Using B2SAFE

About

Document that describes how communities can make use of B2SAFE without running their own iRODS/B2SAFE instance.

Modified: 29 January 2018

Synopsis

This document is targeted at communities who want to make use of B2SAFE but do not wish to deploy iRODS and the B2SAFE module. It explains the functionality that B2SAFE offers in the "using" mode, what software is needed and which questions have to be clarified with an EUDAT centre in order to make use of B2SAFE.

Acronyms

PID: Persistent identifier associated to a file or iRODS collection, usually an EUDAT Handle.

ROR: Repository of Records, (persistent) identifier to the original data object. Can be of any identifier type. If the community does not assign an own identifier the B2SAFE PID will be used.

FIO: First ingested object, the persistent identifier associated to the very first object in in the EUDAT domain. If the the chain has only two elements, the master copy and the first replica, then the PARENT = FIO.

PARENT: Parent PID, the persistent identifier associated to the source object in a replication chain. If the chain has only two elements, the master copy and the first replica, then the PARENT = FIO.

Digital entity: Files and folders

Digital object: Files and folders that carry a persistent identifier and possibly some metadata.

Introduction

B2SAFE offers safe data replication across different data centres. Communities, repositories and data projects can use B2SAFE to distribute valuable data across the EUDAT network in order to keep it safe and to bring it closer to compute infrastructure. In the rest of this section we explain how B2SAFE works.

The underlying assumption of the safe replication of B2SAFE is that the data which needs to be replicated is stored in a repository, the so-called Repository of Records (ROR). In general it is assumed that the data in the repository of records is not to change, i.e. new files may be added to the repository but files do not change over time and are not deleted. The single data files in the repository of records receive a persistent identifier (PID) which is used to link the original data and the replicas. Moreover, the PID is used to store information for integrity checks, e.g. checksums. The ROR is used as a reference to test the integrity between original files and their replicas. This implies that in general we assume that the ROR does not change.

Upon replication from the ROR, new files at other EUDAT sites are created and need to be linked to their respective parent or to their respective originals. To this end B2SAFE makes use of specific PID entries: *ROR, FIO* and the parent PID (*PARENT*). The value in the field *ROR* designates the identifier of the data file in the

EUDAT receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654065.

community repository of records, while the value in the field *FIO* is the PID of the very first ingested in the EUDAT domain. The object's *PARENT* is the PID of the direct parent of the replica. This ensures that we can retrieve the original files of each replica and have reference data objects for integrity checks. To find replicas when given the PID of the original file, the PID field *REPLICA* is used, where all PIDs of direct replicas of the file are stored.

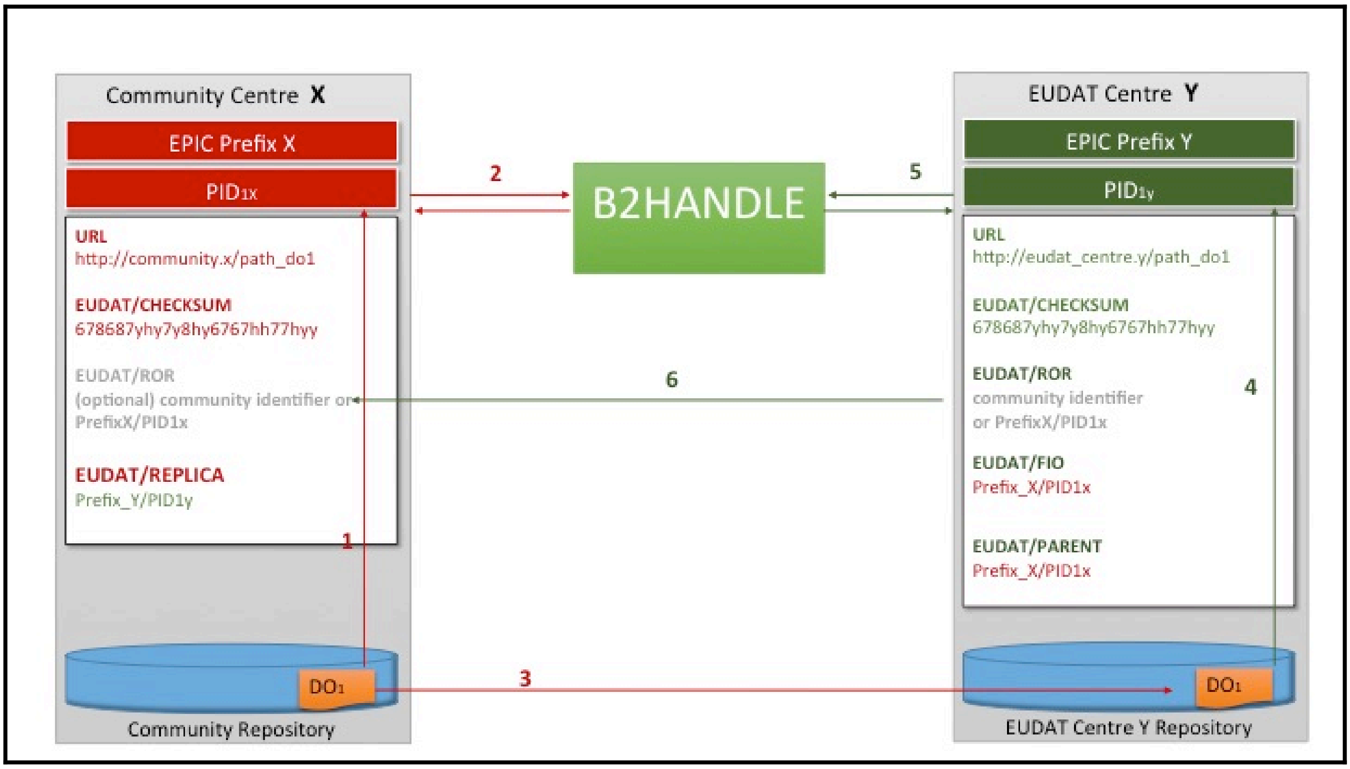

Figure 1: Linking replicas to their original data with PIDs

Figure 1 depicts the flow of linking replicas with their original data. The flow is as follows:

1. The Data Object (DO) is stored at a centre. The figure shows a Community Centre and the workflow discusses it as such, but this is not a requirement.
2. The (community) centre labels the data with a PID which holds information on the data location (URL), its original location (as URL) and some additional data such as the checksum.
3. The data is replicated with B2SAFE to an EUDAT centre Y.
4. Upon replication, the copy of the data object receives another PID. This new PID contains an additional entry ROR, which links to the DO at the (community) centre, a link to the FIO (here we assume the community centre is part of the EUDAT network) and a link to the direct parent, which in this case is the same as the FIO.
5. The PID for the replica is created under a different prefix than the PID for the original DO.
6. The PID of the replica is entered under *REPLICA* in the PID of the original DO.

The replication of files, the generation of PIDs and the linking between originals and replicas are automated by iRODS workflows written as iRODS rules; see the B2SAFE user documentation for some example workflows. For more information on the PID linking in B2SAFE also please refer to the B2SAFE documentation.

In the case of using B2SAFE, the data are moved between the community and one of the EUDAT centres outisde the B2SAFE workflow, and the selected EUDAT centre will ensure ingest of the community data and function as a

ROR from which data can be replicated to other EUDAT centres. This approach spares communities to install iRODS and build the expertise to maintain the iRODS server and define iRODS rules themselves.

General workflow

1. The community needs to select a EUDAT centre that will function as ROR.
2. In collaboration with the EUDAT centre the community needs to define some data policies such as
   - Which other EUDAT centres should the data be replicated to
   - How to transfer any newly added data in the community repository to the ROR
   - Integrity checks
3. Set up a pipeline to transfer the community's data to the EUDAT centre; this can be a push or a pull mechanism. The data can be sent directly to iRODS. In this case the community needs to install the iRODS icommands or, if the EUDAT centre offers B2STAGE, the globus-url tools. Data can also be transferred by other protocols and the EUDAT data centre will ingest the data into their iRODS instance on behalf of the community.
4. Communication of the PID of the ROR to the community. To give the community access to the replicated data, the PIDs of the files in the ROR need to be communicated. EUDAT offers several options to achieve this.

Note, that the steps above illustrate the general workflow. EUDAT is open to work out additional workflows and define additional policies together with the communities as needed.

Technical requirements

- Community:
  - Data transfers: globus-url-copy or any other data transfer protocol that the EUDAT centre supports
  - Database to store PIDs of digital objects in the ROR
- EUDAT centre
  - B2SAFE
  - B2STAGE or other data transfer protocol to receive the community data
  - Database as a means to communicate PIDs to the community

Support

Support for BSAFE is available via the EUDAT ticketing system through the webform.

If you have comments on this page, please also submit them though the EUDAT ticketing system.

Document Data

Version: 1.1

Authors:

Christine Staiger, christine.staiger@surfsara.nl

Giovanni Morelli, g.morelli@cineca.it

Editors:

Themis Zamani, themis@grnet.gr

Kostas Kavoussanakis k.kavoussanakis@epcc.ed.ac.uk

Read more

EUDAT receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654065.