Data curation policies and data provenance in EUDAT Collaborative Data Infrastructure

Vasily Bunakov, Alexander Atamas, Alexia de Casanove, Pascal Dugénie, Rene van Horik, Simon Lambert, Javier Quinteros and Linda Reijnhoudt

Published version information

Citation: V Bunakov et al. "Data curation policies and data provenance in EUDAT Collaborative Data Infrastructure." In: L Kalinichenko, Y Manolopoulos, O Malkov, N Skvortsov, S Stupnikov, V Sukhomlin (Eds.) **Data Analytics and Management in Data Intensive Domains**. DAMDID/RCDL 2017. Communications in Computer and Information Science, vol 822. Springer (2018): 249-263.

DOI: 10.1007/978-3-319-96553-6_18

The final authenticated version is available online at Springer via https://doi.org/10.1007/978-3-319-96553-6_18

Data Curation Policies and Data Provenance in EUDAT Colaborative Data Infrastructure

Vasily Bunakov[1], Alexander Atamas[2], Alexia de Casanove[3], Pascal Dugénie[3], Rene van Horik[2], Simon Lambert[1], Javier Quinteros[4], Linda Reijnhoudt[2]

[1] Science and Technology Facilities Council, Harwell, Oxfordshire, UK
[2] Data Archiving and Networked Services (DANS), The Hague, Netherlands
[3] CINES, Montpellier, France
[4] GFZ German Research Centre for Geoscience, Potsdam, Germany
{vasily.bunakov, simon.lambert}@stfc.ac.uk
{alexander.atamas, rene.van.horik, linda.reijnhoudt}@dans.knaw.nl
{casanove, dugenie}@cines.fr
javier@gfz-potsdam.de

Abstract. The work outlines the development of a data curation and data provenance framework in the EUDAT Collaborative Data Infrastructure. Practical use cases are described, as well as results of defining and implementing data curation policies and data provenance patterns.

Keywords: data curation, data provenance, e-infrastructures, long-term digital preservation, policies.

1 Introduction

EUDAT Collaborative Data Infrastructure (CDI) [1] is a European e-infrastructure of data services and information resources in support of research. This infrastructure and its services have been developed in close collaboration with over 50 research communities spanning across many different scientific disciplines, with more than 20 major European research organizations, data centres and computing centres involved. Researchers, research communities and service providers can use EUDAT data services to manage research data according to their own needs.

The EUDAT services offering [19] has emerged as a result of two consecutive FP7 and Horizon 2020 projects, with the actual services focused on different aspects of data management and data use, and supported by a variety of information technology stacks.

Data curation (or digital curation) is the selection, preservation, maintenance, collection and archiving of digital assets and hence is the essential part of research data management. Sensible data curation requires establishing and developing long-term repositories of digital assets for their current and future use by researchers and wider

society. Collaborative data infrastructures like EUDAT that span across the borders should play a significant role in research data curation.

An important aspect of data curation is data provenance that refers to processes, entities, activities and actors that allow reasoning over data origin, data movement and data transformation. The significance of data provenance in data infrastructures such as EUDAT is caused by a substantial complexity of data workflows that involve multiple data sources, multiple (and interacting) services, and multiple agents, both human and software.

Historically, EUDAT services have been built with only a few considerations for conscious data curation and data provenance, with secure and controlled access to data being one of the major initial goals to achieve. Other aspects of data curation and data provenance started playing a more prominent role when services matured to production stage and became a part of an operational collaborative infrastructure. Specifically, operational requirements of B2SAFE service (that currently offers what long-term digital preservation projects typically call "bit-level" preservation), as well as automated data transfers across interrelated services have made it essential to systematically explore the topic of data curation in EUDAT.

The decision was made to formulate the core approach to data curation with the involvement of two prominent unrelated research communities with substantial amounts of data to manage and then, using these two use cases as a proof-of-concept for clearly formulated data curation activities, get other user communities involved.

Another decision made was to reuse the outputs of the SCAPE project [2] and Research Data Alliance Practical Policy Working Group [3] in order to set up a reasonable data curation framework for EUDAT.

This work expands on the earlier published effort on data curation and policy modelling [20], [21] with the addition of specific data provenance considerations and the actual data policies implementation using the EUDAT data infrastructure. It also offers a consistent vision of the role of data curation and data provenance in research e-infrastructures.

The rest of the paper outlines the core use cases, characterizes the SCAPE and RDA outputs that are deemed to be applicable in EUDAT context, explains mapping of SCAPE policy elements [4] to granular data policies in EUDAT, describes a service for provenance records generation, outlines executable policy implementation effort so far and suggests a semantic approach to modelling data policies. In the end, the role of data curation for the future of e-infrastructures is discussed.

2 HERBADROP use case

2.1 Motivations and relation to EUDAT services

The HERBADROP data pilot [12] aims to offer an archival service for long-term preservation of herbarium specimen images and to develop innovative processes for extracting information from those images. HERBADROP follows the global trend towards scalable industrial-style digitizing of herbaria specimens. It is designed as both an archival service for long-term preservation of herbarium specimen images and

a tool for analysing and extracting information written on the image by using Optical Character Recognition (OCR) analysis, both supported by CINES [6].

Making the specimen images and data available online from different institutes allows cross domain research and data analysis for botanists and researchers with diverse interests (e.g. ecology, social and cultural history, climate change).

Herbaria hold large numbers of collections: approximately 22 million herbarium specimens exist as botanical reference objects in Germany, 20 million in France and about 500 million worldwide. High resolution images of these specimens require substantial bandwidth and disk space. New methods of extracting information from the specimen labels have been developed using OCR but using this technology for biological specimens is particularly complex due to the presence of biological material in the image with the text, the non-standard vocabularies, and the variable and ancient fonts. Much of the information is only available using handwritten text recognition or botanical pattern recognition which are less mature technologies than OCR [13].

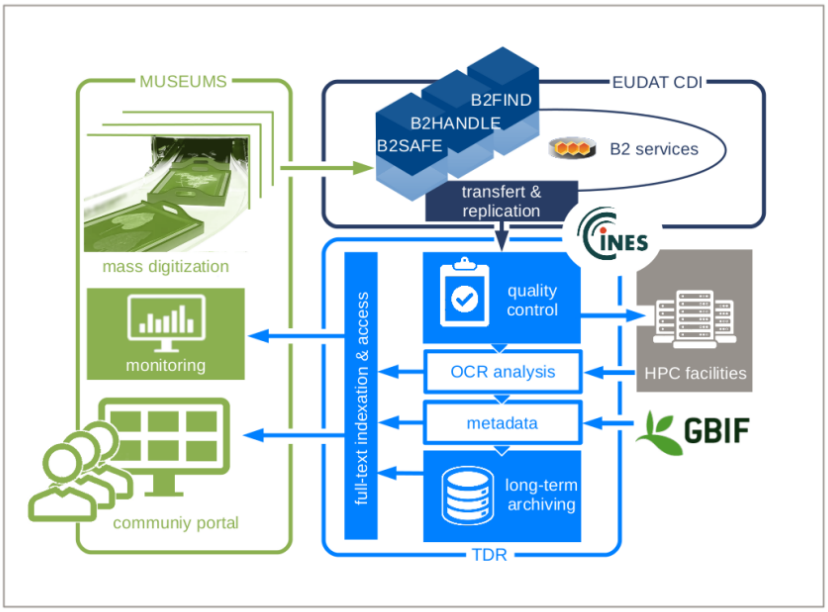

Fig. 1. Data workflow of the HERBADROP data pilot.

The proposed platform is expected to support or even substitute costly manual data input as much as possible. The platform will also curate and enrich metadata resulting from image analysis using optical character recognition (OCR) and pattern matching. Results are exposed as platform independent Web services which can be effectively integrated into herbarium data management systems as well as metadata capture workflows. Since 2016, six European community partners have been involved. Their contribution to the pilot represents a business model that can be potentially replicated

by other institutes. The partners in the HERBADROP data pilot are: Musée National d'Histoire Naturelle (MNHN) – Paris, France; Royal Botanic Garden of Edinburgh (RBGE) – United Kingdom; Botanic Garden and Botanical Museum (BGBM) – Berlin, Germany; Digitarium – Finland; Naturalis Biodiversity Center – Netherlands; Botanic Garden Meise – Belgium.

The EUDAT B2SAFE service is used in the first step of the ingestion process. Existing images of herbarium specimens along with the associated metadata (harvested from GBIF portal) are transmitted to the CINES repository using B2SAFE transfer service. The ingestion into B2SAFE is carried out in accordance with the centralized persistent identifiers (PID) management system used in EUDAT. It is envisaged that discovery and visualization of the data objects will be performed with the EUDAT B2FIND service.

The data workflow in HERBADROP is represented by Fig. 1.

2.2 Data curation scenarios for HERBADROP

One of the requirements from the HERBADROP communities was to implement specific use cases, such as identifying duplicates amongst specimens from the different museums. Another example of policy is long term preservation that involves a number of controls including file format verification and metadata quality. Amongst HERBADROP users, two partners of the community have proposed practical scenarios for data curation: Digitarium [14] and the Royal Botanic Garden of Edinburgh (RGBE).

Scenario proposed by Digitarium (Finland)

Digitarium [14] planned to use Optical Character Recognition (OCR) data to generate metadata based on the label information available for the herbarium specimen. Firstly, a Natural Language Processing based system could be used to do OCR quality check and extract relevant terms. Then metadata could be either automatically generated, or manually inserted through the transcription portal [15] but with the help of OCR data.

More general for EUDAT infrastructure services, Digitarium wanted to utilize and integrate them into the whole digitisation process of natural history biological collections. The data flow goes from the beginning of the digitisation process i.e. imaging, to storage, then to transcription and analysis, until accessing. This involves data storage, high-performance computing resources, and web services in EUDAT.

Firstly, the images from the imaging station can be transferred into EUDAT storage for long-term preservation instantly or in batch. After transferring, HPC can access the images and do OCR to extract label information to generate preliminary metadata. This metadata has to be associated with corresponding images. The data can be openly accessed. However, the access rights of data have to be set up for different purposes, such as endangered species protection.

Secondly, using HTTP APIs, the images and their metadata can be accessible from EUDAT by data-owner portals. Therefore, browsing and transcribing are available.

Updated metadata will be transferred back into the EUDAT B2SAFE service. Different versions of metadata have to be kept.

Thirdly, the metadata is indexed. Therefore, the data can be searched or filtered based on different terms for further scientific usages. HPC resources can be utilized also on the data for different researches.

Scenario proposed by RBGE (the Royal Botanic Garden of Edinburgh) in association with MNHN (Musée National d'Histoire Naturelle) – Paris

The core of the concept of HERBADROP is to harvest metadata from OCR analysis of the text that is a part of herbarium images. The choice has been to proceed to a full text analysis using a Lucene-based engine Elasticsearch[16]. The objective of this approach is to provide a powerful interface for further data curation as part of the preservation process (identifying duplicates, or inducing new taxonomic relations, etc.), see [12].

Safeguarding long-term data storage is an important precondition for reliable access to herbarium specimen information. Thanks to this pilot, it is possible to envisage long-term storage for herbarium specimen images. Moreover, the specimens will be discoverable by the entire scientific community. Thus, undescribed species stored in herbaria can be examined by experts to aid identification and discovery of new species.

Distribution information for species over time can be evaluated and these data could provide evidence of the point in time when an invasive species first occurred in a certain area. Historians could analyse herbarium data to create itineraries for historical characters. The data can be used to calibrate predictive models of the oncoming changes in biodiversity patterns under global threats. This diverse information will be useful for a wide user community including conservationists, policy makers, and politicians.

3 GEOFON use case

The second use case concerns GFZ, the German Research Centre for Geosciences. GFZ provides valuable seismological services in the form of a seismological infrastructure named GEOFON [7].

GFZ is one of the members of the EPOS initiative (European Plate Observatory System) [5] and, in this context, collaborates with other two seismological data centres related to EPOS (KNMI, INGV) in the EUDAT project.

Besides being one of the fastest earthquake information provider worldwide, GEOFON is also one of the largest nodes of the European Integrated Data Archive (EIDA) for seismological data under the ORFEUS umbrella [23], which is a distributed data centre established to (a) securely archive seismic waveform data and related metadata, gathered by European research infrastructures, and (b) provide transparent access to the archives by the geosciences research communities.

The internal structure of GEOFON is based on three pillars:

- A global seismic network operated in close collaboration with many partner institutions with focus on EuroMed and Indian Ocean regions. The network consists of ca. 110 high quality stations, which acquire data in real time [8].
- A global earthquake monitoring system which uses data from GEOFON and partner networks [10]. It publishes most timely earthquake information. First automatic solutions are available few minutes after the events and mostly manually revised later.
- A comprehensive seismological data archive for GFZ and partner networks, for permanent networks as well as for temporary deployments.

For some GEOFON partner networks, GEOFON acts as a data centre saving a replica of the original copy and at the same time as a data distribution centre. Additionally, data from many temporary station deployments are permanently archived at GEOFON, in particular passive seismological experiments of the GFZ Geophysical Instrument Pool Potsdam (GIPP) and the German Task Force Earthquake.

Most of data is open for public access, as well as real-time data feeds when available. However, there is a small amount of data under an embargo period, usually for a limited amount of time (3-4 years).

3.1 Data workflow in GEOFON

GEOFON supports two scenarios for the ingestion of data into its archive: one for permanent networks and one for temporary (and most probably already finished) experiments.

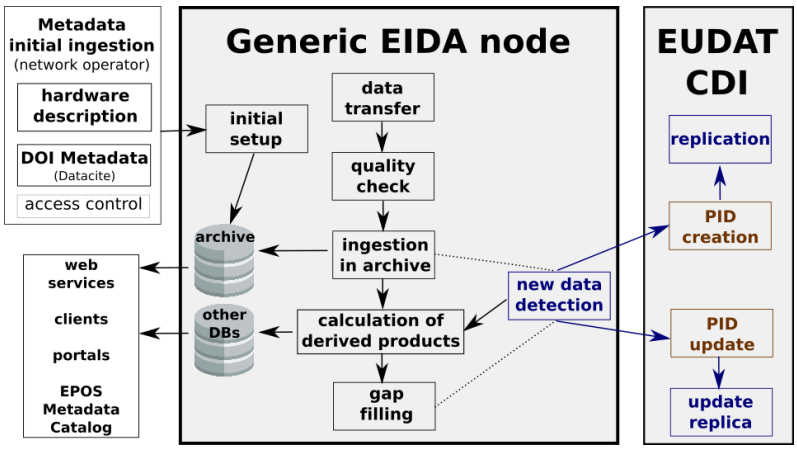

Fig. 2. Data workflow at GEOFON. It also represents the workflow from a generic seismological data centre as the ones under the EIDA/ORFEUS initiative. Boxes in black are generic activities from the data centre. Blue boxes show activities related to the EUDAT service B2SAFE, while brown boxes show the tasks related to B2HANDLE.

Usually, raw data are transmitted to the data centre with the metadata (technical hardware description) to be able to operate with them. In the case of permanent net-

works raw data are received continuously from the stations around the world via satellite using a protocol called SeedLink [17], a real-time data acquisition protocol which works on TCP. The packets of each individual station are always transferred in timely (FIFO) order.

In the case of temporary experiments network operators provide usually, first, the metadata needed to use the data, and in a second phase the data to be archived. Data transmission can be done as in the permanent networks case (SeedLink protocol), or can also be transmitted to the data centre by the network operator using some client-server tools provided by GEOFON, which will automatically do the first quality check of the data format. In some cases, both methods could be used.

A schematic view of the workflow at GEOFON can be seen in Fig. 2. It should be noted that this workflow is also very similar to the ones at other seismological data centres belonging to EIDA/ORFEUS. For instance, the other two data centres piloting EUDAT services (KNMI and INGV).

3.2 Service hosting environment with the inclusion of EUDAT services

Considering the workflow depicted in the previous section, GEOFON introduced some EUDAT services in order to automate and/or improve some of the tasks related to it.

Many services are being provided at GEOFON (e.g. interactive web portals, proprietary protocols to get data or derived products), with two of them (Station-WS and Dataselect) being particularly important, as they are international standards and the core services for the community upon which other services are built. Station-WS serves the information describing the hardware and everything related to the deployment, while Dataselect serves the data.

Two main EUDAT services have been integrated in the GEOFON workflow; namely, B2SAFE and B2HANDLE. The former is used to accomplish many of the Data Management tasks, while the latter is used to manage/store Persistent Identifiers (PIDs).

As the archive is stored on a partition with a particular directory structure, the B2SAFE service “mounts” the archive as an external resource in read-only mode.
One of the main requirements for the Data Policies at GEOFON was the capability to trigger processes based on the inclusion of new data. In the context of B2SAFE, this can be done by means of automatic rules which are executed under certain conditions (e.g. new data ingested).

With the proper rules we can enforce that, after new data is detected by B2SAFE, a certain set of actions is executed. For instance, the derived products can be generated and data can be replicated to a partner data centre from the EUDAT CDI, the Karlsruhe Institute of Technology (KIT). Also, as part of this replication process, persistent identifiers (PIDs) are generated for each file, so that the PID can be used to globally and univocally identify the file.

PIDs are managed and stored by means of the already mentioned service called B2HANDLE, which is based on a Handle Server and other libraries developed within the project. GFZ has a broad expertise on this type of tools and, therefore, we decided

to deploy our own B2HANDLE server and work with our local instance. Each generated PID is stored with a set of key-value pairs called "PID Record". The information in the PID Record allows, among other things, to track other copies of the file in different data centres or validate its integrity by means of pre-calculated checksums.

3.3 Data Policies to apply at GEOFON through EUDAT services

After the formalization of the internal workflows at GEOFON, and the inclusion of requirements from the community and the data centre, we defined a set of Data Policies to be enforced by means of the tools available within EUDAT and new developments, which could be useful for different communities.

Some of them are related to the Replication process. For instance:

- Replicate every new file in the archive to our internal backup server.
- If we are the official provider of the data in a file, replicate it to an off-site partner within the EUDAT CDI.
- Seismological data that do not belong to us but comes from our earthquake early monitoring system should be kept for 6 months only. Data still need to be replicated to the internal server.
- An automated file deletion must not be possible. In case that the system detects that a file should be deleted, an email should be sent to the appropriate operator.

Regarding the access control of the files:

- "Restricted data" must be tagged, with proper access control applied to them.
- Access restrictions can be automatically removed after a period of time (embargo period).
- Data must be able to be accessed via an HTTP API respecting the ACL (Access Control List).

Regarding automatic metadata extraction:

- Metrics derived from the data must be automatically calculated to populate some of our services when new data is ingested.
- Detailed statistics related to the data access should be available for the data owners/creators.
- In case that data are modified (e.g. correcting errors, filling gaps), this information should be available for future use (provenance information) (see section 5).

Regarding the integrity of the stored data:

- A weekly process will select ~2% of the folders in our archive and verify that the synchronization is correct. The idea is that every file will be checked at least once in a year.
- Check that the data are stored in SDS format (SeisComP Directory Structure).
- Start and end time of network/station operation must be available and data outside

this timespan must not be allowed.

The identified relevant policies were gradually implemented using generic EUDAT services and GEOFON-specific software.

4 Mapping of EUDAT data policies to SCAPE and RDA policy curation frameworks

For the design and implementation of data curation actions in EUDAT, the relevant outputs of SCAPE project [2] and Practical Policy Working Group of the Research Data Alliance [3] have been identified. SCAPE outputs are perceived of high quality owing to the advanced thinking that considered long-term digital preservation policies at a granular level suitable for the machine-executable implementation. RDA Practical Policy Working Group outputs are a result of a substantial international collaborative effort including experts in iRODS platform [11], which is the technological foundation of the EUDAT B2SAFE service.

For SCAPE, we used the catalogue of preservation policy elements [4]. This is a systematized compendium of granular policies with examples of what SCAPE called "control policies" (granular statements that are easily translatable to machine-executable functions), and for the RDA Practical Policy Working Group it was their practical policy implementations report [9] that compiled a set of machine-executable functions for iRODS platform [11].

In addition to this top-down retrospective review of the SCAPE and RDA outputs, a bottom-up analysis of control policies applicable to the GEOFON and HERBADROP use case was performed, with a number of control policies identified as prime candidates for implementation in EUDAT B2SAFE. These policies are presented in Table 1.

Table 1. Candidate control policies for implementation by GEOFON and HERBADROP.

Policy category	Control policy	Policy examples
Data replication	Number and location of replicas	Data should be replicated in N locations, including in locations A and B
	Timeframe for replication	Data should be replicated within the next 24 hours after the data ingestion in any particular location
	Data nodes roles	All data nodes are equivalent to read data from, but data can only be initially ingested in node X then replicated over all other nodes

Data integrity checks	The set of checksum algorithms acceptable	Checksum algorithm accepted is MD5
	Periodicity and scope of integrity checks	Calculate checksums for 2% of all data assets every week, with the aim of having the entire data collection checked annually
Data and metadata formats	Data formats accepted	BMP and PNG accepted for images
	Metadata extraction from data	Upon ingestion, file name should be extracted as metadata
	Data format check procedures acceptable	Software package X should be used for data format validation
	Minimal metadata assigned upon data release	PID is a mandatory metadata element
Data access and data reuse	Embargo rules	Embargo period of N years is applied to all PDFs and images
	The set of data licenses recommended upon data release	CC-BY license should be assigned to all data released after the embargo period ends
	Data reuse statistics collection	Number of file downloads should be collected

Then the gap analysis was performed against SCAPE policy elements, to see whether these bottom-up identified control policies allow enough coverage of the extensively defined data curation policy landscape of SCAPE project. SCAPE policy elements catalogue [4] is two-level with Guidance Policies on the top level and Policy Elements on the granular level. An example of Guidance Policy is Authenticity Policy that breaks down to Integrity, Reliability and Provenance as policy elements. Hence control policies in Data Integrity checks category from Table 1 correspond to Integrity policy element of Authenticity Policy in the SCAPE policy elements catalogue.

One noticeable gap discovered through this mapping exercise is the Digital Object lifecycle which was paid due attention to in SCAPE policy landscape but is missing in the current EUDAT considerations. This gap may be hard to address as EUDAT is a

collaborative project that accumulates data from a large variety of research communities with a wide range of digital object types and lifecycles. However, this discovery should inform the future operation of EUDAT services so that they could meet all reasonable (and multi-aspect) requirements for data curation.

5 Provgen: a Web service to generate provenance data

Depending on the conceptual point of view, supplying data with enough provenance information can be seen as a specific data policy or as a track of evidence that specific data policies have been actually implemented. There is a particular challenge for sensible data provenance in data infrastructures, owing to the scale and diversity of data sources, data processing agents (services), and a variety of services implementation.

As the EUDAT services landscape becomes more complete, and also complex, much more information is generated from the execution of data workflows by different actors. Based on this, we designed and implemented a "Provenance generation service" called Provgen, which could act as a central point to generate and collect all the information generated not only from the EUDAT CDI services, but also from other external services which take part in the data workflow.

One of the main requirements of this service was to be capable of being used on almost any reasonable data workflow which could be designed and contacted by any service. Therefore, it had to be highly configurable to the different needs of the users, strongly decoupled from the other products, and with a minimum set of software dependencies.

To fulfil the requirement of flexibility to adapt to different data workflows, we designed a templating system, where templates can be loaded by the operator of the system. Templates are in Notation3 format and each template is the result of the design of a certain Provenance record type generated at a particular point in the workflow. For instance, a template for the creation of a Persistent Identifier (PID) to identify a data file can be expressed as the RDF shown in Fig. 3.

To decouple Provgen from other components of the EUDAT CDI and avoid software dependencies, we decided to specify and expose an API [22] which offers the following functionality over HTTP calls:

- see available templates in Provgen including the specification of each template (expected variables and their description),
- instantiate a template to generate provenance information,
- show the documentation of the system,
- see details on how Provgen was configured.

Given any particular data flow, a user of Provgen needs to do two tasks to configure it before it is available for use:

- identify the point(s) at which some Provenance information should be generated. For each of these points, a template must be generated with all the needed attributes and copied to the "templates" folder into the system,

- include in the code related to these points of the workflow a call to the Provgen API specifying the template and instantiating all the required variables.

Dublin Core and PROV [24] elements are used for the expression of RDF template, with an addition of a Provgen-specific element for the expression of persistent identifiers.

When a client needs to generate a record according to the template, it calls the API specifying the template and a set of key-value pairs. If all variables are specified and the types are correct, the user will get a Provenance record as a response to the HTTP call. Optionally, the provenance record is generated and stored on the server in a triple-store backend.

With this setup, all current and future EUDAT services (as well as external consumers) will be able to use Provgen in order to generate provenance records and provide users with a unified view of provenance information across different services in EUDAT and beyond.

More details and a longer description on Provgen component will be available in [21].

6 Implementing data versioning policy in EUDAT B2SAFE

Versioning is essential for proper data curation and provenance. Software version control git-like systems are not able to deal efficiently with large size binary datasets that are one of the focuses of EUDAT B2SAFE – robust, safe and highly available service for storing large-scale data in community and departmental repositories. This is why a new service-specific versioning functionality has been designed and implemented, which can be considered a working example of an executable data policy, and a role model for other executable data policies.

B2SAFE is based on the iRODS middleware [11] and is inherently capable to operate on very large datasets distributed across multiple domains. Research communities employ B2SAFE service primarily to replicate datasets across different data centres. Information about locations of replica and their other properties is maintained and queried using Persistent Identifiers (PIDs). The so-called "PID record" keeps information as a set of key-value pairs and can store, for example, checksum, timestamp, location, etc. of a given dataset. Internally, the Handle system [26] manages PIDs in the B2SAFE system.

The versioning code is written as a custom iRODS rulebase using iRODS Rule language and is available at GitHub as a part of B2SAFE service core code [27]. In our implementation of versioning functionality for B2SAFE service, a version of digital object is understood as a timestamped copy of the object. Versions are kept in a separate directory created only for storage of versions.

As illustrated by Fig.3, PID record of each version contains a URL of the version and a few other fields:

- EUDAT/DO_VERSION_NUMBER field keeps the version's number;

- EUDAT/FIXED_CONTENT field set to “true” meaning that content modification of the version is not allowed;
- EUDAT/CHECKSUM field containing checksum of the version file to be used for integrity validation;
- EUDAT/PROFILE_VERSION is set to one - as a legacy EUDAT value;
- EUDAT/TIMESTAMP contains a timestamp when the version has been created;
- EUDAT/FIO is a reference to the First Ingested Object (FIO) stored in the EUDAT/FIO field;
- EUDAT/VERSION_SOURCE contains a reference to direct source file which version has been created;
- EUDAT/REVISION_OF and the EUDAT/WAS_DERIVED_FROM fields refer to the previous and next version, respectively; a reference to the Repository of Records (ROR) stored in the
- EUDAT/ROR that refers to the original of the version in the Repository of Records (ROR).

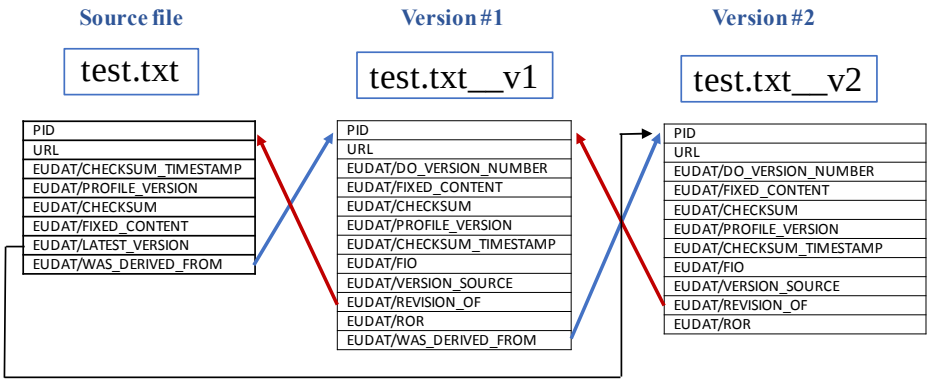

Fig. 3. B2SAFE versions cross-linking.

An issue of practical importance is how to retrieve an older (previous) and a newer (next) version with respect to a current version. In the B2SAFE versioning service, for navigation between versions, references to older and newer versions are kept in the fields called, according to the semantics of PROV-O ontology [24], EUDAT/REVISION_OF and EUDAT/WAS_DERIVED_FROM, respectively. Hence, the new version contains a reference to the previous version and vice versa, i.e., versions are cross-linked for ease of traversing through them (as in Fig.3). Moreover, for convenience, PID record of the source file has a reference to its latest version.

The described versioning system can create versions not only of a single data object, but also recursively versions of every data object located within a collection (optionally including sub-collections). The versioning functionality can also be applied to a file or collection of files which do not have registered PID. Furthermore, maximum number of versions of a file can be set to keep disk usage under control. By default, the B2SAFE versioning service retrieves the latest written version of a file

from a specified versions repository to a given directory. A version number should be specified if an older version of the file is required.

7 Conclusion and further work

Analysis of data curation requirements of two use cases: HERBADROP and GEOFON has been performed, coupled with the retrospective review of the elaborated data curation policies from a dedicated EU project (SCAPE) and practical (machine-executable) policies that were the output of the dedicated RDA working group. A set of granular control policies have been identified and implemented in two use cases, and a gap analysis of these policies has been performed against the SCAPE catalogue of policy elements.

The conceptual framework developed and requirements analyzed supported the actual implementation of data versioning policy in EUDAT B2SAFE service and the development of Provgen component [21] for generation of data provenance records.

After the set of identified policies was applied in the two use cases that have been involved in their formulation, the same policy framework could be applied in a larger number of research communities associated with EUDAT.

The scope of projects and initiatives in data curation and long-term digital preservation can be extended beyond SCAPE and RDA working groups; this specifically applies to popular functional models of digital preservation like OAIS [18] that have not been thoroughly evaluated so far for their potential application in EUDAT.

Sensible modelling of data policies and, specifically, overcoming conceptual and technological gaps between textual policy formulation and machine-executable policies require further considerations and experiments. One possible approach could be modular policy modelling supported by semantic Web technology [28].

Should the practical project opportunity arise, the body of knowledge acquired through the outlined EUDAT data curation and data provenance task can be applied to the existing or emerging data infrastructures in order to make them “data curation-centric”, in opposition to the current most popular approach when data services are implemented first, with data curation requirements and implementation considered second.

For EUDAT Collaborative Data Infrastructure, this approach can be considered as addressing two major challenges:

- making a fully-fledged long-term digital preservation solution out of EUDAT B2SAFE service that looks technologically sound and supported by a distributed network of research partners with a certain level of the IT governance culture, yet lacks consistent data curation vision and enough implementation of data policies;
- replacement of “vertical” or “silos” data curation concepts, requirements and technology that are specific to particular data pilots with “horizontal” concepts, requirements and technology that will be universal across the entire data infrastructure traversing and supporting a variety of data pilots.

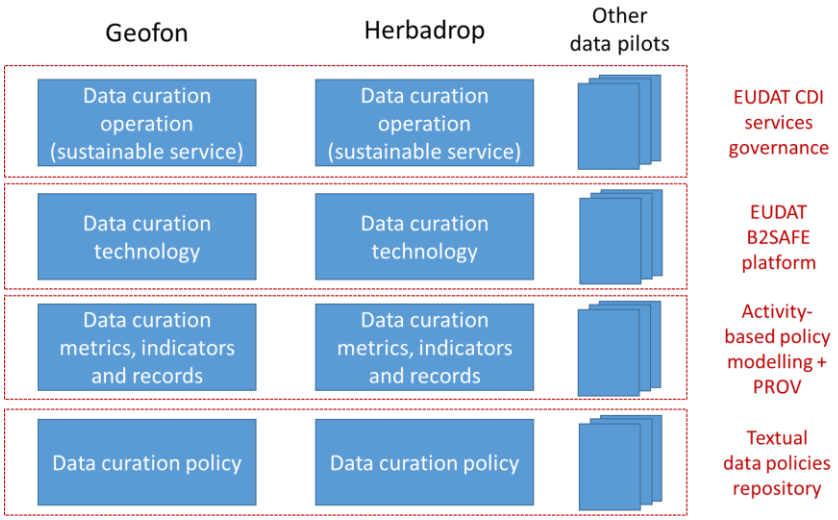

Fig. 4. Data curation-centric vision of the EUDAT Collaborative Data Infrastructure.

This vision is represented by Fig. 4 with the layers that correspond to different aspects of curation-centric data infrastructure. The right column suggests "providers" for each layer, e.g. for the operational governance, organizational structures of EUDAT CDI [1] could be used. This vision will require substantial organizational effort and sustainable sources of funding in order to be delivered and supported in the long term.

Acknowledgements

This work is supported by EUDAT 2020 project that receives funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No. 654065. The views expressed are those of authors and not necessarily of the project.

References

1. EUDAT Collaborative Data Infrastructure. https://www.eudat.eu/eudat-cdi
2. SCAPE: Scalable Preservation Environments. http://scape-project.eu/
3. Research Data Alliance Practical Policy Working Group. https://www.rd-alliance.org/groups/practical-policy-wg.html
4. SCAPE Catalogue of Preservation Policy Elements. http://scape-project.eu/wp-content/uploads/2014/02/SCAPE_D13.2_KB_V1.0.pdf
5. EPOS: European Plates Observing System. https://www.epos-ip.org/
6. CINES: French national IT center for higher education and research. https://www.cines.fr/en/

7. Hanka, W., Kind, R. The GEOFON Program. Annals of Geophysics v. 37, n. 5, Nov. 1994. ISSN 2037-416X. doi:10.4401/ag-4196 (1994)
8. GEOFON Data Centre (1993): GEOFON Seismic Network. Deutsches GeoForschungsZentrum GFZ. Other/Seismic Network. doi:10.14470/TR560404 (1993)
9. Practical Policy Implementations Report. http://dx.doi.org/10.15497/83E1B3F9-7E17-484A-A466-B3E5775121CC
10. Hanka, W., Saul, J., Weber, B., Becker, J., Harjadi, P., Fauzi, and GITEWS Seismology Group. Real-time earthquake monitoring for tsunami warning in the Indian Ocean and beyond. Nat. Hazards Earth Syst. Sci., 10, 2611-2622, doi:10.5194/nhess-10-2611-2010 (2010)
11. iRODS: Integrated Rule-Oriented Data System. https://irods.org/
12. Haston, E., Chagnoux, S., Dugénie, P. Herbadrop – Long-term preservation of herbarium specimen images. Proceedings of the second Eudat User Forum. Rome (2016).
13. Dugénie, P., Chagnoux, S. EUDAT Data pilot Herbadrop. Second interim Herbadrop Data Pilot report (2016)
14. Digitarium: service centre for high performance digitization. http://digitarium.fi/en
15. DigiWeb+ digitization platform. http://digiweb.digitarium.fi/
16. Elasticsearch search and analytics engine. https://www.elastic.co
17. SeedLink protocol and tools overview. http://ds.iris.edu/ds/nodes/dmc/services/seedlink/
18. Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book). Issue 2, June 2012. CCSDS (The Consultative Committee for Space Data Systems), Washington DC (2012)
19. EUDAT services. https://www.eudat.eu/services-support
20. Bunakov, V. et al. Data curation policies for EUDAT collaborative data infrastructure. In "Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017)". CEUR Workshop Proceedings Vol-2022, urn:nbn:de:0074-2022-6, 72-78 (2017)
21. Bunakov, V., Quinteros, J., Reijnhoudt, L. Data Provenance Service Prototype for Collaborative Data Infrastructure. Submitted in ALLDATA 2018: The Fourth International Conference on Big Data, Small Data, Linked Data and Open Data. http://www.iaria.org/conferences2018/ALLDATA18.html
22. Provgen API specification. https://raw.githubusercontent.com/javiquinte/provgen/master/swagger.yaml
23. ORFEUS: Observatories and Research Facilities for European Seismology. http://www.orfeus-eu.org/
24. PROV ontology. https://www.w3.org/TR/prov-o/
25. EUDAT B2SAFE service. https://www.eudat.eu/b2safe/
26. Handle.Net registry. http://handle.net
27. B2SAFE service core code. https://github.com/EUDAT-B2SAFE/B2SAFE-core/tree/versioning
28. Bunakov, V. Data policy as activity network. In "Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017)". CEUR Workshop Proceedings Vol-2022, urn:nbn:de:0074-2022-6, 79-86 (2017)