![x]

EUDAT Primer

About

Introduction to EUDAT for Community Data Managers and the general public.

Modified: 14 October 2020

Synopsis

This document explains the steps that Data Managers and scientific communities need to take in order to join the EUDAT infrastructure or to use its services. It also includes a brief introduction to the services for the end-users and concludes with an explanation of why EUDAT is important.

For community data managers and system administrators: There are two modes of integration with EUDAT:

1. "join" EUDAT: the datacentre/institute becomes a part of the EUDAT network. To this end it is necessary to adopt the EUDAT Collaborative Data Infrastructure (CDI).
2. "use" EUDAT: an EUDAT datacentre ingests the community data and ensures its storage and replication. EUDAT will also expose several web-interfaces and APIs to its services, which will allow a looser, but still feature-rich connection with EUDAT.

In both cases the community will benefit from access to EUDAT's primary services – B2SAFE for safe replication, B2ACCESS for user authentication and authorisation, B2FIND, the metadata harvesting and cataloguing service, B2SHARE for publishing, storing and sharing of smaller data sets, B2STAGE for transferring data to high-performance compute servers and B2DROP for storing, synchronising and exchanging data with colleagues and team members. Joining EUDAT is more demanding, but it allows to benefit from the complete set of EUDAT datacentre features such as: choice of datacentres to federate with by means of B2SAFE; use of an own PID service within the EUDAT domain; optimised data transfer within the EUDAT CDI; and tighter control of data management policies (replication, authorisation) through direct management of core services at a site.

Joining EUDAT requires the installation and configuration of a minimum software stack (the EUDAT CDI). Furthermore, it requires the community to dedicate some resources (storage, compute and man-power) to the project. In practice, three packages are required: iRODS deployed at the site; access to a system for PIDs (or B2HANDLE); and (optionally) deployment of the B2STAGE DSI package. Configuration requires: access to disk space and connection to other iRODS installations in EUDAT; monitoring configuration; user-access configuration; and connection to the metadata catalog. The definition of the Common Services Layer Interface (CSLI) API for Community Data Managers to integrate with EUDAT through the use mechanism is under development at the time of writing this document.

For end-users: The EUDAT services B2ACCESS, B2SHARE, B2STAGE, B2FIND and B2DROP all offer interfaces for end users, and we cover these interfaces below.

EUDAT Services

EUDAT has developed a service stack that forms the Collaborative Data Infrastructure (CDI). The services are as follows:

1. B2SAFE, Replicate Research Data Safely
2. B2STAGE, Get Data to Computation
3. B2FIND, Find Research Data
4. B2SHARE, Store, Share and Publish Research Data
5. B2DROP, Sync and Exchange Research Data

Figure 1. The EUDAT Services

In addition, a set of EUDAT core operational services, essential for the management of the CDI have been defined as follows:

1. B2ACCESS (identity and authorisation), easy-to-use and secure Authentication and Authorisation platform
2. B2HANDLE (persistent identification management), a service to register persistent identifiers called Handles to data objects and retrieve data objects via these identifiers, serving a purpose similar to DOIs for papers
3. B2HOST, a Service Hosting Framework which allows communities to deploy and operate their own applications and data-oriented services on machines next to the data storage location

Most of these components are depicted in Figure 1 and described in more detail in section Service Descriptions below. We then discuss the options for your community to join EUDAT. The document also discusses how to use the components via user-interfaces.

Service Descriptions

The following table gives a brief overview of the functionality offered, the technologies involved and the interfaces exposed by the EUDAT services.

Service	Purpose	Description	Technology	Interface
B2SAFE	Policy-based data management	Core data management service allowing for the automatic, rule and policy-driven replication of data across a federation of EUDAT CDI datacentres (either community or non-specific)	• IRODS (v3.3, v4.1) • EUDAT iRODS microservices.	• IRODS-to-iRODS (via microservices) • Federation controlled by iCAT databases
B2STAGE	Dynamic replication for processing	High-performance data movement service which allows data to be staged into and out of the CDI to, for instance, external high-performance computing services	• EUDAT DSI GridFTP server	• GridFTP (external) • HTTP API • GlobusOnline (external) • iRODS (SR)
B2SHARE	Storing, sharing and publishing research data	User-facing service which provides an easy way to upload, tag and share research data. Furthermore, uploaded data are made citeable via PIDs.	• Invenio plus EUDAT customisations	• Web portal (accessible via HTTPS) • HTTP REST API
B2FIND	Searching and accessing research data	Service exposing metadata catalogue harvested across EUDAT, through a user-friendly, web-based search portal and a standard API.	• EUDAT-customised CKAN (search portal with facetted search functionalities) • OAI-PMH (to harvest metadata from communities)	• Web portal (accessible via HTTP) • API (CKAN)
B2DROP	Storing, synchronising and exchanging research data	A service for storing, synchronising and exchanging dynamic research data with colleagues and team members	• OwnCloud	• Web portal (accessible via HTTPS) • WebDAV
B2ACCESS	Identity and authorisation	A federated cross-infrastructure authorisation and authentication framework for user identification and community-defined access control enforcement	• Unity • EduGain	• OpenID, SAML, x.509, Social • OAuth 2
B2HANDLE	Persistent Identifier management	Core PID registration service based on the global Handle system	• Handle system server (deployable locally or usable centrally) • Handle system server client toolkit.	• Handle system server client toolkit
B2HOST	Service Hosting Framework	Service available on four providers to allow computation close to stored data.	• EUDAT RCT	• Web portal (accessible via HTTPS)

![x]

"Joining" EUDAT or "using" EUDAT?

This document makes a clear distinction between two types of engagement with EUDAT: tight integration by joining the CDI and becoming one of EUDAT's datacentres; and loose interaction, using EUDAT's services through the Common Services Layer Interface (CSLI), which is currently under construction. This is illustrated schematically in Figure 2.

In this schematic we represent a community datacentre as a blue data disk accompanied by a green "community software stack" lozenge, representing tools and other software components particular to that community. A "generic" datacentre is depicted without a green software stack. The orange lozenge represents the core software stack for an "EUDAT Node", including all software required to deliver the core services. The smaller orange square on the boundary of the CDI represents the Common Service Layer Interface, the interfaces that the CDI presents to external users. It can be thought of as the CDI external API.

The other CDI subsystems depicted include the PID registry and the B2FIND metadata catalogue core metadata store and harvester service.

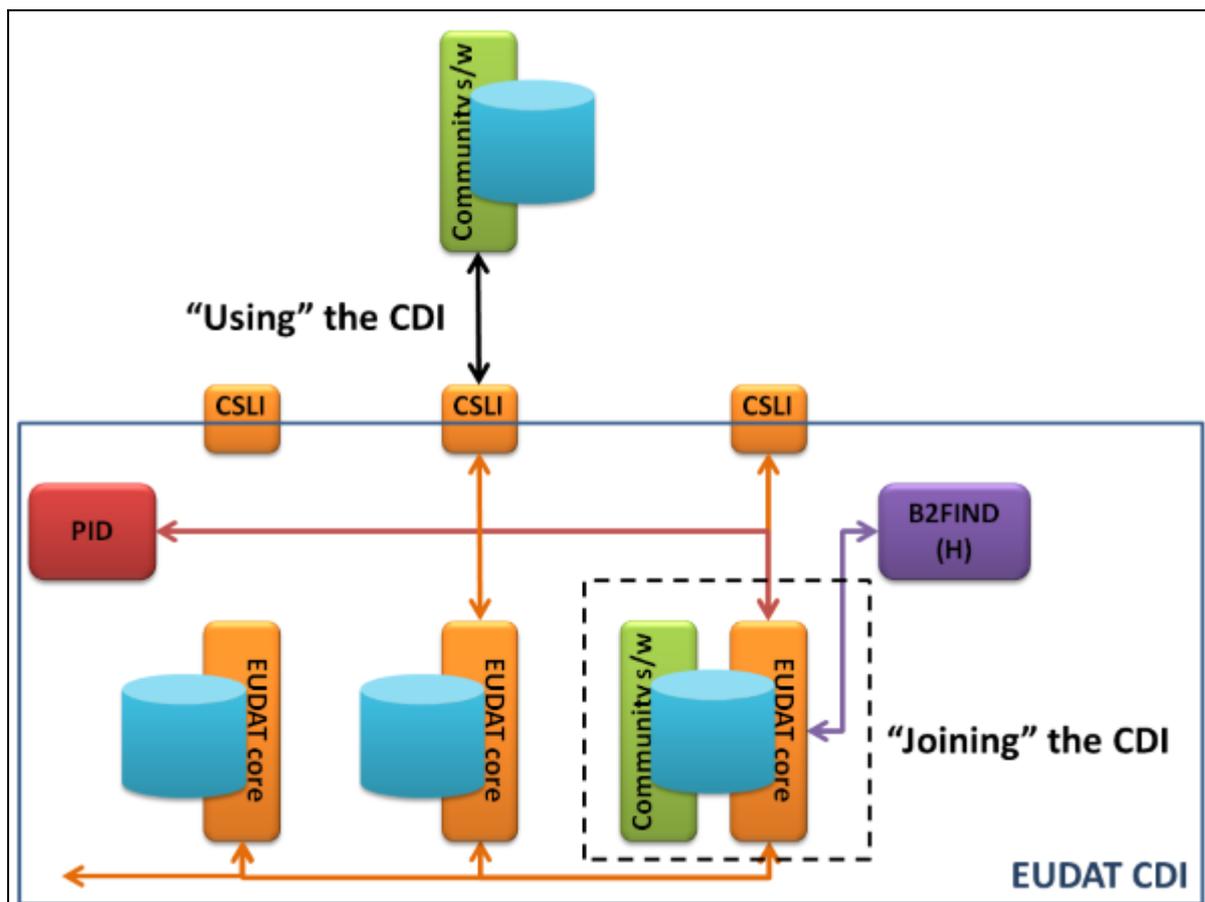The arrows indicate the primary interactions between CDI subsystems (both control and data channels).

Figure 2. Two possible methods of interacting with the EUDAT CDI – joining it (tight integration with all core services) or using it (looser interaction through basic services).

Joining the EUDAT CDI

Joining the EUDAT CDI implies tight integration with all *core* services by installing and configuring the CDI software stack:

1. Installing the EUDAT CDI software stack
2. Configuring the EUDAT CDI software stack, including making three principal federation connections with other CDI subsystems:
   - Federation with other datacentres to enable the safe replication (B2SAFE) service
   - Connection to the central EUDAT B2HANDLE service. This is hierarchical, so centres that wish to deploy their own running PID registration/resolution service for performance reasons are free to do so provided these are configured correctly as part of the EUDAT registration domain
   - (Optional) Connection to the EUDAT metadata harvest service over the OAI-PMH protocol, which enables the B2FIND service
   - (Optional) deployment of an iRODS-compliant, GridFTP server
   - (Optional) offer of computational resources close to the data store
3. Registering the service(s) and enabling monitoring and collection of metadata

When these steps are completed, a datacentre has fully joined the EUDAT CDI and can benefit from the features of tight integration such as:

- Choice of datacentres to federate with for B2SAFE
- Use of an own PID service within the EUDAT domain
- Optimised data transfer within the EUDAT CDI
- Tighter control of data management policies (replication, authorisation) through direct management of core services at their site

In addition to providing significant User Documentation, EUDAT supports new communities wishing to join EUDAT. Dedicated teams are ready to work with the community to successfully complete joining projects. Please contact us with a support request for further information.

Using the EUDAT CDI

Datacentres and data repositories can make use of EUDAT's B2SAFE service to safely replicate their data and thus guarantee their stability and availability. This gives the communities control over replication, but it requires to install iRODS and the B2SAFE module. Starting from late 2016, EUDAT actively promotes use of B2SAFE, whereby the communities select a EUDAT centre to host and replicate their data.

End users like scientists can employ EUDAT's services – B2FIND (metadata), B2SHARE (publishing, storing and sharing of data sets) and B2DROP (sharing, synchronising and storing of smaller, dynamic data sets) and B2STAGE (staging data in/out of a EUDAT site) – without the tight integration required when joining the CDI. To enable this, EUDAT develops and implements a series of simple interfaces to interact with its primary core services. These interfaces are aggregated in the Common Services Layer Interface (CSLI) which contains web-services and an HTTP API.

The CSLI services will be underpinned by the core services within the CDI, and so users making use of the service will benefit automatically from PID registration, B2SAFE etc. However, by this looser interaction users of the CDI will have little or no control over the devlopment of core features and their configurations. E.g. cannot define fine-grained authorisation nor make (detailed) choices on the geographical location of ingested data.

For the researchers themselves, there are user interfaces which enable to use the EUDAT CDI. The usage of the following tools is highlighted below and in dedicated pages:

- B2ACCESS (Primer | Doc): authenticate to EUDAT using a variety of credentials
- B2SHARE (Primer | Doc): web-based publishing, storing and sharing of scientific data sets
- B2STAGE (Primer | Doc): fast data transfer into and out of EUDAT data nodes
- B2FIND (Primer | Doc): search and access data in EUDAT
- B2DROP (Primer | Doc): storing, synchronising and exchanging dynamic research data in the Cloud
- B2HANDLE (Primer | Doc): using PIDs to to reference objects stored in EUDAT
- EUDAT Monitoring Service (Primer | Doc): health status of the EUDAT services
- B2HOST (Primer | Doc): computation close to the data

How to join EUDAT

A technical condition to become an EUDAT-associated datacentre is to install the B2SAFE software stack. Furthermore, data centres need to identify resources they want to contribute to the EUDAT project and its communities. Below we highlight actions to be carried out to join EUDAT (links take you to the corresponding section in the Primer):

- Deploy iRODS, the technology behind EUDAT safe replication B2SAFE
- Acquire a Handle prefix, to allow registering of your data
- Configure B2SAFE, to enable safe replication of your community data
- Request disk space at EUDAT providers, where your community data are replicated to
- Configure Monitoring of your resources
- Manage B2ACCESS to define the users and groups that are allowed to perform certain actions in B2ACCESS-enabled services
- (Optionally) Enable metadata harvesting, to allow users to find and access your community data
- (Optionally) Deploy and configure an iRODS-compliant GridFTP server and a GridFTP client to allow data to move efficiently in and out of your servers
- (Optionally) Contribute PaaS or IaaS services for the EUDAT community to use

Pointers about further information about these actions are available in this chapter. Please also use the online form or the EUDAT Helpdesk (requires log in) to ask for specific technical assistance.

Deploy iRODS

iRODS is a data grid software system, used in EUDAT for data management and data replication between geographically distributed centres. Its main strengths are its policy-based data management facilities and its data-interfacing and sharing capabilities. iRODS is open source under a BSD license, it is community-driven, has a simple installation procedure, is hardware-agnostic, and works on all major operating systems.

EUDAT's large scale iRODS environment is organised into zones where each datacentre hosts one or more zones. Each zone is a distinct administrative domain with its own users, groups and access rights. A zone defines a unique set of hierarchical namespaces for all the objects that it contains, in the same way as a file system. Also, a zone has its own unique special database (the iCAT) where the metadata for all the zone's objects and information on users are stored. Data exchange between zones can be accomplished via zone federation. This procedure makes a remote zone visible and accessible from the local zone, in accordance with the security settings of the remote zone.

A powerful iRODS feature is the rule engine, an automation mechanism that facilitates administration and policy enforcement. Rules are described in a custom but simple language and can be set to run automatically on certain events (e.g. on ingesting a file). E.g. upon data replication between zones, i.e. copying data to another administrative domain, the iRODS rules implemented by EUDAT create new PIDs for the data in the destination

zone and link them to the respective PIDs of the original data objects. By doing this, EUDAT can keep track of replicas.

In depth technical documentation about [deploying and configuring iRODS](#) is available from the EUDAT user documentation site. There is also a page with an [example iRODS zoning configuration](#) and information about the [use of dCache](#) in connection with iRODS.

Acquire a Handle prefix

EUDAT uses persistent identifiers (PIDs) to identify and link data objects. For this reason EUDAT requires that the persistent data of each datacentre be associated with PIDs that can be uniquely attributed to the datacentre.

EUDAT has adopted Handle-based PIDs. The [Handle System](#) is a software infrastructure offering general purpose identifier registration and resolution services. The Handle system uses an open set of protocols, which are designed to allow data curators to create and manage Handles (identifiers) of digital resources in a distributed way. The protocols also allow the users to locate, query the metadata and resolve the PID to a digital resource.

A PID is a reference, an opaque string. A PID in the Handle System is composed of a prefix, a slash character ('/') and a suffix. The prefix denotes the owner of all PIDs created under that prefix. A prefix (and all its PIDs) are hosted at a resolution subsystem (i.e. a Handle server; note that one Handle server can serve many prefixes). A Handle is associated with a container which includes metadata, called the PID Record or Handle Record. The PID can, with the help of these metadata, be resolved to the data object it refers to.

A community joining EUDAT first needs to [acquire a prefix](#). With the prefix acquired, there are two options: the datacentre can run its own Handle system to manage and resolve PIDs; or it can pass the details of its prefix to an EUDAT partner to manage it on their behalf.

Further detail on acquiring a Handle prefix and managing it is available on the [B2HANDLE user documentation page.](#)

Configure B2SAFE

The deployment of iRODS and the acquisition of the Handle prefix are the two prerequisites to facilitate safe replication within the EUDAT network. EUDAT has developed software that integrates with iRODS and links the process of replication with that of synchronising the persistent identifiers of data objects and their associated replicas across the involved datacentres.

The configuration of B2SAFE is discussed on the [dedicated user documentation page](#).

Configure B2ACCESS

B2ACCESS allows EUDAT users to authenticate themselves using a variety of credentials. Users then get access to the EUDAT services, with varying levels of privileges in these services. B2ACCESS is based on the [Unity ID Management technology](#). The process of handling user registration and giving them access to services is carried out on the [B2ACCESS administration portal](#). More information on how to manage B2ACCESS, including a defintion of Groups and Attributes, is available from the [dedicated User Documentation page](#).

Integrate with the B2FIND Metadata Catalogue

Metadata are important for data management, as they describe the dataset that they accompany. Systematically structured and searchable metadata also allow efficient discovery of the data, and they are essential for scientific research in particular. EUDAT offers the B2FIND metadata service, which provides a comprehensive joint metadata catalogue and a powerful discovery portal. Integration with B2FIND is optional, but recommended. It requires that the community run suitable data servers that allow pulling the provided records via an internet protocol. B2FIND prefers as technology the OAI-PMH protocol, as it is a standardised, widely used and easy-to-use protocol, but can work with other formats too. Integration with B2FIND is easy to achieve, as discussed on the [dedicated user documentation page](#).

Deploy and configure an iRODS-compliant GridFTP server (B2STAGE)

This is an optional step, which will allow your site to efficiently ingest large data resources and associate them with iRODS. The service is called [B2STAGE](#). There are two aspects involved in enabling B2STAGE at your site: deploying the EUDAT Data Storage Interface (DSI) module and adding support for Handles and B2SAFE. Support for B2HANDLE and B2SAFE is prepackaged in the respective modules discussed above. The DSI module is based on the established GridFTP package, which EUDAT has enriched with functionality so as to integrate it with iRODS.

You can find in-depth instructions for the installation of B2SAFE in the [EUDAT user documentation](#).

Contribute PaaS or IaaS services (B2HOST)

This is an optional step, which allows communities to deploy and operate their own applications and data-oriented services on computers next to your data storage location. The service is called [B2HOST](#). There are two aspects involved in enabling B2HOST at your site: registering your interest; and configuring the EUDAT Site and Service Registry to publicise your resource offer. More detail on [joining B2HOST](#) is available from the user documentation.

How to Use EUDAT

EUDAT allows communities to use B2SAFE, i.e. to have an EUDAT site host and replicate the community data. This is desribed in more detail [below](#). EUDAT is also preparing a lightweight webservices interface known as the Common Services Layer Interface (CSLI) which will allow simple, slightly limited integation of Community datacentres with EUDAT. More information will appear here as the CSLI is rolled out.

Additionally, EUDAT services provide user interfaces that allow access to some of the functionality. They are as follows:

- The B2ACCESS service ([Primer](#) | [Doc](#)) for end-users to authenticate themselves with EUDAT services using a variety of credentials.
- The B2SHARE service ([Primer](#) | [Doc](#)) for scientists and end-users to independently deposit and publish small data sets into EUDAT either via a web user interface or a REST API.
- The B2STAGE service ([Primer](#) | [Doc](#)) allows users to efficiently move large data sets out of/into EUDAT, integrated with PID capabilities.
- The B2FIND service ([Primer](#) | [Doc](#)) uses metadata to catalogue and locate EUDAT data.
- The B2DROP service ([Primer](#) | [Doc](#)) allows researchers to store, share and dynamically synchronise data.
- The B2HANDLE service ([Primer](#) | [Doc](#)) allows end-users to reference objects stored in EUDAT using long-term stable references.
- The EUDAT Monitoring system ([Primer](#) | [Doc](#)) provides an at-a-glance view of the status of the EUDAT

---

services.

- The EUDAT B2HOST service ([Primer](#) | [Doc](#)) allows your community members to execute computation close to their data.

These are described below.

Using B2SAFE

Using B2SAFE is much simpler for the community than joining EUDAT. The details differ from community to community, but the general workflow is the same.

First, the community needs to select a EUDAT centre that will host the data. Then the rules for replication need to be agreed, as do the mechanisms for ongoing upload of new data. With these in place, EUDAT can pull or accept a push of data from the community, using one of many possible transfer protocols. Finally, as access to the EUDAT-hosted digital objects requires knowledge of PIDs, a mechanism is put inplace for the community to be aware of their PIDs.

More information, including general background to the replication process and the list of technical prerequisites, is available from the [dedicated User Documentation page](#).

Using B2ACCESS

B2ACCESS is the EUDAT federated cross-infrastructure authorisation and authentication framework for user identification and community-defined access control enforcement. Before accessing EUDAT services, users and OAuth clients need to register on the B2ACCESS portal. The following log-in options are supported:

- User's Home Organisation Identity Provider
- Social account (e.g. Google, Microsoft Live and Facebook)
- B2ACCESS ID

B2ACCESS is based on the [Unity ID Management technology](#). Users can easily register with B2ACCESS on the [online portal](#). Further information is available on the [dedicated User Documentation page](#).

Using B2SHARE

B2SHARE is a web-based service for publishing, storing and sharing small data sets, intended for European scientists. The service stores the data at a trusted repository with national backing, in order to provide a professionally managed and supported IT environment.

B2SHARE is designed to be easy to use and currently supports the following functionalities:

- access control
- registration of a PID for any uploaded data collection (can be one file or several files)
- exposing checksums for each of the uploaded files
- and transition of all metadata information to the B2FIND metadata service. B2SHARE enforces the inclusion of metadata accompanying the deposited data, so as to increase the value and facilitate sharing

A screenshot of the tool is available below.

Figure 3. B2SHARE record creation screen

B2SHARE is accessible online. Users can also upload, download and search data in B2SHARE via APIs. There are two APIs, an OAI-PMH API for metadata harvesting and a HTTP REST API for search, retrieval and upload of data. Additionally, B2SHARE is integrated with B2DROP, which allows user to upload data from B2DROP into B2SHARE.

Of particular interest is the EUDAT License Selector, available through B2SHARE. The License helps users pick a suitable license for access to their data.

To ensure the availability of the data in B2SHARE, the service providers have agreed on a Memorandum of Understanding with the EUDAT consortium and keep the data accessible for at least 2 years after the official end of the EUDAT project.

Monitoring of the service is provided via the EUDAT central monitoring facilities.
B2SHARE documentation is available from the EUDAT documentation website; note also the separate documentation page on the REST HTTP API. Please visit our training site on GitHub for B2SHARE hands-on training material.

Using B2STAGE

The B2STAGE service allows data staging, i.e. data transfer into and out of EUDAT datacentres. Data staged into EUDAT are assigned a unique Persistent Identifier (PID). B2STAGE exposes two protocols for staging data, as follows:

- GridFTP (via the EUDAT Data Storage Interface) is aimed at large data transfer and numerous files. It allows for third party transfers. The target group are power users who need to access data in B2SAFE and move them to compute sites.
- HTTP is for small and medium files. The HTTP API allows for access to B2SAFE metadata. The target group are community developers who want to integrate data and features from B2SAFE into their community-specific applications.

The curent B2STAGE implementations are only limited to files managed by the EUDAT B2SAFE service. The data

staging functionality of B2STAGE is realised by extending the iRODS system with: a GridFTP interface, implemented by the EUDAT Data Storage Interface (DSI) Component; and with a separate HTTP API. Integration with the iRODS technology causes incoming files to be assigned a Persistent Identifier (PID) and be passed on to the B2SAFE service for safe replication to another EUDAT site, as per the agreed rules for replication for the user's community.

There are no special clients needed for the HTTP API. Most commands can be issued with the standard curl command line tool or even via the web browser. For the GridFTP interfce any existing client supporting the GridFTP protocol can be employed for B2STAGE, including globus-url-copy, Globus On Line, UberFTP, gTransfer, etc, as per the figure below. Among the available clients there is also the EUDAT File Manager which supports a range of transfer protocols (e.g. GridFTP, FTP, native iRODS, etc.) and provides an intuitive and user-friendly interface. A personal certificate (X.509) is required to access the service.

Figure 4. EUDAT B2STAGE options

Using B2FIND

B2FIND offers an interface with which scientists can search for interesting datasets. Communities can make their metadata searchable in B2FIND by making them harvestable via the standard OAI-PMH interface. Each community decides which metadata are made available for EUDAT. A sophisticated framework ensures that metadata providers are harvested regularly to always display complete and up to date information. EUDAT provides an optimised translation from community metadata schema to standard facets in the B2FIND metadata catalogue.
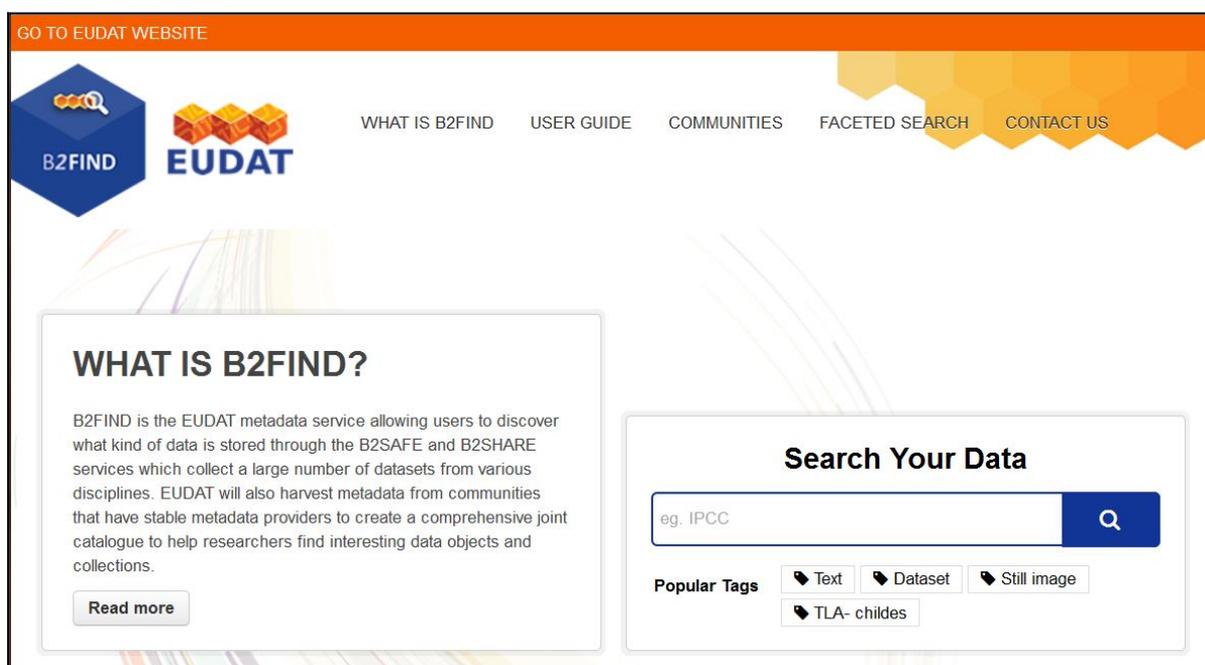

Figure 5. B2FIND Homepage

The B2FIND metadata catalog is acessed via the web page at http://b2find.eudat.eu (see figure 5). User documentation and a description how to make metadta harvestable is available from our website.

Using B2DROP

B2DROP is a cloud-storage based service for storing, synchronising and exchanging dynamic research data with colleagues and team members. It allows researchers to upload, download or remove data, or automatically synchronise data with files on several file systems. The service is intended for the long-tail and still volatile data which can change and are still subject to active research e.g. drafts of research papers. Therefore, B2DROP offers versioning of all ingested files but does **not** attach persistent identifiers to them. B2DROP is hosted at the Jülich Supercomputing Centre, which guarantees that your research data stay in Europe. Daily backups of all files in B2DROP are taken and kept on tape. Data are encrypted on transmission through the exclusive use of the https protocol for data transfer. Each B2DROP user is allowed 20GB of storage.

B2DROP is based on the ownCloud technology. It is available at https://b2drop.eudat.eu. More information is available on our dedicated User Documentation page.

Using B2HANDLE

B2HANDLE is the EUDAT distributed service for management of Persistent Identifiers. It allows communities to associate their EUDAT digital objects with stable identifiers and enables users to reference these objects throughout and past the data lifetime. Users can interact with the service by means of its graphical and programmatic HTML interface, or they can download and use the B2HANDLE Python library.

B2HANDLE is developed based on the established Handle system. More information is available on the B2HANDLE User Documentation page.

Using the EUDAT Monitoring Service

The provisioning of reliable and performing services in a distributed environment requires a monitoring system which allows the service providers as well as responsible operations teams to be informed about the status of the infrastructure and to be notified if any service failure or significant performance drop is detected. One of the key aims when designing the EUDAT monitoring service was to provide a simplified view for community members or people only interested in the status of some of the services.

Figure 6. The entry page of the EUDAT monitoring service

We show the monitoring service entry page an example in Figure 6. This view is light on details, but useful to a user since it enables the determination of the overall status of a service at a glance, rather than having to know which hosts and components are involved in the provision of a front-line service. More information is available on the dedicated user-documentation page.

Using B2HOST

A number of EUDAT providers offer computational resources at their datacentres, through the B2HOST service. These resources are dedicated to help data-handling in EUDAT, i.e. general computations on data are not permitted to the EUDAT users. Examples of valid use include cases where extra compute power might be needed when the required data are too large to be transmitted; or when data licences do not allow the data to leave their datastore. More information of how to use B2HOST is available from the user-documentation page.

More information

You can get more information about the EUDAT services from the following sources:

- The User Documentation pages linked from the User Documentation main page
- The online training presentations
- The hands-on training material on GitHub

Feel free to also use the EUDAT support form for more comments and questions, including comments on User Documentation.

![×]

Why EUDAT Matters

EUDAT's objective is to build a Collaborative Data Infrastructure (CDI) as a pan-European solution to the challenge of data proliferation in Europe's research communities. EUDAT will allow researchers to share their data within and between the communities. The expectation is that the services offered through the CDI will foster innovative, multidisciplinary research. EUDAT aims at providing a data management solution that will be affordable, trustworthy, robust, persistent and easy to use. Why does this matter?

Meeting the Grand Challenges. Europe and its nations are faced with daunting economic, demographic, social and environmental pressures. These Grand Challenges include an ageing population, environmental degradation, the loss of biodiversity, the growing demands for food, water and energy, and the need to respond rapidly to emerging global threats such as pandemics and bioterrorism. The surest way to meet these challenges is to tap into the huge potential of ideas, resources and people across national borders. This necessitates pan-European, interdisciplinary collaboration to help "virtualise" research – to make access to data and research ideas as transparent as possible – and long-term strategic thinking and economies geared towards innovation that will deliver new solutions and create new wealth.

The importance of data. All efforts aimed at addressing the key societal and economic challenges in both research and industry generate increasing amounts of data. It is estimated that the amount of data produced each year is greater than the sum of all that previously created. The growth of data has outpaced the development of tools to deal with it. Today, research results need to be stored for the long term in order to be processed within collaborative research projects. There is therefore an emerging and pressing need to provide infrastructures that will meaningfully integrate new types of data to be collected in the future.

The data infrastructure landscape, in Europe and beyond, is currently fragmented and ill-equipped to deal with these challenges. EUDAT is a direct response to this fragmentation. EUDAT's CDI will provide clear benefits to European researchers struggling to make best use of data. The EUDAT's CDI is also expected to provide the data organisation framework and some the "core services" in the construction of the e-infrastructures planned by many ESFRI approved infrastructures.

Improving data integration – reversing the fragmentation of research. The European research ecosystem consists of a large number of organisations and scientific communities, each producing large volumes of diversified data. The EUDAT CDI by adopting a data model abstraction capable to unify the organisation of the different data types is expected to increase the efficiency and maintain the competitiveness of European research.

Increasing our ability to incorporate and exploit new types of data. Addressing new problems and applying new research methodologies leads to the creation of new types of data, often by combining data from different research communities. Storing and cataloguing data in the EUDAT CDI will enhance our abilities to use and share novel, research results.

Consolidating European resources. The collection, curation, storage, archiving, integration and cross-community, trans-national deployment of modern research data is an immense challenge that can no longer be handled by a single organisation or by one country alone. The EUDAT CDI will facilitate the process of extracting optimum value from current and planned investments in this area.

Improving Open Access to research data. EU research policy requires that all public research results should be available publicly; see, for example, the ERC Scientific Council Guidelines for Open Access. The EUDAT CDI will make Open Access easier to achieve for depositors, and will thus help drive the Open Access agenda across Europe.

Linking efficiently to European compute infrastructure. Large-scale research data often need large-scale computing power to process and analyse them. The EUDAT CDI will make use of both technology and policy approaches to create efficient and effective links between research data and the European high-performance and high-throughput computing infrastructures provided by PRACE, EGI and others.

EUDAT matters because data matter, and data matter as our keys to unlock solutions to the Grand Challenges facing Europe in the twenty-first century.

Document Data

Version: 1.12.1

Authors:

Kostas Kavoussanakis, kavousan@epcc.ed.ac.uk

Cristina Manzano, c.manzano@fz-juelich.de

Emanuel Dima, emanuel.dima@uni-tuebingen.de

Rob Baxter, r.baxter@epcc.ed.ac.uk

Carl Johan Håkansson,cjhak@kth.se

Heinrich Widmann, widmann@dkrz.de

Hans van Piggelen, hans.vanpiggelen@surfsara.nl

Christine Staiger, christine.staiger@surfsara.nl

Read more