Data Curation and Provenance for the EUDAT Collaborative Data Infrastructure

Alexander Atamas[1], Rene van Horik[1], Linda Reijnhoudt[1], Javier Quinteros[2], Vasily Bunakov[3]
[1] Data Archiving and Networked Services (DANS), The Hague, the Netherlands
[2] GFZ German Research Centre for Geoscience, Potsdam, Germany
[3] Science and Technology Facilities Council (STFC), Didcot, the United Kingdom

Two Use Cases as a Proof-of-Concept for Data Curation Activities

HERBADROP offers an archival service for long-term preservation of herbarium specimen images and extracts information from those images.

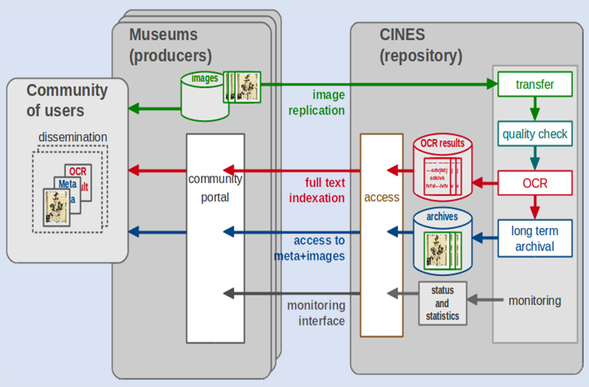


Figure 1. Data workflow of the HERBADROP data pilot.

- B2SAFE service is used in the first step of the ingestion process into the EDTR
- B2SAFE transfer service is engaged to transmit existing images of herbarium specimens along with the associated metadata to the CINES repository
- B2HANDLE is used to manage and store Persistent Identifiers

GEOFON is a seismological data centre providing long-term access for seismic waveforms as well as a data provision centre.

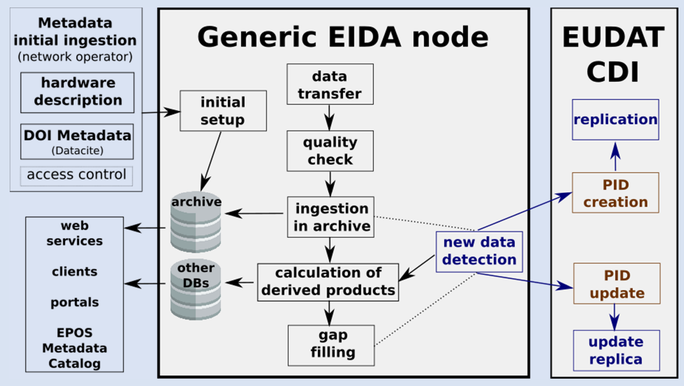
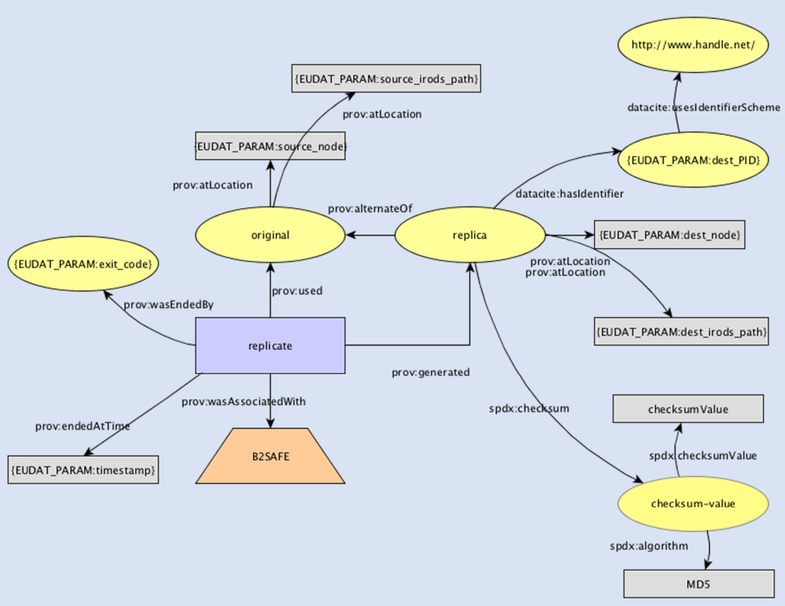


Figure 2. Data workflow from the GEOFON pilot.

- B2SAFE is employed to accomplish data management tasks
- B2SAFE actions are executed after new data is detected
- B2HANDLE is used to manage and store Persistent Identifiers

Provenance of Research Data



Provenance of research data is important for tracking origins, ownerships and modifications of data over the data lifecycle. The concept of provenance guarantees that the data sets made available for sharing and exchange are reliable and hence all data transformations and results obtained using the data sets could be reproduced.

- An HTTP template-based service collecting provenance information is developed to be used by any service from the EUDAT CDI portfolio
- The service defines an API for the clients to generate provenance data based on particular templates (Notation3 format) based on the PROV Ontology (PROV-O), which are made available by its operator
- The gathered provenance data then can be queried by any interested EUDAT service by using HTTP protocol

Figure 3. The template of B2SAFE replication provenance. It depicts entities as yellow ovals, activity as blue rectangle and agent as orange trapezium. The curly brackets denote the parameters used by the template.

Versioning of Research Data

Versioning is required because research data often undergo rather intensive lifecycle. Individual data sets are modified, updated or recalculated. Hence, in the scientific community, there is a need for versioning functionality for proper data curation and provenance of research data. A version of digital object is understood as a timestamped copy of the object.

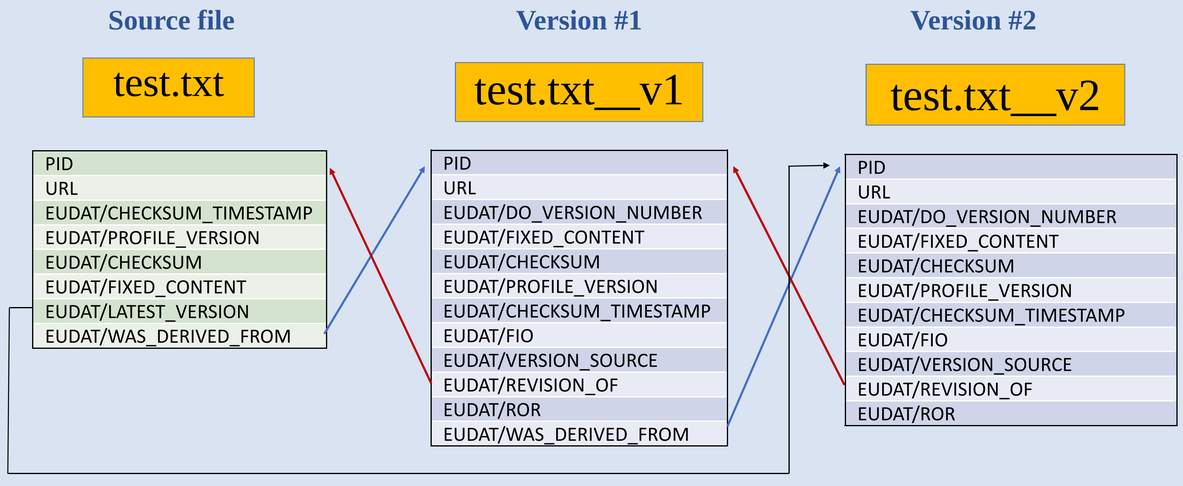


Figure 4. Data workflow of the B2SAFE versioning.

- Versions are stored in a separate directory created only for storage of versions.
- Persistent Identifiers (PID) are employed to identify and access versions independently of physical storage location
- PIDs are generated and used as references within the framework B2HANDLE based on the handle system (http://www.handle.net) to reveal actual URL of a version.
- Every version is made cross-linked with the previous version for the sake of easy navigation between versions