Data Curation Policies for EUDAT Collaborative Data Infrastructure

© Vasily Bunakov[1], © Alexia de Casanove[2], © Pascal Dugénie[2], © Rene van Horik[3], © Simon Lambert[1], © Javier Quinteros[4], © Linda Reijnhoudt[3]
[1] Science and Technology Facilities Council, Harwell Campus, United Kingdom
[2] CINES, Montpellier, France
[3] Data Archiving and Networked Services (DANS), The Hague, Netherlands
[4] GFZ German Research Centre for Geoscience, Potsdam, Germany
vasily.bunakov@stfc.ac.uk, casanove@cines.fr, dugenie@cines.fr, rene.van.horik@dans.knaw.nl, simon.lambert@stfc.ac.uk, javier@gfz-potsdam.de, linda.reijnhoudt@dans.knaw.nl

Abstract. The work outlines an approach to the development of a data curation framework in the EUDAT Collaborative Data Infrastructure. Practical use cases are described as well as provisional results of defining granular data curation policies with high potential for their machine-executable implementation.
Keywords: data curation, 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 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. The major EUDAT services [19] are:

- B2ACCESS – identity and authorization service;
- B2HANDLE – service for assigning and managing persistent identifiers;
- B2DROP – service for secure and trusted data exchange;
- B2SHARE – service for sharing small-scale "long tail" data;
- B2SAFE – robust, safe and highly available service for storing large-scale data in community and departmental repositories;
- B2STAGE – service for managing data transfers between EUDAT storage and high-performance computing;
- B2FIND – service for data discovery across the EUDAT infrastructure (data catalogue).

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.

Historically, EUDAT services have been built with only a few considerations for conscious data curation, with secure and controlled access to data being one of the major initial goals to achieve. Other aspects of data curation 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 B2DROP, B2SHARE and B2FIND 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.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

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.

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, describes mapping of SCAPE policy elements [4] to granular data policies in EUDAT, and sets directions for further works on data policies in EUDAT.

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 metadata 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, both supported by CINES [6], by using Optical Character Recognition (OCR) analysis.

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].

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, five European community partners[1] have been involved. Their contribution to the pilot represents a business model that can be potentially replicated by other institutes.

The EUDAT B2SAFE service is used in the first step of the ingestion process. Existing images of herbarium specimens along with the associated metadata 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 Figure 1.

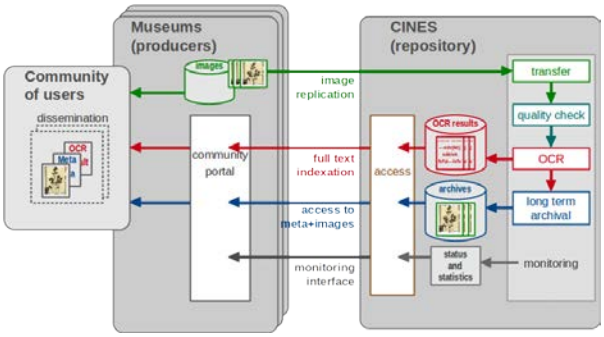

Figure 1 Data workflow of the HERBADROP data pilot

2.2 Data curation scenarios for HERBADROP

The HERBADROP communities have expressed their wish to implement specific use cases such as identifying duplicates amongst specimens from the different museums. This kind of requirement is very useful to improve EUDAT services. 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] would like 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 will be used to do OCR quality check and extract relevant terms. Then metadata will 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 would like to utilize and integrate them into the whole digitisation process of natural history biological collections. The data flow goes from the beginning

[1]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

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] to research and better understand our complex system Earth.

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[2] umbrella, 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 [9]. 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 data are 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.

Usually, raw data is transmitted to the data centre with the metadata (technical hardware description) to be able to operate with it. In the case of permanent networks raw data is received continuously from the stations around the world via satellite using a protocol

[2] Observatories and Research Facilities for European Seismology (http://www.orfeus-eu.org/)

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 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 valid for many of the seismological data centres belonging to EIDA/ORFEUS. For instance, the other two data centres piloting EUDAT services (KNMI and INGV).

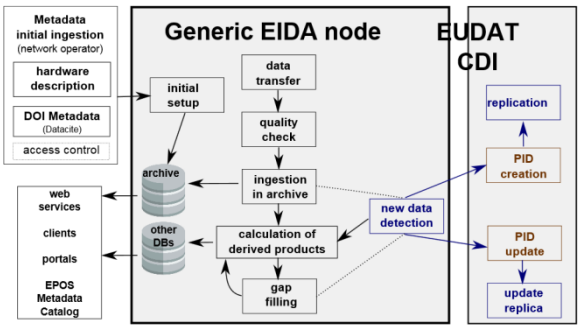

Figure 2 Data workflow from GEOFON. It also represents the workflows from a generic seismological data centre as the ones under the EIDA/ORFEUS initiative. Boxes in black are generic activities from the data centre. Blue boxes show activities related to the EUDAT service B2SAFE, while brown boxes show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

In both cases, permanent and temporary networks, data go through some quality checks after being received. When data are sent in real-time there is a first control by sorting the records before actually ingesting them into the archive (~1 day after reception). After 4–6 weeks, for stations that still have the buffered data, a gap filling process is started.

When data have been bulk uploaded to the data centre by the network operator, it is immediately checked to exclude overlaps. In this case, as all available data is copied off-line, there is no need to check for problems related to real-time transmission, like gaps and proper order of records, as they are checked by the automatic archiving tools.

In the case that the data is under an embargo period, the access control list is created or updated. After completion of the last steps, data is opened through standard access protocols.

The internal organization of the archive is based on an approach called SeisComP3 Data Structure (SDS). This means that files are stored under a predefined directory structure, which uses the codes from the network/station/channel used to record the data as well as the year. The continuous time series are stored in a standard seismological format called Mini-SEED. The time series are split in daily files for each recording sensor and, therefore, files are closed when the day finishes. At that moment, "new" data (recently closed files) can be processed to obtain derived products from them. For instance, quality metrics on the data or detailed availability information, which are offered to our users by means of a Web service.

Once the data is archived users can make use of any of the services provided by GEOFON to retrieve it. Considering that there are different services which can provide the data to the users, the usage statistics is centralized in one database to be able to analyse the impact of the data on the community regardless of the method used to retrieve it.

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 most of the Data Management tasks, while the latter is used to manage/store Persistent Identifiers (PIDs).

As the archive is stored in a directory structure from a partition, 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 is 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 in 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 does 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;
- file deletion must not be possible in an automated way. 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 and proper access control must be 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).

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 is stored in SDS format;
- start and end time of network/station operation must be available and data outside this time span must not be allowed.

The identified relevant policies are being 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] that is a technological foundation of the EUDAT B2SAFE service.

For SCAPE, we used the catalogue of preservation policy elements [4] that 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.

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 and long-term digital preservation.

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

5 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 as candidates for implementation in two use cases, and a gap analysis of these policies has been performed against the SCAPE catalogue of policy elements. A similar gap analysis should be performed against the RDA practical policies catalogue, in order to see what existing iRODS implementations can be reused for the creation of machine-executable policies in EUDAT B2SAFE service.

After the set of identified policies is applied in the two use cases that have been involved in their formulation, the same policy framework should be applied in a larger number of research communities associated with EUDAT through its pilot programme.

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 we feel have not been thoroughly evaluated so far for their potential application in EUDAT.

The major result of these works is going to be a conceptually and terminologically consistent catalogue of machine-executable policies for EUDAT services that will be explicitly mapped to requirements of the participating research communities, as well as to mature data policy frameworks developed by EU projects and international collaborations dedicated to data curation and long-term digital preservation.

The EUDAT data policies catalogue will serve then both as guidance for machine-executable policy implementations and as a validation tool to ensure the compliance of EUDAT CDI services to high-level policies of data curation and long-term digital preservation. This should allow to promote certain EUDAT platforms such as B2SAFE from their current status of “bit-level” data management solutions topolicy-driven services where the actual set of policies can be configured according to a particular use case.

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, 37 (5), Nov. 1994. ISSN 2037-416X. doi:10.4401/ag-4196

[8] GEOFON Data Centre (1993): GEOFON Seismic Network. Deutsches GeoForschungsZentrum GFZ. Other/Seismic Network. doi: 10.14470/TR560404

[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, pp. 2611-2622 (2010). doi:10.5194/nhess-10-2611-2010

[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. Proc. of the second Eudat User Forum. Rome (2016). https://www.eudat.eu/communities/ long-term-preservation-of-herbarium-specimen-images

[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). EUDAT services. https://www.eudat. eu/services-support

[19] EUDAT services. https://www.eudat.eu/services-support