client scripts within EUDAT are replaced with the B2Handle library and the Handle System is upgraded to version 8, which the B2Handle library can connect to directly for most tasks. The EPIC APIv2 is kept operational to perform search/reverse lookup queries.
- Bottom row: In the final setup, the remaining functionality that EUDAT requires from EPIC APIv2 is replaced with servlet components deployed within the Handle System v8 embedded servlet container, thereby fully eliminating the dependency on EPIC software.

The PIDs subtask gathered feedback on intermediate and final versions of the library, particularly from B2SAFE (as that service is the most important user of the library) but also from other EUDAT services (including B2SHARE and B2STAGE), to make sure that all the EUDAT activities working with PIDs can use the library and that all their requirements are adequately addressed. Preliminary and final versions of the library were distributed to the interested EUDAT services and test server accounts were also provided to ease the planned roll-out in the second year of this phase of the project.

To cover concerns raised at the kick-off meeting over the long-term sustainability and quality of the PID service software components, a professional Continuous Integration (CI) system was configured by GRNET⁶³ for automated building, testing and reporting. The CI workflow is fully integrated with the public GitHub repository hosting the B2Handle library, including on demand pull request testing and automated publishing of documentation.

In collaboration with the Monitoring Activity in WP6, B2HANDLE was chosen as the first service that would implement its monitoring probes for the new monitoring service. The B2HANDLE functional tests will include: CREATE, READ, UPDATE and DELETE (CRUD) of PIDs through the EPIC API components and the resolution and mirroring capabilities of the Handle components.

During this period, a new monitoring probe was developed by GRNET that tests the CRUD capabilities using the ePIC client library and an existing probe developed by SURFsara⁶⁴ was reused in order to test the resolution capability. In the second year of this phase of EUDAT, the B2HANDLE Library will replace the old ePIC client and the individual tests within the probe will be refactored and enhanced.

5.2.2. Coordination Activities

In view of standardization efforts at the Research Data Alliance (RDA), the PIDs subtask initiated an action to document and unify the contents of PID records within EUDAT and to improve interoperability with possible external services as described by the approved outputs of the RDA Working Groups on PID Information

⁶³ <https://grnet.gr/>

⁶⁴ <https://www.surf.nl/en/about-surf/subsidiaries/surfsara/>

Types⁶⁵ and Data Type Registries⁶⁶. The unification of the contents of PID records is facilitated by the use of the new library which makes it easier to manage specific PID record values. This action consisted of a series of video calls concentrating on assembling a normative document for EUDAT PID records. The subtask aimed to gather feedback from all concerned services, particularly B2SAFE and the upcoming HTTP API development. The document was handed over for approval by the EUDAT Technical Committee, and the plan is to roll out changes to the EUDAT services during the second year of this phase of the project.

A discussion with longer-term impact concerns the possible integration of Unified Resource Name (URN) support with the B2Handle library or other EUDAT services. The URN specifications are currently undergoing a major revision by IETF⁶⁷. However, the specifications have not yet passed the final steps of the IETF approval workflow, and therefore, the subtask will keep monitoring these efforts throughout the second year of this phase of EUDAT.

5.2.3. Outlook

In December 2015, the Handle System v8.1.0 server software was finally released by CNRI. DKRZ and SURFsara have already run tests on the version that was released and have deployed it on test servers. The final release now enables EUDAT to roll out the new software stack during the second year of this phase of the project, preferably in a coordinated action in spring 2016 together with support from WP6. The roll-out is expected to be a major effort and will require careful planning and further testing, and may also include additional development due to the significant changes to the existing software stack.

In general, the development of the B2Handle library will continue during the second year of this phase of EUDAT with a focus on improving the quality of the package as part of an overall maintenance effort. New requirements and features requested by B2 services will be discussed and implemented as required. Some of the concrete next steps that will be taken include publishing the library through the Python package index (PyPI) and extending the technical and end user documentation.

The discussions regarding the document about the contents of normative PID records will continue as well. The next step after finalizing the document will be to apply the changes in practice, that is, to adapt the existing B2 service components dealing with PID records to the norm – possibly as part of a general roll-out – and to discuss the initiative and coordinate it with similar efforts on a larger scale such as those of the RDA.

Additional work may be done as part of the interaction with the EUDAT research communities. Some communities may express direct needs for PID services, and eventually also call for dedicated tools to help with operative PID management. Discussions about such support tools and standardized processes have been going on in the past within EUDAT, and also under the umbrella of the RDA, and this subtask will contribute to these discussions and may pick up specific ideas for future tool development if time allows.

5.3. Data Curation and Provenance

The objective of the Data Curation and Provenance subtask is to improve the quality of the EUDAT CDI by defining and automating clear policies for data curation and provenance. These policies should convince consumers (and other stakeholders) that the data and related services in the CDI meet the requirements that relate to their concerns. This task supports the extended focus of EUDAT from preserving bit streams to curating digital objects and their contents.

The work on this task is expected to start in April 2016. The following sections provide an overview of the expected tasks and outcomes.

5.3.1. Expected Outcomes

The subtask will provide EUDAT and its members/users with:

- an overview and analysis of policies,
- demonstrators that show that/how policies can be executed using the CDI, and
- awareness of the role of data curation policies in the EUDAT CDI and their impact on the CDI.

⁶⁵ <https://rd-alliance.org/groups/pid-information-types-wg.html>

⁶⁶ <https://rd-alliance.org/groups/data-type-registries-wg.html>

⁶⁷ <http://ietf.org/>

The progress of these outcomes will be reported in the periodic deliverables D5.1, D5.2 and D5.3.

5.3.2. Challenges and Opportunities

The challenges this task faces include the following.

- Data curation is a relatively new topic in the domain of EUDAT. It brings in other practices, language and culture.
- Policies cover a domain from high-level and strategic to low-level and technical aspects. This task is based in a relatively technical domain. The challenge is to involve enough strategic stakeholders.
- The maturity of policies over different domains and organisations differs. Making a clear policy for a distributed and diverse and international community will be challenging.

The opportunities open to this task include:

- efforts in data policy managers and rule based policy execution using iRods,
- efforts in certification (Task 2.2), and
- building expertise amongst the partners.

5.3.3. Strategy

The overview of and analysis of policies will require the following.

- A framework to organise and analyse policies is needed to relate to stakeholders, connect high-level with (machine-) executable policies, relevant services, and so forth. We will continue the work done in the SCAPE project⁶⁸. It will be important to have a clear understanding of the scope and ambitions of EUDAT, the CDI and its services.
- An inventory of existing policies in EUDAT (explicit and implicit), of normative policies (as implied by, for example, the Data Seal of Approval) and of policies that are desired by external (client-) communities is also needed. For this, we will continue on from the work done by the RDA Practical Policies⁶⁹ group, by WP4 on community engagement and by EUDAT's Task 2.2 on Trust and Certification.

The demonstrators will be created using the following steps.

- The requirements for demonstrators will be defined and appropriate policies will be selected. The important requirements are (a) that the demonstrators need to help selected stakeholders to understand the policies, and how these can be automated, and (b) it should be feasible to implement the policies using the existing software.
- Selected high-level policies will be translated into executable policies and implemented in B2SAFE and/or B2SHARE. To do this, the Data Curation and Provenance task will collaborate with the relevant EUDAT services.
- The policies will be deployed and tested. The policies should be registered in the EUDAT Data Policy Manager.

Awareness will be increased by:

- providing clear definition and explanation of the term 'policy',
- engaging with EUDAT partners, tasks and workgroups during the inventory of policies, and reporting on and presenting outcomes of that, and
- engaging with client communities by developing a training module with WP3.

It is expected that the additional focus on data curation and policies will require a cultural change, which should be handled accordingly.

⁶⁸ http://www.scape-project.eu/wp-content/uploads/2014/02/SCAPE_D13.2_KB_V1.0.pdf

⁶⁹ <http://dx.doi.org/10.15497/83E1B3F9-7E17-484A-A466-B3E5775121CC>

6. DATA PROCESSING AND ANALYSIS

The Data Processing and Analysis task continues the work from the first phase of the EUDAT project by providing enhancements to the B2STAGE service, through which data is ingested into the EUDAT infrastructure and data movement to and from HPC services is facilitated. In addition, the work in this area is providing a Python library, which will also benefit the other EUDAT services. The design of the RESTful HTTP API is nearing conclusion, which will provide a general API for storage and retrieval to be used by the other EUDAT B2 services. Of the new work areas that have been introduced, the semantic web task has been making progress throughout this first year of this phase of the project. Work on other new tasks relating to workspaces and big data analysis is due to commence in the next project year.

6.1. Data Staging (B2STAGE)

B2STAGE is a service designed to facilitate transfers of large data sets between EUDAT storage resources and high performance computing (HPC) workspaces in a reliable, efficient, and lightweight manner. The service supports the following functionalities:

- transferring large data collections from EUDAT storage facilities to external HPC/HTC facilities for further processing,
- ingesting the results from computations into the EUDAT infrastructure, and
- in conjunction with the B2SAFE, replicating research community data sets and ingesting them onto EUDAT storage resources for long-term preservation.

The B2STAGE service comprises four different packages:

- the GridFTP iRODS Data Storage Interface (iRODS-DSI) to enable fast data transfer through the GridFTP protocol,
- the HTTP RESTful API service to enable the up-/download and management of digital assets via an HTTP (The HTTP API is defined in collaboration with the HTTP Storage and Federation task (9.1.2) and will be reported in the D9.2 deliverable.),
- the Data Staging Script to instrument data transfers via a command-line tool, and
- the EUDAT python library to provide users with programmatic access to EUDAT services.

6.1.1. GridFTP iRODS-DSI

GridFTP is a high-performance, secure, reliable data transfer protocol which provides remote access to data stores. There are many different types of data storage systems – from standard file systems to arrays of magnetic tape. To make it possible for GridFTP to support differing underlying storage systems, the GridFTP server can be extended by implementing an interface called the Data Storage Interface (DSI) without modification to the core server code.

The GridFTP iRODS-DSI was developed making use of the iRODS C API to facilitate the interaction between the GridFTP server and the iRODS server as the underlying storage system. Once this has been built and configured, the iRODS-DSI will give user the ability to manage data on an iRODS server through any GridFTP client by passing a valid iRODS path to it. Figure 28 illustrates the workflow: every time a transfer request is sent to the GridFTP server, the iRODS-DSI takes care of managing the operation (upload, download, list, delete) and interacting with the iRODS server.

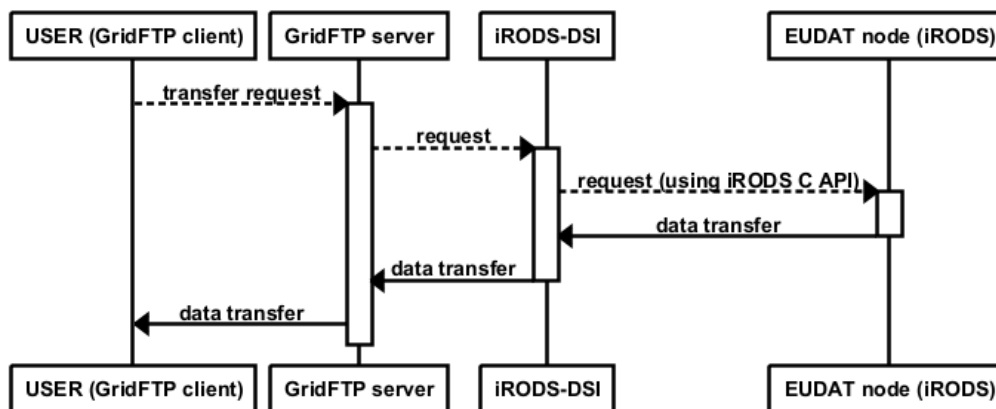


Figure 28: Data transfer with iRODS-DSI

During the first year of this phase of the project, the functionalities offered by iRODS-DSI module were extended by adding the ability to use PIDs for listing and retrieving resources: this gives makes it possible for users to list and download data by passing either a PID or an iRODS path as input.

When a PID is given as input, the iRODS-DSI tries to resolve it and perform the requested operation using the URI returned by the Handle server (see Figure 29). The operation (listing or downloading) will be correctly performed only if the URI returned by the Handle server is a valid iRODS path pointing to the iRODS instance which the iRODS-DSI is connected to.

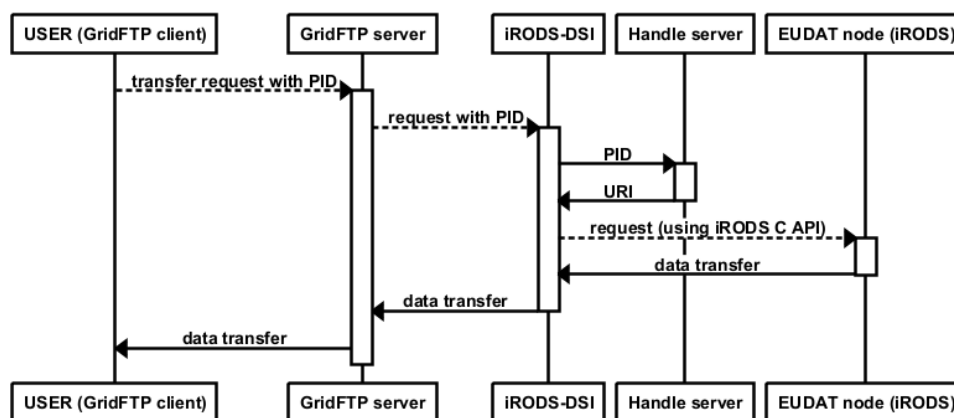


Figure 29: Data transfer with iRODS-DSI passing a PID as input

6.1.2. Data Staging Script

The Data Staging Script (DSS) is a Python script providing users with the capability to stage data using the associated PID and interacting with the Globus OnLine⁷⁰ (GO) service to perform the actual data transfer. This script can be used to transfer data between an EUDAT data node (equipped with GridFTP iRODS-DSI) and an external GridFTP server and is able to handle data that is qualified through its iRODS physical path or via PIDs. As mentioned previously, the script makes use of the GO service (via a Python API): the Data Staging Script instruments the transfer request and, subsequently, GO manages the operations without any user intervention. The general steps of the GridFTP dataflow are explained in section 3.4.3. Over the past year, the Data Staging Script has been updated to be compatible with the new iRODS rules for managing PIDs.

6.1.3. EUDAT Python Library

The goal of the EUDAT Python library is to provide users with programmatic access to EUDAT services. It aims to support interactions with all the EUDAT services ranging from the ingestion of new data sets and their discovery and retrieval, to their analysis. The EUDAT library will furnish Python APIs through which users will be able to create their own clients to carry out specific tasks, creating workflows which involve different EUDAT services.

⁷⁰ <https://www.globus.org/tags/globus-online>

Therefore, the EUDAT library should become the primary entry point to access EUDAT services via a programmatic python-based interface. The library will serve to couple the interfaces of the different EUDAT services thus facilitating their interactions and their integration into the systems of the existing research communities. In this way, the library will become the "mediator" across all the interfaces provided by other EUDAT services covering the full data life cycle.

In the first year of this phase of the project, the development of the library focused on the B2STAGE and B2FIND services. Through the library it is now possible to manage Globus GridFTP endpoints, to start third party data transfers⁷¹ of digital entities, objects and collections, and to check the transfer status. To perform these kinds of operations, the EUDAT library makes use of the Globus Online Python APIs. Moreover, the library makes it possible to find datasets by using research community names or tags – this takes advantage of the REST APIs exposed by CKAN (the technology which B2FIND is based on). A test suite was written to verify the correct operation of the running code (using non-regression tests).

Throughout the duration of the EUDAT project, the Python library will be extended to interact with other EUDAT services: in particular, in the next months, integrating the library with B2SHARE and B2ACCESS will be investigated.

6.2. Semantic Web Services

The aim of the Semantic Web Services task is to design and implement a set of prototypes for Semantic Web/Linked Data services. These services will make it possible to extend the metadata descriptions of files in EUDAT, and offer a flexible data classification system using domain-specific keywords and a unique platform for aggregating data from multiple scales and domains. Several building blocks that will support this process are being considered, for example, an Ontology Look Up service⁷², a term lexicon to store new terms that are missing in the current ontologies (similar to Neurolex⁷³) and a service to publish ontologies that are stored in B2SHARE.

The first semantic web service that has been addressed is a data annotator, which enables users to attach additional metadata to files and link the files together in a format that can be used by machines. The previous phase of EUDAT⁷⁴ identified a need for such a service and developed a first proof of concept based on a user-case provided by LTER⁷⁵. This service was called B2NOTE (<http://b2note.bsc.es>) and it provided an interface and services that made it possible for users to map textual information from different text-containing file types (text, csv⁷⁶ and pdf) with defined domain-specific vocabularies. This preliminary work was used to define the initial system requirements for the annotation service, which needed to have the ability to:

1. create RDF triples using external ontologies,
2. store the annotations (RDF triples),
3. publish and query annotations,
4. be integrated with other EUDAT services,
5. be integrated with community architecture, and
6. have a scalable infrastructure.

To refine these system requirements, user scenarios relying on B2SHARE, B2FIND and B2DROP were created. This made it possible to investigate the potential integration of the Semantic Web services with the other EUDAT services, and to identify four main uses of the Semantic Web services: manually linking files with semantic tags and other files, semi-automatically annotating the contents of files hosted within the EUDAT CDI, annotating external documents (a legacy use case), and creating, using and publishing new aggregated datasets.

⁷¹ GridFTP allows for the execution of the client process to be at a separate location from both the source and destination locations. A *third-party transfer* lets a researcher at a remote site initiate a data transfer from Site A to Site B, where he has access to data and/or computing facilities.

⁷² <http://www.ebi.ac.uk/ontology-lookup/>

⁷³ http://neurolex.org/wiki/Main_Page

⁷⁴ <http://www.eudat.eu/semantics-2nd-eudat-conference>

⁷⁵ <http://www.lter-europe.net/>

⁷⁶ https://en.wikipedia.org/wiki/Comma-separated_values

The analysis was based on the view that users want to create either private annotations (to organise relevant data within the EUDAT CDI) or public annotations (for crowdsourcing). These user scenarios led to an initial set of user requirements for the annotation service, namely that it should have:

1. an easy-to-use user interface with
 - a. auto-completion (manual annotation),
 - b. an interface for annotation curation and management,
 - c. automation of annotations using learning agents,
 - d. an intuitive search facility based on annotations, and
 - e. a means for visualizing the contents of text-based files for annotation,
2. the ability to create links with files within CDI and/or outside of the CDI,
3. a method for tracking the annotation provenance that is associated with a user account, and
4. a facility for retrieving annotated data and annotations.

(Note that, as the development of these services follows an agile-like principle, this list of requirements will be re-evaluated in the light of user input on the first prototypes of this service.) These requirements were then used to create development specifications.

Substantial development effort then went into investigating existing platforms (such as EBI⁷⁷, Europeana⁷⁸, and Bioportal⁷⁹), testing different technological “bricks” for building this service, and designing the user interface. For the back end technologies and architectures for storing, retrieving and publishing annotations, attention was focused on RDF triple stores, as well as alternative storage technologies like MongoDB⁸⁰ and Neo4J⁸¹. The first version of the development specifications used OpenVirtuoso⁸² for CRUD operations and the integrated SPARQL⁸³ endpoint for searching but several major limitations were identified. Nevertheless, as this platform is a “standard” for publishing RDF triples through a SPARQL endpoint, EUDAT will use it to publish stable annotation datasets – which means that the triples need to be stored on another back end. Although SQL back ends are extremely efficient, NoSQL⁸⁴ and MongoDB were chosen as NoSQL offers much greater flexibility (in particular for tracking the provenance of annotations), and the open source version of MongoDB offers advantages including (a) being compatible with Django (the Python-based web framework used for the development of the initial B2NOTE web interface) and (b) using JSON which is the optimal format for using the existing JSON serialization of the RDF data model (JSON-LD).

For the web front end, the functionalities of the Django web framework and the integration with MongoDB were investigated. Both the Python and JavaScript⁸⁵ libraries are being used for auto-completion and the well-established SolR⁸⁶ software was chosen as the indexing engine. The challenging aspects with this were to find ways to efficiently harvest and index terms from various ontologies and to also optimize the response time for auto-complete.

In terms of the existing data models and formats for creating annotations, the JSON-LD⁸⁷ serialization of RDF as well as the Open Annotation model⁸⁸ and the W3C PROV model⁸⁹ were investigated – the aim was to use those standards for storing the annotations. Also, to fulfil an additional requirement from one of the EUDAT Data Pilot case studies, there was work done on a possible serialization of csv that could be used to annotate csv files.

During the course of investigating these various aspects, a first prototype of the architecture was built at the BSC facilities. Currently an OpenVirtuoso instance and a SolR instance are running on a dedicated Virtual

⁷⁷ <http://www.ebi.ac.uk/>

⁷⁸ <http://www.europeana.eu/portal/>

⁷⁹ <http://bioportal.bioontology.org/>

⁸⁰ <https://www.mongodb.org/>

⁸¹ <http://neo4j.com/>

⁸² <http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/>

⁸³ <https://www.w3.org/TR/sparql11-query/>

⁸⁴ <https://en.wikipedia.org/wiki/NoSQL>

⁸⁵ <https://en.wikipedia.org/wiki/JavaScript>

⁸⁶ <http://lucene.apache.org/solr/>

⁸⁷ <http://json-ld.org/>

⁸⁸ <http://www.openannotation.org/spec/core/>

⁸⁹ <https://www.w3.org/TR/prov-overview/>

Machine (VM). This VM is rather small, however, there are plans to test the required characteristics for a full system at a later date. Additionally, the Django-MongoDB framework is installed on a second VM, where the proof of concept is hosted.

After the preparatory investigative phase, development of the initial version of the B2NOTE service began. That version focused on a simplified scope where the user could manually create semantic tags using the auto-completion function (user requirement 1a). The user interface is rather simple and enables users to create annotations that are automatically stored in the MongoDB back end (system requirement 2) and retrieved from the DB to be visualized and managed. The users can actually delete annotations from the list of annotations (user requirement 1b). To store annotations and provenance information (user requirement 3), this first prototype may use a simple data model where each term is stored as an Internationalized Resource Identifier (IRI), with a label (which is used in the autocomplete function) and simple provenance information (such as the user, creation data and modification data). The next version of the data model will be extended to incorporate the Web Annotation Data model, which is integrated with the PROV model. The current version of the service is available here: <http://b2note.bsc.es/devel/>.

As the main aim of this service is to make it possible to annotate data elements within the CDI, we have also worked on integrating it with B2SHARE. In order to integrate B2NOTE seamlessly with other EUDAT services (system requirement 4), it was proposed that B2NOTE should be a widget that could be called by the UI of the different EUDAT services. After discussions with the developers of the different services, the technical solution that was selected is the use of web iframes. This approach was tested and integrated into the first prototype. The result is accessible here: <https://86.50.168.182/record/8>. As this approach is rather simple and generic, it could be used with any research community web interface (system requirement 5).

This version of B2NOTE is a preliminary proof of concept, which can be used as a demonstration service to be tested and validated by the users. The first version was delayed by several months due to a recruitment issue that has since been resolved. Further development of the service is under consideration, including extending the functionalities (such as publication in triple store, searching for annotated data and annotations, sharing/publishing annotations, and adding lexicon functionalities) and the number of indexed ontologies, developing a RESTful API (to access the annotation services), an Ontology lookup service and annotation spaces for individual users, benchmarking backend storage technologies, assessing smart or semi-automated annotation via automatic learning ,and integrating it with the other B2 services.

6.3. Outlook

In the Data Processing and Analysis service, the developments in the coming project year are concentrated on the further development of the EUDAT library, extending the support for GridFTP and data transfers via Globus Online and File Transfer Service⁹⁰ (FTSv3), implementing searches via B2FIND and the integration with the B2ACCESS service. Also the pre-production and the first production release of the HTTP API service is planned, along with supporting up- and downloads of the digital assets to/from the B2SAFE service. On the developments in the semantic web service task, a pilot release of the B2NOTE service is planned. The research communities that will test the B2NOTE service will be able to use the pilot instance. Also work on integrating the B2NOTE service with the other B2 services will start. The tasks about Workflows and Workspaces, and on Big Data Analysis tools are planned to start. The Workflows and Workspaces task will concentrate on the uptake of the Generic Execution Framework which is being developed in task 8.4.5 in the Data Life Cycle across Communities JRA work package. In the second year of this phase also the task to assess the usage of the B2 services (for example, B2SAFE and B2DROP) as workspace areas will start up. The Big Data Analysis tools task will assess how data stored within the CDI can easily be transferred to big data analysis systems, for example, a Hadoop⁹¹ or array database system run in a cloud computing environment. At the time of writing, both of these tasks are in the process of defining their work plans.

⁹⁰ <https://svnweb.cern.ch/trac/fts3/wiki>

⁹¹ <http://hadoop.apache.org/>

7. USER DOCUMENTATION AND TRAINING MATERIALS

User documentation is a task that continues from the first phase of the EUDAT project. Its scope covers the following:

- collecting relevant information and documents provided by other work packages and tasks,
- consolidate the form, style and content of the documentation in such a way that there is a consistent set of documents that are suitable for the targeted audience,
- monitoring the user documentation to ensure that it is up to date and still targeting the right audiences, and
- supplying augmented content for training material related to EUDAT's services and their functionality (which is new in the second phase of EUDAT).

The User Documentation activities were relocated from Work Package 6 to Work Package 5 in the second phase of EUDAT, in order to bring them closer to the developers. The aim of this action was to co-locate the team that co-ordinates and edits user documentation, with the pool of developers who provide the raw material. Additionally, in this phase of EUDAT, the User Documentation team is in charge of producing training material for the core EUDAT services, in particular but not limited to software tools (like sandboxes) for training. The rationale behind this move is to leverage the user documentation material and the requisite relationship with the developers so as to be able to provide more technical material in an economical fashion. It was also envisaged that embedding these activities into Work Package 5 would provide access to more technical skills, which are required for production of the requisite software tools. An additional aim was to foster better collaboration with the Work Package 3 team, which includes the Training and Communication team. With the exception of the Task Leader at EPCC⁹², all the other team members are new to this task, including seasoned EUDAT colleagues from SURFsara and DKRZ, and fresh eyes from GRNET. The team effort is a modest 0.6 FTE (0.3 FTE at EPCC, and 0.1 FTE at each of the other three sites), which requires careful co-ordination and prioritisation of work given the wide scope of the activities that are involved. The new team was inducted on the 23rd of March 2015 and all members contributed to the slides for the 60-minute, EUDAT2020 kick-off slot that were presented on the 26th of March. Seventeen team calls have taken place between April and December 2015, co-ordinating work and tracking progress.

The past year has shown the clear benefit from relocating these activities to Work Package 5. Production of user documentation material has remained steady, with seven new documents being generated, while all 22 existing documents have been revised and restructured in response to reviews from the new team members and the Communications team. The departure from Work Package 6 has been handled well, with the team agreeing on a viable plan with the Task 6.3 Enabling Team to embed user documentation into their activities. Collaboration with the Training Team (WP3.2) has been fruitful, with good compliance with the agreed deadlines. Alignment with Work Package 4 is planned to ensue in the second year of this phase of EUDAT, motivated by the new projects from the Call for Collaboration.

In summary, the user documentation and training materials activities act as a hub between EUDAT's trainers, communicators, developers and operators. The rest of this section discusses the highlights in the user documentation and training material lines of work in respective sub-sections, including an outline of the future plans. Emphasis in the second year of this phase of the project is expected to shift to training, in line with the training team schedule.

⁹² <https://www.epcc.ed.ac.uk/>

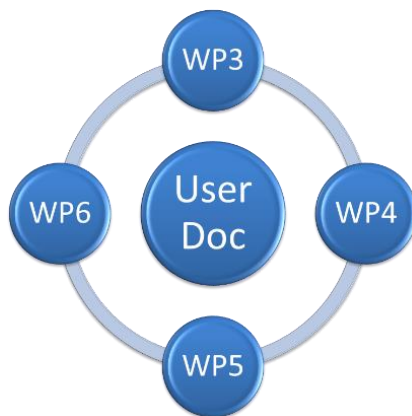


Figure 30: User documentation is a hub of EUDAT activity

7.1. User Documentation

The user documentation activity inherited a successful framework from the first phase of EUDAT and has worked on improving that. User documentation is still presented around the Engage-Deploy-Use triptych, and delivered over the EUDAT website. A technical change is that, as proposed by the Web team, some of the material is now generated directly on the Drupal Content Management System that the Web team uses, and the team was trained on the job to use this technology. The Task Leader presented the previous document-upload process to the Web team and worked with them to update the process for Drupal. Over time, other members of the team led the more technical aspects of the collaboration. Of particular service to the User Documentation team is the new Drupal facility to compare document versions. The Web team also updated the URLs of the pages to match their position in the sitemap and added breadcrumbs, also keeping the old URLs.

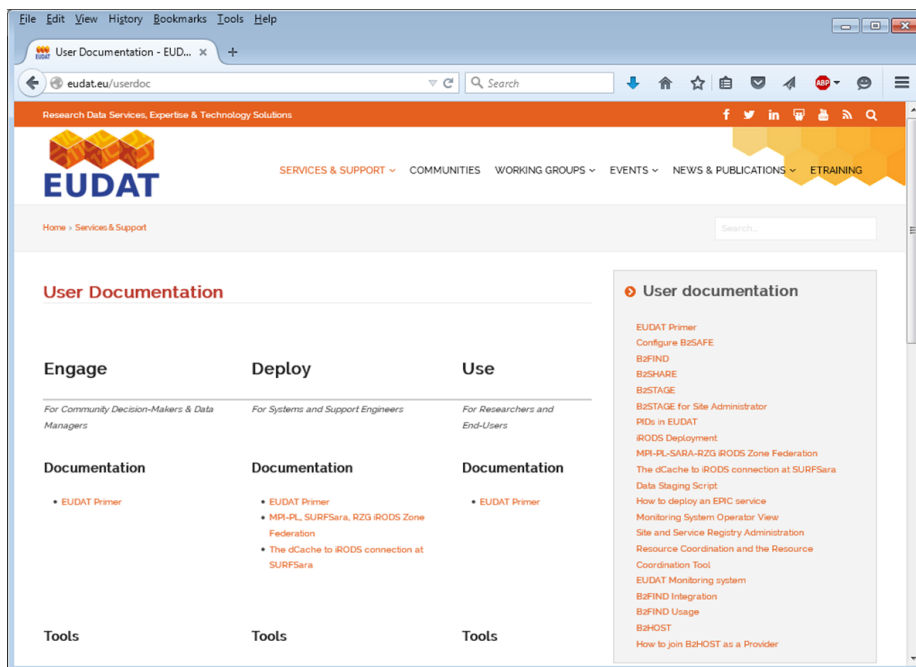


Figure 31: The User Documentation web page, implementing the Engage-Deploy-Use model.

Collaboration with the Web team to improve the Drupal tooling is ongoing, with significant progress having been made so far. In practice, many of the developers still prefer to draft material on Confluence⁹³, and this is something that the user documentation team accommodates to encourage their contribution. Other minor changes to the workflow have taken place so as to streamline the publication process. All changes were documented on the team wiki. They include lessons learnt from the team activities thus far, notably

⁹³ <https://www.atlassian.com/software/confluence>

standards for presenting figures and tables, which led to developing new CSS⁹⁴ styles to be applied across all documents.

It is also still true that, for reasons of economy, the core team cannot encompass experts from all of the diverse EUDAT services, and thus must rely on the developer team to deliver the primary material. At the start of this phase of the EUDAT project, the service building and operations groups agreed that a new service or an update to existing could only be handed over to operations or released to the end user when they were accompanied with an appropriate level of user documentation. There has been mixed success in achieving this thus far, although EUDAT has made clear progress in improving the standard of documentation available with releases.

The documentation material has been enhanced and updated, with excellent support from the developer teams. All documents have been slightly restructured to improve readability and to make them easier to navigate. Where big gains were possible with small changes, these were made as well. Figures were also updated on pages where services have been updated in accordance with the new design principles.

At the end of December 2015, the documentation portfolio included 22 documents, of which seven are new in this phase of EUDAT and many have been updated significantly. In order to prioritise work, the team contacted the Service Area Managers and assessed the state of the existing documentation and the developers' plans for releases. The new documents concern the new services – B2FIND, B2DROP and B2ACCESS – as follows.

- Three new B2FIND documents were published, piloting the new approach to develop separate documents for Community Managers, Developers and Users, and were produced working exclusively on Drupal. The new process of delivering documentation was also followed for the first time for these documents.
- The document discussing B2DROP was designed to be very light, since it is needed only for completeness as the tool is graphical, well established and should be self-explanatory. In practice, the decision to document the security and capacity specificities of the EUDAT2020 B2DROP deployment was vindicated, but the team used B2DROP for its interactions with the training team and found nuances that were hard to follow. It was thus decided to document a complete workflow, a decision that was received well by the B2DROP team, as was the documentation of the B2DROP (ownCloud) desktop clients. The work on these aspects also largely influenced the content of the B2DROP material, validating the decision to co-locate user documentation and core-services training material.
- B2ACCESS was the first service where the rule that software and draft documentation should be delivered in tandem was put into practise. The user documentation team contributed not only editorial work, but also suggestions on the product. A further two B2ACCESS documents will be published in early 2016.

Additionally, two B2HOST documents, that had been drafted during the later stages of the previous phase of EUDAT, were published, while other older documents were revised and archived without publication, as they are no longer of interest.

The EUDAT Primer was extensively reworked in light of remarks from the new members of the team. Significant amounts of text were rewritten and re-factored. The document has been repeatedly updated to include new services and reflect changes to existing ones. The team also made use of Drupal facilities to link content appropriately across the document and along the user documentation body of work, something that was not possible when Confluence was used for document preparation. The “Configure B2SAFE” document was also updated to the new B2SAFE version 3 and iRODS 4, as was the “iRODS deployment” document. It is worth highlighting that the user documentation team played an important role in the Operations discussion about iRODS versions, supporting the now adopted view that sites should be encouraged to upgrade to iRODS 4. Other updates have also taken place, for example, Jülich⁹⁵ updated their B2HOST details on the relevant document.

⁹⁴ <https://www.w3.org/Style/CSS/>

⁹⁵ http://www.fz-juelich.de/ias/jsc/EN/Home/home_node.html

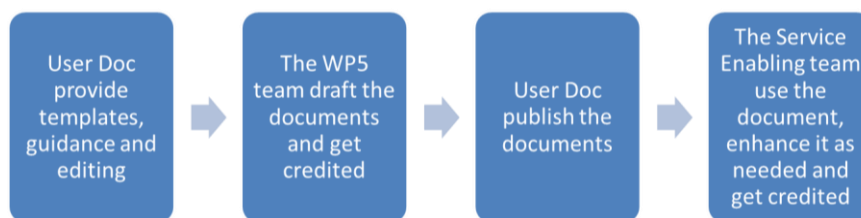


Figure 32: The user documentation workflow engages developers and operators.

Also of interest is the relation between the documentation task and Work Package 6. A user documentation talk formed the bridge between Work Package 5 and Work Package 6 in their face-to-face sessions during the meeting in Amsterdam from the 6th to the 9th of October 2015. In that talk, we secured buy-in from the Service Enabling team for a methodology to consistently use and constantly update the user documentation material. The idea is that the Service Enabling team uses the material available on the EUDAT website for the enabling projects, and that they provide feedback and ideas for its improvement. A Documentation queue was subsequently set up as a formal mechanism for communication. This model closes the loop between the developers and operators and allows feedback into the documentation from the colleagues with day-to-day, hands-on experience with the software.

7.2. Training Materials

After an initial exchange of information and goals alignment at the kick-off meeting for this phase of EUDAT, collaboration with the training team started in earnest at the face-to-face meeting on the 21st of September, where a timeline was agreed for the user documentation team to update the training material. As part of this exercise, the following presentations were updated and delivered (generally on time) to the training team in Work Package 3: “EUDAT Description”, “B2FIND Usage”, “B2FIND integration”, “B2SAFE Usage”, and “B2STAGE”. A B2SHARE presentation has also been developed and is under review by the service developers. Additionally, a hands-on B2SAFE tutorial, which includes information about PID and iRODS integration, was developed – it was delivered in December 2015 and was received well.

Although the training material was in good shape generally, the updates of the presentations extended beyond updating the style to the new templates, to include updates to the wording, strengthening of the key messages and validation of the information being conveyed. All drafts by the user documentation team were validated by the service area experts before being passed on to the training team.

Communication between the two teams is ensured by the respective Task Leaders attending each other’s calls and meetings, and monitoring email from the relevant mailing lists.

7.3. Future Plans

The user documentation activities will remain proactive in identifying new services and changes to existing services that require documentation. There is already contact with the B2HANDLE developers, as updates to the service are expected in the near future, including rebranding, and the documentation team has also approached the people working on B2NOTE indicating our availability to collaborate on documentation and training. It is also expected that updates to the B2SAFE material will be required, as the Service Enabling team follows the new documents and feeds back suggestions for enhancements and improvement. There are plans to create a new “Use B2SAFE” document early in 2016, which will be in line with the newly favoured approach for research communities to deposit software onto a EUDAT core site rather than installing it themselves. The model will be piloted by the new project from the Call for Collaboration, and it is expected that the liaison with Work Package 4 will pick up as a result of this activity. Updates to B2STAGE and B2SHARE will also take place and the expectation is that the new material for B2ACCESS will be published early in 2016, and then be updated the following year.

Generally the emphasis of the documentation and training work is expected to shift to training, in line with the planned training events⁹⁶. A particular challenge that will be addressed soon involves hosting the Virtual Machines required for training, as exemplified by the new B2SAFE training material.

⁹⁶ D3.2 Training Plan M13-M36

8. SERVICE BUILDING REQUIREMENTS AND ROADMAP

EUDAT is a user-driven project in which research communities and researchers define the functionalities that are provided within the EUDAT CDI network. To support the adaptation of the CDI network to new requirements in an agile way, a service building requirements procedure has been established in which new requirements are submitted, assessed, and then approved or rejected. The service building requirements procedure is described in section 8.1. After a new requirement has been approved, it will ultimately find its way into the EUDAT Service Building Roadmap.

Via the Service Building Roadmap, information is communicated about planned releases of new functionalities within existing and new services. This information is essential for both research communities and project enablers, because it defines when certain functionality or new services will become available. It is also vital for the service building team, because the roadmap specifies plans on a general level for the developments that need to be undertaken. The roadmap is described in more detail in section 8.2.

8.1. Service Building Requirements Procedure

The services that are provided and offered to EUDAT's research communities and users are defined by and jointly developed in collaboration with the research user communities. The research communities are involved in the definition and development process of services to ensure that the services developed by EUDAT match the requirements of the research communities. Defining the requirements for building services is a multi-staged process.

1. In the first stage of this process, the research communities and users express their needs for certain functionality – this can be done either by identifying gaps in existing services or by proposing a whole new service. At this stage the requirements are described by the research communities and/or users in the vocabulary of the communities and/or users.
2. EUDAT is about providing common data services and tools, which are useful to multiple communities and which can be used in a cross disciplinary manner. The second stage of the process of defining requirements is therefore to identify the common aspects in the requirements that have been identified. These common features are aggregated and translated to common requirements. The common requirements are described in terminology familiar to EUDAT and to the CDI network and CDI services. These descriptions are not yet at the level that is required for development to proceed, but can be understood so that the requested functionality and its implications can be assessed and then a decision to approve or reject the requirement can be taken.
3. After a requirement is approved (see below), the requirement must be translated into a technical description that is understandable for a developer. This is done in stage 3.

The process for managing the service building requirements procedure and approving or rejecting requirements is the responsibility of the Technical Committee. The Service Area Managers (SAMs) and development coordinators are responsible for assessing the technical implications and determining the amount of effort that would be needed to implement a proposed requirement. Depending on the result of this assessment, the Technical Committee approves or rejects each proposed requirement. When a requirement is approved, its priority level is defined and the requirement is then planned within the EUDAT Service Building Roadmap.

Figure 33 shows the process of moving from a community requirement to a service building requirement.

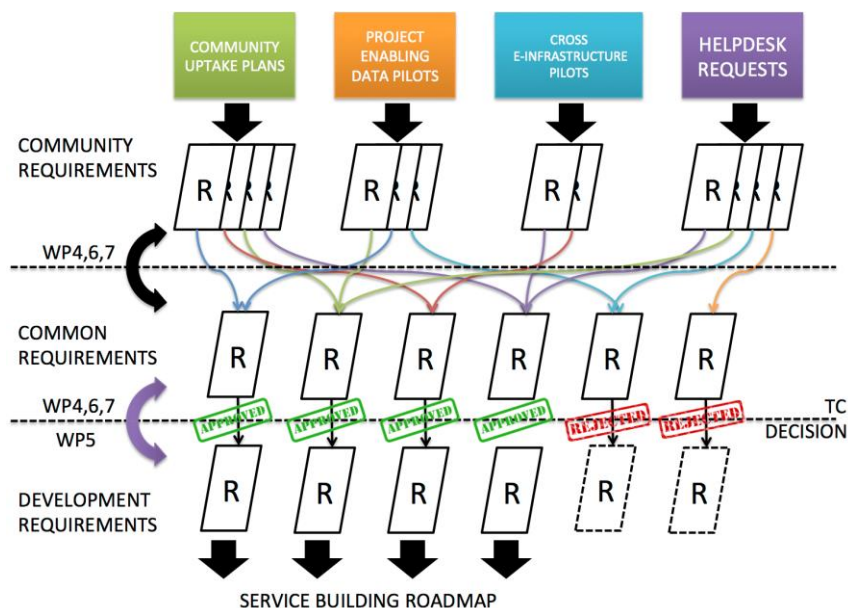


Figure 33: Logical diagram to define service building requirements

In the following sections the stages of defining service building requirements are described in more detail.

8.1.1. Research Community Requirements

In EUDAT there are different channels through which the research communities can express their requirements in respect to how the CDI network is used and the building blocks that are offered, and to identify gaps in the current services where additional functionality is required. The communities can communicate their requirements in the following ways.

- **Uptake plans:** To organize the uptake of the CDI services by the interested research communities, uptake plans have been discussed and agreed upon with the research communities EPOS, CLARIN, LTER, ELIXIR, VPH, ENES and ICOS, which are part of the current phase of the EUDAT project. In these uptake plans, the communities have identified the services they are going to use as pilots. In this process, any relevant requirements and gaps are identified to steer the evolution and development of the EUDAT common services.
- **Data pilots:** To engage and involve research communities that are not currently partners in the project, EUDAT is organizing calls for collaboration. In the first year of this phase of the EUDAT project, EUDAT organized the first call for data pilots. The call invited research communities to test, integrate and/or further develop the CDI building blocks in collaboration with EUDAT. The call resulted in 24 proposals being accepted for pilots.
- **Cross e-infrastructure pilots:** In this phase of the project, EUDAT is engaged with other e-infrastructure providers (such as PRACE and EGI). EUDAT is working with these other e-infrastructure providers to offer joint pilot studies are defined to create bridges between the different types of e-infrastructures (for example, for storage and high performance computation) and thus stimulate cross usage of the different types of services and e-infrastructures. For this EUDAT is collaborating with PRACE on the PRACE DECI calls, in which researchers can put in one proposal requesting the usage of EUDAT services or resources combined with access to PRACE high performance computing resources.
- **Helpdesk:** The EUDAT Helpdesk is the main channel through which research communities and users of the EUDAT CDI network and CDI services can engage with the CDI support organization. Via the helpdesk, users can report issues and bugs in services, and can also ask general questions or propose new requirements. Depending on the nature of the issue ticket that is reported via the helpdesk, the ticket is forwarded within the CDI support organization. Proposals for new requirements or new services are then picked up within the service building requirements process.

The different channels through which requirements for new services can be proposed are managed by different parts of the EUDAT project organization. Requirements that are identified from the uptake plans come from efforts to engage with the research communities (for example, by WP4), whereas the EUDAT

personnel working on enabling aspects of the project (namely WP6) identify requirements that come from EUDAT's data pilots. Cross e-infrastructure requirements are identified as part of the pilots run within our cross e-infrastructure engagement (by WP7). The operational team (WP6) operates the Helpdesk, but many technical people from both the operations (WP6) and development (WP5) teams keep a close watch on the submitted issue tickets, which can generate additional requirements.

8.1.2. Common Requirements

As mentioned in the second bullet point of the service building requirements section (8.1), EUDAT is about providing common data services and tools that are useful to multiple research communities and which can be used across different research disciplines. From the requirements for various types of data services that have been identified through different channels, commonalities are identified. These commonalities are translated to common requirements that should be supported within the EUDAT CDI network and CDI services. An example of this was the request from research communities to extend the support for descriptive metadata within the CDI network. Because metadata support is not a simple requirement, this request resulted in many discussions at the TC level because the necessary changes would have a big impact on the architecture of the CDI and on several developments within the different services, as well as involving the development of a new building block, the metadata store operated at each CDI node.

The challenge in identifying and describing the requirements for common services is to abstract and aggregate requirements from research communities to produce common requirements that suit all the communities. If too much abstraction or aggregation occurs during the process of coming up with a common requirement, the resulting common requirement may not match the requirements of all the communities that were interested in that service, and will consequently be far less useful. In these cases, it is best to submit multiple requirements instead of a single requirement. EUDAT personnel that are involved with the community engagement activities are responsible for identifying commonalities and proposing new requirements based on those commonalities to the service building requirements process.

To describe the specific requirements for building a new or adapting an existing service, a template wiki page (see annex F) is defined on the EUDAT confluence wiki that can be copied. For each proposal for a new service requirement, some basic information (for example, title, short description, requestor contact and community information, goals and impact) must be provided. The description of the requirement must identify the services that are involved or impacted by the requirement. If a requirement consists of multiple aspects that would involve developments in different areas (for example in the areas of PIDs, policies and/or AAI) the requestor can specify sub-requirements for each of the different aspects.

8.1.3. Development Requirements

When a common requirement for a new service or a modification to an existing service is approved, that common requirement must be translated into specific requirements that are described from a development point of view and hence can be understood by a developer. To support the development of a single (sub-)requirement, it can happen that a particular requirement will result in multiple development requirements across multiple services. Depending on the platform that is used to develop or track issues (for example, Github or JIRA), the development requirements may be submitted as separate issues (as happens with Github) or as tickets (as in JIRA). To maintain the relationship between the development requirements and the service building requirements that are submitted, the reference(s) to the development requirement is (are) listed in the service building requirements on the EUDAT wiki.

8.2. Service Building Roadmap

Within the many research community engagement activities within the EUDAT2020 project (such as developing uptake plans, or working on data projects, or project enabling), gaps in the existing services are identified which result in a demand to develop certain functionality. Although the service building activities work package has the largest number of person months (PM) allocated to it, the time available for development effort is limited. Therefore the requirements that are approved are then prioritized and planned. Because certain uptake or enabling work depends on the availability of certain functionality it becomes essential from the service building point of view to communicate about the planned releases within services and to coordinate the activities.

To communicate about the planned releases within the B2 services, a service building roadmap has been developed. The roadmap is organized in three periods for each service, where the first two periods are defined as the first two half-year periods and the final third period is defined as the next six-month period and beyond. The periods are defined on the basis of the project months in this phase of the EUDAT project. The current roadmap is defined for the periods: M13 (March 2016) – M18 (August 2016), M19 (September 2016) – M24 (February 2017) and M24+ (March 2017 and beyond). Defining the roadmap in this manner provides an outlook of a 1½-year period in the service building activities. The full roadmap can be found in Annex G.

9. CONCLUSIONS

The EUDAT2020 project is the continuation of the first phase of the EUDAT project and started on the 1st of March 2015. The end result of the previous phase of the project was the starting point for this phase and laid down the basis of the EUDAT CDI network with the B2 service suite. This deliverable provides an overview of the service building activities conducted within the first year of this phase of the EUDAT project.

The main objectives of the service building activities within this phase of EUDAT are to consolidate the development of the CDI common building blocks (with special focus on integrating the building blocks further), to develop new services and tools to address the full data life cycle, and to define an architectural blueprint of the CDI. In close relation to the development of the services, the service building work package is also responsible for providing technical user documentation and contributing to the development of training material.

The service building activities are organised within three *service areas* (namely data access and re-use, data preservation, and data processing and analysis), each of which is led by a dedicated *service area manager* who steers the developments.

To organize the collaboration for these activities, knowledge transfers and information flow between the technically oriented work packages – such as Community Requirements and Engagement (WP4), Service Building (WP5), Operations (WP6), Cross e-Infrastructure Services (WP7), Data Life Cycle across Communities (WP8), and Technology Exploration (WP9)) – the EUDAT Technical Committee was established. The TC consists of the work package leaders of the aforementioned work packages and the services area managers of the three service areas.

When it comes to defining the blueprint for the EUDAT CDI architecture, the basis of the CDI data model and of the CDI layered architecture has been defined. In the course of the currently ongoing discussions about the *CDI collaboration agreement*, and in regard to the concepts of *using* and *joining* the CDI, the steps for becoming an *interoperable* or an *integrated* CDI node have been defined. In addition, the current status quo with regard to the discussions about *how to support metadata* has been provided, as that relates to the discussions of the CDI architecture. The discussions about the CDI architecture and about supporting new functionalities and services within the CDI network are ongoing discussions. To support the requests for new functionalities and services, a service building request procedure has been put in place. To communicate about releases within the B2 services suite, a roadmap has been defined which is updated every half year.

The main focus on the development side of the service building activities was to consolidate the developments on the existing B2 services and to focus on integration. Highlights of these developments are as follows.

- **B2SHARE:** a new architectural design based on Invenio 3, support for role-based authorization, editable metadata, and integration with B2ACCESS and B2DROP
- **B2DROP:** branding of the B2DROP web interface, automated deployment and integration with B2SHARE and B2ACCESS
- **B2FIND:** enhancing the ingesting of metadata and improve and generalizing the semantic mapping
- **B2ACCESS:** first release and handover of the B2ACCESS service, integration with B2SHARE and with IdP providers such as eduGAIN and social IdPs (like Facebook, Google, or Microsoft), and extensive user documentation to simplify the integration with the B2ACCESS service
- **B2SAFE:** first release of B2SAFE with iRODS v4 support, refactoring of the policies to support iRODS v4, beta release of the Data Policy Manager, assessing integration with B2ACCESS, and providing support for metadata
- **B2HANDLE:** branding of the B2HANDLE service, support for Handle v8, unifying the PID scripts into the B2Handle library, initiating standardization of the EUDAT PID record and policies
- **B2STAGE:** supporting data transfers on the basis of PIDs via GridFTP, the EUDAT Python library with support for data transfers and the search for data in B2FIND, and the definition of the HTTP API

In addition to the developments on the existing B2 services, new developments were initiated: **B2NOTE**, as an uptake from the JRA activities from the first phase of the EUDAT project, and the **Data Type Registry** as a

result of interest from the research communities and as an uptake from the RDA PID Information Type and Type Registries working groups.

In the second year of this phase of the EUDAT project, a number of important functionalities are planned and new services will be introduced. There will be special focus on further integration of the B2 services (particularly B2ACCESS, B2SAFE, B2SHARE, and B2DROP). With the introduction of the local metadata store, the support for managing descriptive metadata within the CDI will be extended. To extend the support for accessing data within the CDI network, there will be additional focus on authorization in B2ACCESS and the related services.

At the beginning of the second year of this phase of the project, an important release of the B2SHARE service, version v2.0, with extended functionality for supporting metadata, PIDs and versioning is planned. Also a number of new services are being introduced: an HTTP RESTful API service to enable up-/downloads via HTTP, the production release of the Data Policy Manager, and pre-production releases of the Data Type Registry and the B2NOTE service.

The full roadmap of the service building activities can be found in Annex G.

ANNEX A. TOGAF ARCHITECTURE DEVELOPMENT METHOD

TOGAF is an open standard⁹⁷ which is widely adopted in Europe for defining Enterprise Architectures and has been developed by the “The Open Group Architecture Forum”⁹⁸. TOGAF is a flexible standard which can be adapted to a specific context. It may be used freely by any organization wishing to develop enterprise architecture for use within that organization.

TOGAF consists of a cycle model for defining, implementing, governing and monitoring an Enterprise Architecture. Figure 34 shows the concepts used within TOGAF for the cycle model and processes.

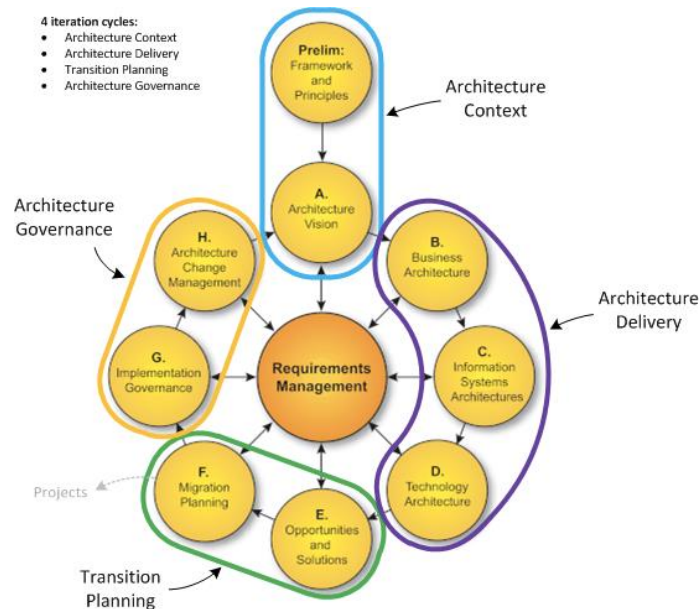


Figure 34: TOGAF Architecture Development Method

A.1. TOGAF Architecture Domains

Within TOGAF, the following four architecture domains are defined to describe the Enterprise Architecture on the different levels of an enterprise:

- **Business Architecture**, which describes the structure and interactions between the business strategy, organization, functions, business processes and information needs,
- **Data Architecture**, which describes the structure and interaction of the enterprise’s major types and sources of data, logical data assets, physical data assets and data management resources,
- **Application Architecture**, which describes the structure of and interaction between the applications as groups of capabilities that provide key business functions and manage the data assets, and
- **Technology Architecture**, which describes the structure and interaction of the platform services, and logical and physical technology components.

⁹⁷ <http://www.opengroup.org/subjectareas/enterprise/togaf>

⁹⁸ <http://www.opengroup.org/getinvolved/forums/architecture>

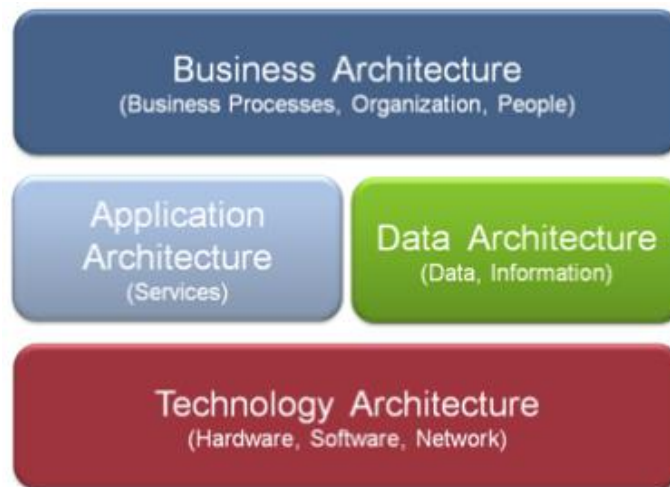


Figure 35: TOGAF Architecture Domains

A.2. TOGAF in Relation to Other Frameworks

TOGAF for defining Enterprise Architectures can easily be tailored and used next to other frameworks, for example ITIL⁹⁹ or FitSM¹⁰⁰ for service management, or PRINCE2¹⁰¹ for project management. In this context TOGAF is complementary to the other frameworks and vice versa.

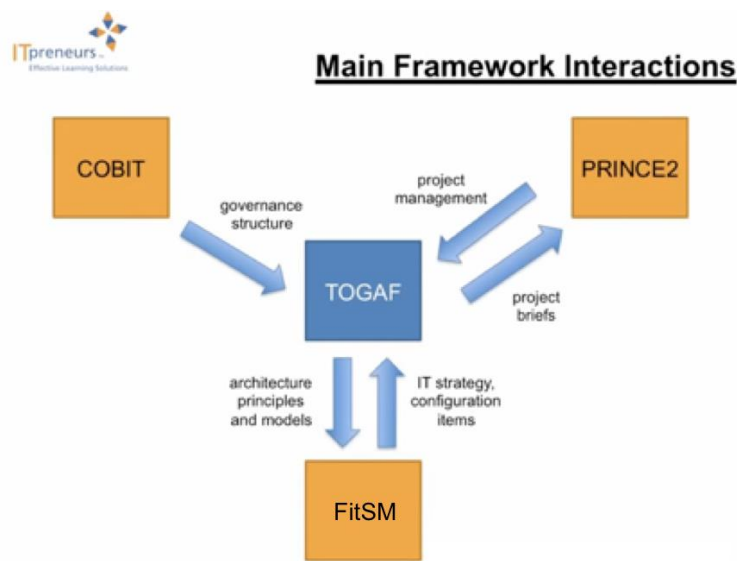


Figure 36: TOGAF in relation to other Frameworks

⁹⁹ <https://en.wikipedia.org/wiki/ITIL>

¹⁰⁰ <http://fitsm.itemo.org/>

¹⁰¹ <https://en.wikipedia.org/wiki/PRINCE2>

ANNEX B. EUDAT CDI DATA MODEL TERMS AND DEFINITIONS

Table 1: EUDAT CDI data model terms and definitions

Data type	Abbreviation	Description	RDA defined
Bitstream	B	A bitstream is a sequence of bits that encodes a specific informational content, either stored on some media or being transferred under control of protocols	Yes
Metadata	MD	Metadata contains descriptive, administrative, contextual and provenance assertions about the properties	Yes
Structural Metadata (Manifest)	MF	Structural metadata is a manifest (e.g. METS and BagIt) describing the structure, content and the relationship between individual objects or entities	No
Persistent Identifier	PID	A persistent identifier is a long-lasting ID represented by a string that uniquely points to a DO and that is intended to be persistently resolvable to access meaningful, current state information about the identified DO	Yes
System metadata	SYS MD	Information that describes those current properties of the DO that are relevant for proper management and access.	Yes, as “state information”
Digital Entity	DE	A digital entity is anything that can be represented by a bitstream.	Yes
Digital Object	DO	A digital object (DO) is represented by a bitstream, is referenced and identified by a persistent identifier and has properties being characterized by metadata.	Yes
Digital Collection	DC	A digital collection is an aggregation which contains DOs and DEs. The collection is identified by its PID and described by its metadata. Therefore a digital collection in itself is also a digital object.	Yes
Digital Package	DP	A digital package is a digital object or a collection of digital objects or entities packaged into a single bitstream	No
Digital Asset	DA	A digital asset is not a specific kind of digital entity, - object, -collection, or –package, but is used to depict all for mentioned into a single word	No

ANNEX C. DESCRIPTIVE METADATA LEVELS

Table 2: Definition of descriptive metadata levels

Metadata levels	Description	Action
EUDAT supported metadata templates	In the B2SHARE service in collaboration with research communities metadata templates are defined to ease the user in describing uploaded data. The metadata templates are also being supported via the HTTP API service.	Metadata is automatically interpreted; a metadata entity is stored in B2SAFE, a copy of the metadata is stored in the local metadata store for harvesting.
Community defined Interpretable metadata	Research communities have defined metadata concepts that are specific to its community or science domain. Commonly these are described in an interpretable format (such as XML) or can easily be extracted via common tools (e.g. NetCDF or HDF5).	Metadata is automatically interpreted; the bitstream is managed as a metadata entity in B2SAFE, a copy of the metadata is stored in the local metadata store for harvesting.
Community Identified metadata objects	Community metadata is described in an uninterpretable format, but is identified as a metadata entity during upload. Uninterpretable means, either binary which requires domain specific tools to do metadata extraction or is at a size which is not suited for human interpretation.	Metadata cannot be interpreted; the bitstream is managed as a metadata entity in B2SAFE, minimum metadata description is stored in the local metadata store for harvesting.
Un-interpretable & unidentified metadata	Community metadata is described in an uninterpretable format and is not identified as metadata entity during upload. In this case the bitstream is stored as a digital entity.	Metadata is not identified; the bitstream is managed as a data entity in the workspace domain.

ANNEX D. DPM TESTBED ENVIRONMENT

Table 3: Overview of the testbed environment to test the Data Policy Manager

Virtual Machines	Operating System	RAM (GB)	VCPU	Disk (GB)	Software Environment	
Cineca dmp1	CentOS 6.7 (x86_64)	2	2	100	Software	Version
					iRODS	4.1.7
					B2SAFE	github master
					iRODS webdav	git master branch snapshot 15 July 2015
					DPM client	git master branch
Cineca dmp2	CentOS 6.7 (x86_64)	2	2	100	Software	Version
					iRODS	4.1.7
					B2SAFE	github master
					iRODS webdav	git master branch snapshot 15 July 2015
MPCDF dmpdev0- eudat.esc.rzg.m pg.de	SUSE Linux Enterprise Server 11 SP3 (x86_64)				Software	Version
					iRODS	4.0.3
					B2SAFE	git master branch, commit on Oct 14, 2015

ANNEX E. B2ACCESS ATTRIBUTES

Table 4: Overview of the B2ACCESS attributes

Name	Mandatory	Multi-Valued	Description
urn:oid:2.5.4.49 distinguishedName	YES	2	The DN as described above (it occurs twice, once with the OID as attribute name and once with distinguishedName)
unity:persistent	YES	1	The persistent identifier as described above
urn:oid:2.5.4.3 cn	YES?	0..2	Common name. Occurs twice.
urn:oid:1.2.840.113549.1.9.1 username	YES?	0..2	Principal. A single value of the form user@domain, where domain is (typically) a DNS-like subdomain representing the security domain of the user (e.g., "osu.edu") and user is generally a username, NetID, UserID, etc. of the sort typically assigned for authentication to network services within the security domain. Occurs twice.
Email	YES	1	email address
memberOf	NO	0..*	A list of strings denoting group memberships of EUDAT communities, roles in EUDAT itself, and roles associated with EUDAT services E.g. /CLARIN, /ENES for the communities, /EUDAT-Staff for members of the EUDAT project, and /eudat:b2safe, /eudat:/b2share for people with roles associated with the services.

ANNEX F. SERVICE BUILDING REQUIREMENTS DESCRIPTION TEMPLATE

General specification			
Nr: #	Title:	Status (III):	
Requestor:		Email:	
Community(ies):		Phone number:	
Date:			
Goal:			
Impact:			
References to uptake plans:			
1			
2			

Requirements specifications:

1	Requirement short description:		Priority (IV):	
Reference to detailed description				
Service dependencies (II)				
Nr:	Service Name	Dependency description		
1				
2				

2	Requirement short description:		Priority:	
Reference to detailed description				
Service dependencies				
Nr:	Service Name	Dependency description		
1				
2				

Requirements assessment:

Requirements assessment (I), to be filled in by the development coordinators						
Req.	Development coordinator	Comment	Feasibility	Estimation Effort	Planning	Status (III)
Nr						
Req.	Development coordinator	Comment	Feasibility	Estimation Effort	Planning	Status
Nr						

Requirements approval:

Requirements Approval (I), to be filled in by the Technical Committee coordinator						
Req.	TC coordinator	Comment	Feasibility	Estimation Effort	Planning	Status (III)
Nr						
Requirement tracker (V):						
1						
2						
Req.	TC coordinator	Comment	Feasibility	Estimation Effort	Planning	Status
Nr:						
Requirement tracker:						
1						
2						

- i. If there are more requirements please add additional requirement sections
- ii. Service dependencies: one of the B2 services suite name or New for a new service
- iii. Status: Submitted, Review, Accepted
- iv. Priority: High, Medium, Low
- v. Requirements tracker: http reference links to JIRA ticket of Github issues

ANNEX G. SERVICE BUILDING ROADMAP M13

Table 5: Service Building Roadmap M13

Service	M13-M18 Mar 2016 - Aug 2016	M19 - M24 Sep 2016 - Feb 2017	+M24 Mar 2017 -
B2SHARE	<ul style="list-style-type: none"> •Release of B2SHARE 2.0 •Support Digital Object Identifiers (DOIs) •Support metadata management community data managers •Support download statistics •Support for Puppet and Docker •Integration with B2DROP •Pilot with B2SHARE as Metadata Store 	<ul style="list-style-type: none"> •Support authorization management •Support versioning •Support for cloud storage services (e.g. DropBox and Google Drive) •Support for more storage back-end solutions (e.g. cloud-storage and object stores) •Integration with B2SAFE •Pilot integration with B2NOTE and DTR 	<ul style="list-style-type: none"> •Extend integration with B2SAFE •Integration with B2NOTE and DTR •Integration with the EUDAT Generic Execution Framework (GEF) •Support for metadata extraction and data exploration via the GEF
B2DROP	<ul style="list-style-type: none"> •Integration with B2SHARE and B2ACCESS •Support for Puppet and Docker 	<ul style="list-style-type: none"> •Pilot with federation B2DROP (e.g. OpenCloudMesh) 	
B2FIND	<ul style="list-style-type: none"> •Continued community integration •Improved user experience •Resolve granularity issues •Integration with B2NOTE •Performance and scalability improvements 	<ul style="list-style-type: none"> •Customisation of GUI for communities •Prototype of SRU interface •Extend harvesting methods (OGC / CSW) •Improved search functionalities for hierarchical search and taxonomies •Improved semantic mapping 	<ul style="list-style-type: none"> •SRU interface in production •Integrate with EUDAT CDI Metadata Store •Integration with Data Type Registry
DATA TYPE REGISTRY (new)	<ul style="list-style-type: none"> •Pilot with Data Type Registry •Collaborate with pilot communities on further evaluation and adaptation of the DTR 	<ul style="list-style-type: none"> •Pre-production release DTR service •Define B2SHARE metadata templates •Define PID profiles •Branding DTR as B2 service •Pilot integration with B2 services 	<ul style="list-style-type: none"> •Production release DTR service •Integration with B2 services
B2ACCESS	<ul style="list-style-type: none"> •Integration with B2SHARE, B2DROP, B2SAFE for authentication •Integration B2STAGE Library and HTTP API for authentication •Integration with Data Project Coordination Portal (e.g. DPCP) •Installation packages & distributed setup •Pilot for distributed authorisation (e.g. XACML) •Pilot IdP integration with PRACE and EGI 	<ul style="list-style-type: none"> •Extend integration with external community sites •Basic IdP Integration with PRACE and EGI •Support basic distributed authorisation •Pilot integration with DTR 	<ul style="list-style-type: none"> •Extend integration with external community sites •Full IdP integration with PRACE and EGI •Improvements based on change requests from communities and partners

Service	M13-M18 Mar 2016 - Aug 2016	M19 - M24 Sep 2016 - Feb 2017	+M24 Mar 2017 -
B2SAFE Core	<ul style="list-style-type: none"> •Support for metadata ingest •Improved performance •Integration with B2ACCESS •Handle v8 support 	<ul style="list-style-type: none"> •Authorization •Local metadata store and harvesting •iRODS v4.2 support 	<ul style="list-style-type: none"> •Support for data packages (e.g. SIP and AIP) •Message bus system
B2SAFE Data Policy Manager	<ul style="list-style-type: none"> •Pre-production release DPM •Full data life cycle replication support •Integration with B2ACCESS •Integration with Data Project Coordination Portal 	<ul style="list-style-type: none"> •First production release DPM •Reviewed policy schema's •Delegated authorization support •Initial support curation policies •Extend support for other services 	<ul style="list-style-type: none"> •Replacement of DPM agent with HTTP API •Support for hierarchical policies
B2HANDLE	<ul style="list-style-type: none"> •Support for Handle V8 •Support for EPIC v2.5 •Standardized PID profiles •PoC central PID catalog, assess requirements 	<ul style="list-style-type: none"> •Release generic Handle library (e.g pyhandle) •Release EUDAT PID library •Pre-production release central PID catalog, revisit requirements 	<ul style="list-style-type: none"> •Production release PID catalog •Pilot integration with DTR
B2STAGE GridFTP	<ul style="list-style-type: none"> •Support PID data retrieval 	<ul style="list-style-type: none"> •iRODS v4.2 support •Support for digital objects 	
B2STAGE HTTP	<ul style="list-style-type: none"> •Pre-production release HTTP API service, •Upload/downloads basic entities in B2SAFE •Integration with B2ACCESS 	<ul style="list-style-type: none"> •Production release HTTP API service •Support for metadata •Support for digital objects •Support for authorization 	<ul style="list-style-type: none"> •Support for digital packages •Support for digital collections
B2STAGE Library	<ul style="list-style-type: none"> •Beta release EUDAT Python library •Support GridFTP •Support data transfers via Globus Online •Support B2FIND •Integration with B2ACCESS 	<ul style="list-style-type: none"> •Production release EUDAT Python library •Support HTTP API •Support data transfers via FTS •Support B2HANDLE 	<ul style="list-style-type: none"> •Extend support HTTP API functions •Further integration B2 services suite
B2NOTE	<ul style="list-style-type: none"> •Demo release B2NOTE, available for testing 	<ul style="list-style-type: none"> •Pilot release B2NOTE 	<ul style="list-style-type: none"> •Integration with B2FIND, B2SHARE and B2HANDLE

ANNEX H. GLOSSARY

Term	Explanation
AAI	The Infrastructure and services to provide authentication and authorisation
ANDS	The Australian National Data Service organisation
API	Application Programmable Interface
ARK	Is a URL based multi-purpose identifier for information objects of any type
BagIt	Is a hierarchical file packaging format designed to support disk-based storage and network transfer of arbitrary digital content.
Bitstream	A bitstream is a sequence of bits that encodes a specific informational content, either stored on some media or being transferred under control of protocols
B2ACCESS	Brand of the EUDAT service for federated authentication and authorisation
B2DROP	Brand of the EUDAT trusted cloud storage service
B2FIND	Brand of the EUDAT central metadata catalogue
B2HANDLE	Brand of the EUDAT persistent identifier service
B2HOST	Brand of the EUDAT service to deploy community applications close to the data storage location
B2NOTE	Brand of the EUDAT service to manage semantic annotations
B2SAFE	Brand of the EUDAT service via which the data management policies are implemented within the CDI network
B2SAFE DPM	Brand of the EUDAT B2SAFE data policy manager via which community data manager are able to manage policies within the CDI network from a central portal
B2SHARE	Brand of the EUDAT easy-to-use data repository service
B2STAGE	Brand of the EUDAT comprehensive set of API's and tools to access data managed within the CDI network
B2 service suite	Aggregation name of the EUDAT B2 services
CDI node	Generic of Thematic service provider who has signed the CDI collaboration agreement either as interoperable or integrated partner
CDI	EUDAT Collaborative Data Infrastructure
CDI Gateways	Services or service endpoints (e.g. API or WUI) which are part of the Access Layer of the CDI layered architecture.
CLARIN	CLARIN (Common Language Resources and Technology Infrastructure) ERIC organisation which provides easy and sustainable access for scholars in the humanities and social sciences to digital language data (in written, spoken, video or multimodal form), and advanced tools to discover, explore, exploit, annotate, analyse or combine them, wherever they are located.
CLDR	Provides key building blocks for software to support the world's languages, with the largest and most extensive standard repository of locale data available
CLI	Command Line Interface
CNRI	Corporation for National Research Initiatives, is a non-for-profit organization which develops and maintains the Handle and Cordra technologies

Cordra	Technology for the creation of, and access to digital objects as discrete data structures with unique, resolvable identifiers (e.g. data types), developed and maintained by CNRI
Digital Asset (DA)	A digital asset is not a specific kind of digital entity, -object, -collection, or – package, but is used to depict all for mentioned into a single word
Digital Collection (DC)	A digital collection is an aggregation which contains DOs and DEs. The collection is identified by its PID and described by its metadata. Therefore a digital collection in itself is also a digital object.
Digital Entity (DE)	A digital entity is anything that can be represented by a bitstream.
Digital Object (DO)	A digital object (DO) is represented by a bitstream, is referenced and identified by a persistent identifier and has properties being characterized by metadata.
Digital Package (DP)	A digital package is a digital object or a collection of digital objects or entities packaged into a single bitstream
Docker	Is an open source technology that automates the deployment of applications inside software containers
DOI	Persistent identifiers within the 10.xxxx prefix namespace domain, issued at a registration agency which is a member of the International DOI Foundation (IDF)
DONA	Foundation to provide management, software development and other services for the technical coordination, evolution, application and other use in the public interest around the world of the Digital Object (DO) Architecture, including its logical extensions and follow-ons, with a mission to promote interoperability across heterogeneous information systems.
DPM	B2SAFE Data Policy Manager
DSPACE	Is an open source repository software package typically used for creating open access repositories for scholarly and/or published digital content
eduGAIN	The eduGAIN service interconnects identity federations around the world, simplifying access to content, services and resources for the global research and education community. eduGAIN enables the trustworthy exchange of information related to identity, authentication and authorisation (AAI).
ENES	European Network for Earth System modeling developing a common climate and Earth system modeling distributed research infrastructure in Europe
ePIC	Consortium of European partners in order to provide PID services for the European Research Community, based on the handle system (TM, http://www.handle.net/), for the allocation and resolution of persistent identifiers
ERIC	European Research Infrastructure Consortium
FitSM	Standard for lightweight service management in federated IT infrastructures
FTS	File Transfer Service developed by CERN
Globus Online (GO)	Service provided and operated by the Globus organisation to manage data transfers
GraphDB	Technology that implements a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data
GridFTP	Is a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks

GOCDDB	Technology to maintain general information about sites and service endpoints, is used as site and service registry within the CDI network
Handle	Technology for the creation of, and access to resolvable unique identifiers, is developed and maintained by CNRI
HDF5	HDF5 is a data model, library, and file format for storing and managing data
HTTP	Is an application protocol for distributed, collaborative, hypermedia information systems
iCAT	Central database in an iRODS zone
IdP	Organizational Identity Provider within an identify federation
Invenio	Is a free software suite which enables you to run your own digital library or document repository on the web, which is maintained by an internal collaboration led by CERN
IRI	Internationalized Resource Identifier
iRODS	Technology for flexible policy based data management of files and metadata that span storage devices and locations, developed and maintained by the iRODS consortium
ISO639	Is a standardized nomenclature used to classify all known languages
Manifest	A manifest (e.g. METS and BagIt) describing the structure, content and the relationship between individual objects or entities
Metadata	Metadata contains descriptive, contextual and provenance assertions about the properties
METS	is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema of the World Wide Web Consortium
NetCDF	Is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.
OAIS	Is an ISO reference model for an Open Archival Information System
OAI-PMH	The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) which is a low-barrier mechanism for repository interoperability.
OAuth	Is an open standard for authentication, commonly used as a way for internet users to log into third party websites
OLA	Operational Level Agreement
OpenID	Is an open standard and decentralized authentication protocol
OpenStack	Is a free and open source software platform for cloud computing to provide Infrastructure-as-a-service
ownCloud	Enterprise file sharing solution for online collaboration and storage, developed by the ownCloud company
Persistent Identifier (PID)	A persistent identifier is a long-lasting ID represented by a string that uniquely points to a DO and that is intended to be persistently resolvable to access meaningful, current state information about the identified DO
Puppet	Is an open-source configuration and management tool
PURL	Persistent URL

RDA	The Research Data Alliance (RDA) is a community-driven organization from in the European Commission, the United States Government's National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation with the goal of building the social and technical infrastructure to enable open sharing of data.
RDA-DFT	The RDA Data Foundation and Terminology working group
RDF	Resource Description Framework (RDF) is a family of the World Wide Web Consortium specification originally designed as a metadata model, commonly used as a general method for conceptual descriptions or modelling of information
Registered Data Domain	Data domain in which digital objects are stored and managed such that data carrying associated descriptive metadata are discoverable and referable/retrievable by persistent identifiers.
RESTful	Representational State Transfer (REST) is a software architectural style of the World Wide Web. Systems or services that conform to the constraints of REST can be called RESTful.
SAML	Security Assertion Markup Language is an XML-based, open-standard data format for exchanging authentication and authorisation data between parties, in particular, between an identity provider and a service provider.
Service Area Manager	The <i>service area manager</i> steers the developments and acts as the liaison towards the other work packages for each of the services within his/her own service area.
Service Catalogue	User/customer facing list of all live services offered along with relevant information about these services. Note: The service catalogue can be regarded as a filtered version of the service portfolio that is offered to customers/users.
Service Management Framework (SMF)	Is the organization, the organizational processes and services to operate and to provide support within the CDI network
Service Management Infrastructure (SMI)	Services to operate and to provide support within the CDI network
System metadata	"metadata" information that describes those current properties of the DO that are relevant for proper management and access.
Service Portfolio	Internal list that details all the services offered by the service provider (those in preparation, live and discontinued). Note: The service portfolio includes meta-information about services such as their value proposition, target customer base, cost and price, risk to the provider, service level agreements offered and operational level agreement supporting them.
Service Provider	Organisation or federation or part of an organisation or federation that manages and delivers a service or services to customers
SLA	Service Level Agreement
SRU	Is a standard XML-based protocol for search queries, utilizing CQL – Contextual Query Language
Swift	Technology to implement a highly available, distributed, eventually consistent object/blob store, is part of the OpenStack software stack

Technical Committee	Governance body to organize the collaboration, knowledge transfers and information flow between the technical-oriented work packages (e.g. Community requirements and engagement (WP4), Service building (WP5), Operations (WP6), Cross e-infrastructure services (WP7), Data life cycle across communities (WP8), Technology exploration (WP9))
TOGAF	The Open Group Architecture Framework, which is a method to define enterprise architectures
Unity IDM	Technology for identity, federation and inter-federation management, is used within the B2ACCESS service
URL	Uniform Resource Locator
URN-NBN	Bibliographic identifiers function as names for objects that exist both in print and, increasingly, in electronic formats
Webdav	Web Distributed Authoring and Versioning is an extension of the HTTP that allows clients to perform remote web content authoring operations
Workspace Data Domain	Data domain is used to store and manage temporary working-copies of data, transient bitstreams of digital entities or objects which are not subject to data management operations yet
WUI	Web based User Interface
W3C	World Wide Web Consortium
XACML	The eXtensible Access Control Markup Language (XACML) is an OASIS open standard that defines a declarative access control policy language implemented in XML and a processing model describing how to evaluate access requests according to the rules defined in policies
XML	Extensible Markup Language is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable
X.509	is an <u>ITU-T</u> standard for a public key infrastructure (PKI) and Privilege Management Infrastructure (PMI)