End-to-end research data management workflows:
A case study with Dendro and EUDAT

Fábio Silva, Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro

INESC TEC—Faculdade de Engenharia da Universidade do Porto
{ffjs1993,ricardo.amorim3,joaoaguiarcastro,joaorosilva}@gmail.com,
mcr@fe.up.pt

Abstract. Depositing and sharing research data is at the core of open science practices. However, institutions in the long tail of science are struggling to properly manage large amounts of data. Support for research data management is still fragile, and most existing solutions adopt generic metadata schemas for data description. These might be unable to capture the production contexts of many datasets, making them harder to interpret. EUDAT is a large ongoing EU-funded project that aims to provide a platform to help researchers manage their datasets and share them when they are ready to be published. DataPublication@U.Porto is an EUDAT Data Pilot proposing the integration between Dendro, a prototype research data management platform, and the EUDAT B2Share module. The goal is to offer researchers a streamlined workflow: they organize and describe their data in Dendro as soon as they are available, and decide when to deposit in a data repository. Dendro integrates with the API of B2Share, automatically filling the standard metadata descriptors and complementing the data package with additional files for domain-specific descriptors. Our integration offers researchers a simple but complete workflow, from data preparation and description to data deposit.

1 Introduction

An unprecedented growth in data production is compelling institutions to implement infrastructures to make these resources available in the long run [6], while funding institutions require projects to make data available as specified in Data Management Plans, which are becoming mandatory. The challenges range from enabling researchers to deposit and describe their data early in the research projects, to ensuring the long term preservation of project results upon their completion.

Although the expertise in managing publication records can be seen as a starting point when designing applications for data management, recent studies reveal that adapting the existing tools to the new requirements often yields limited capabilities that may render the infrastructure unfit for research data management (RDM) [2].

To promote interoperability, RDM platforms are often compliant with metadata exchange protocols and offer interfaces to enable the integration with other platforms. This is the case with Dendro, a platform to assist researchers in the organisation and

description of their datasets. Dendro can export the prepared datasets to any institutional repository, ideally one that can leverage domain-specific metadata to improve data visibility and increase the potential for reuse of the datasets [7].

EUDAT[1] is an European initiative that aims to create a centralized solution for data management in several research settings, ranging from the publication of datasets in the long tail to storing and delivering large datasets in specialised high-performance environments. EUDAT offers modules for data management and communication, along with a comprehensive API to simplify integration with established infrastructures.

This work is focused on the long tail of science [5], where small research groups from diverse domains need straightforward processes for data deposit and long-term preservation. We describe the integration of Dendro with the EUDAT e-infrastructure to compose an RDM workflow that can be easily integrated into the regular research processes.

2 Data Management for Reuse

Along with the increasing open-access demands, research institutions can benefit from timely disclosure of their outputs. As with research papers, published research data can be cited and provide credit to their authors. Moreover, by enabling other researchers to reuse data, institutions contribute to research transparency and increase their own visibility. Data reuse implies that the researcher can fully grasp both the origin and the context of production for the dataset [3].

A workflow that covers the entire data lifecycle is therefore required, to couple data and metadata from the start and provide a clear record of the data production process. The initial stages are often characterized by datasets being created and updated, making flexible staging platforms ideal to manage such resources. Complementary tools such as electronic laboratory notebooks have also shown promising results in motivating researchers to actively describe their data [1].

On the final stages of the workflow—deposit and dissemination—there is concern with the existence of sufficient metadata, so that the dataset can be located, interpreted and reused. The main issues are often related to capturing the context of production of the datasets, which often means dealing with multiple metadata schemas, while ensuring the compatibility with domain-level metadata, a problem that is often undervalued in emerging platforms [2].

In some research areas infrastructures are already in place, with well established data sharing guidelines, often specified in Data Management Plans. Repositories for these scenarios are commonly tailored to existing local needs, and often rely on datasets that follow a consistent structure across diverse projects. This is not common in institutions that deal with several projects at the same time and have to cope with heterogeneous datasets as well.

Widely used platforms for institutional repositories—such as DSpace and ePrints—have been adapted to handle research data. This solution satisfies the data access requirements, as the existing dissemination protocols, namely the OAI-PMH, are natively supported. The main issue with data publication is description. In large disciplinary repositories, the task can be committed to a curator with expertise in the domain. This approach is not viable for repositories in the long tail, dealing with many domains. Here,

[1] `https://www.eudat.eu/`

the bulk of the description task has to be assigned to researchers and data creators. To add this to their regular activities, new tools and workflows are required.

Data repository platforms such as Figshare[2] and Zenodo[3] have come to offer simpler, yet extensive interfaces to allow data deposits to be completed by the researchers [2]. Involving researchers in the management of their data is a wise step, taking advantage of their knowledge on the domain to generate accurate description for the data.

3 Dendro and EUDAT

Providing researchers with data management tools for the whole research process is expected to improve the quality of their data, to generate more and more specific metadata, and to make more datasets reach the publication stage. In the proposed workflow, Dendro contributes to the data organisation and description components, while EUDAT is used as the publication platform.

3.1 Dendro

Researchers need data management tools early in the research workflow, namely to capture domain-level metadata. In some cases, data description is already a part of the research routine, sometimes including laboratory notebooks as means of personal organization. The laboratory notebooks hold valuable metadata records which are expensive to produce; they serve as inspiration for more efficient tools to capture metadata in increasingly digital workflows and processes.

Projects also involve teams of several researchers; digital platforms can provide researchers with collaborative environments where they can represent their domain-specific metadata into structured and standards-compliant records. Dendro[4] focuses on creating comprehensive descriptions with domain-specific terms and providing collaborative features. It supports researchers on their data management routine [7], and can be seen as an extension of their workspace. When the project comes to an end—or anytime the researchers choose—Dendro exports the final package, containing both data and metadata, to almost every data repository.

Dendro has a data model based on ontologies that can be regarded as conceptual representations of domains and may group descriptors from several metadata schemas. The ontologies are built in collaboration with the research teams, assessing their description needs and capturing the domain terminology that will enable their peers to interpret the datasets [4].

3.2 EUDAT

EUDAT proposes an integrated environment that addresses several requirements of researchers with respect to data processing, description and deposit. It is an array of platforms, including modules for data processing and refinement, data preservation, collaboration and dissemination. The services are offered in compliance with European

[2] `https://figshare.com/`
[3] `http://zenodo.org/`
[4] `http://dendro.fe.up.pt/demo`

guidelines on open access and research data disclosure. These capabilities make EUDAT a strong candidate for institutions that need to provide a Data Management Plan when applying for European research grants, but also for those that are looking for a platform for daily use by researchers. The existing modules are:

- **B2Drop**—stores and synchronizes data, providing collaborative tools using a Dropbox metaphor.
- **B2Share**—facilitates data deposit by researchers or institutions in some of the major domain repositories, e.g. CLARIN for linguistics or GBIF for biodiversity; some fields are required for deposit, and will be used as metadata; depending on the target repository, some more specific fields can be added; a unique identifier is assigned to the dataset.
- **B2Safe**—replicates research data; its features include policy rules, management of identifiers and integrity checking.
- **B2Stage**—offers computational resources to help researchers refine their data; it handles the exchange of data between EUDAT's storage resources and High-Performance Computing workspaces.
- **B2Find**—supports data discovery; using the OAI-PMH protocol, it gathers metadata from external repositories and B2Share, and exposes the results to users through a search interface.
- **B2Access**—handles federated authentication across all modules.

EUDAT is provided as a service, an approach that can reduce the impact of deploying a data management platform when compared to institution-supported solutions. The growing support community and a broad network of partners all over the EU contribute to the visibility of EUDAT in the data management landscape. The additional modules for large-scale storage and computing might also help institutions without sufficient funds to access such capabilities.

3.3 Integrating Dendro with EUDAT

Dendro allows researchers to download selected project folders and deposit the generated package in B2Share, automatically filling in the required metadata fields. The researcher immediately obtains a URL for the deposited dataset, which also includes a unique identifier that can be cited. The gathered metadata is pre-processed to filter descriptors that are recognized by the platform—such as `title`, or `description`. Dublin Core descriptors are exported through the existing API. The complete metadata record with all descriptors is exported as an RDF file that can later be ingested by other platforms to facilitate the interpretation and use of the dataset. At this stage, Dendro can export the results to the EUDAT platform in two ways:

- **Via B2Share**—through Dendro's interface, the researcher exports the project. The API of B2Share is used. The researcher chooses to deposit data in a personal account (by providing a personal authentication token) or in the default one.
- **Via B2Find**—Dendro exposes project metadata via an OAI-PMH server, with varying levels of access to data and metadata. This is done automatically by a script that gathers metadata from projects and XML files and exposes them to OAI-PMH harvesters.

The first approach is appropriate when the project data and metadata can be disclosed. This is often a decision of the project manager and ensures that datasets remain closed and are only exported to the EUDAT platform when ready. The second case covers scenarios where researchers cannot directly disclose their data, usually during the research project or when embargo periods are in place. Only the metadata is exposed and any external access requires authorization of the researchers. In this case, the dataset remains on the Dendro platform, and three levels of access control were implemented to address these constraints:

- **Private**—neither project data nor metadata is to be shared, addressing scenarios where the dataset contains sensitive or private data;
- **Public**—metadata is exposed via OAI-PMH protocol. The project's URL redirects users to a page that allows them to see the project structure and download it. This can be useful for projects in the public domain or containing institutional information that requires datasets to be visible to the community;
- **Metadata only**—metadata is exposed via OAI-PMH. However, the project's URL redirects to a page where the user can request access to the project. This level is used, for instance, when it is interesting for the researchers to reveal the project status, associated contacts and other metadata, but disclosure is postponed.

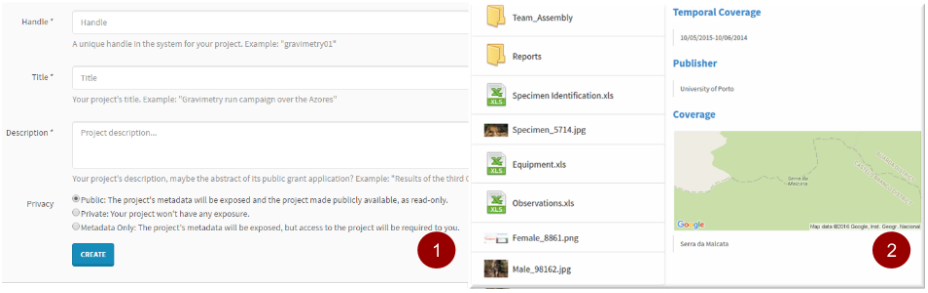

Fig. 1: **1**.Defining project visibility. **2**.Dataset in view-only mode.

The Dendro interface was adapted to accommodate the implemented features. In Figure 1, number **1** illustrates the creation of a project, where the researcher can choose a privacy level, which can be updated at any time. Number **2** has a view of a project configured as public, as it is systematically updated by the project team. Along with the data, external parties can access the associated metadata record.

4 Conclusions

This work proposes a workflow for research data, from creation to description and publication. Dendro was used for data organisation and description, and EUDAT for data publication. The integration of Dendro and EUDAT raised concerns with respect to metadata. The EUDAT API provides a set of Dublin Core descriptors, leaving out other domain metadata recorded in Dendro. To overcome this limitation we associated

the complementary descriptor values as a file in the dataset package. Although they do not contribute to search in EUDAT, the metadata record exported by Dendro is compliant with existing metadata schemas, and the ontologies used are also published. Metadata can therefore be used by more advanced systems or applications. The integration of Dendro with B2Share for deposit and B2Find for search gives researchers flexibility to disclose their data according to the permissions they have set[5].

The two integration paths manage to export data and metadata within the Dendro environment, using EUDAT, one of the current solutions for European public data repository. This experiment has responded to the pressing needs of researchers who need a simple process for data deposit and publication. Two lines of work are ongoing: the release of Dendro as open-source code and the test of a public data repository managed by the University of Porto. More experiments to test these solutions with researchers are required. Moreover, work in this line depends on the European infrastructures and on forthcoming national and European policies.

5 Acknowledgements

This work is financed by the ERDF—European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project POCI-01-0145-FEDER-016736.

References

1. Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. LabTablet: Semantic metadata collection on a multi-domain laboratory notebook. In *Metadata and Semantics Research Conference Proceedings*, volume 478. Springer, 2014.
2. Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, 2016.
3. Massimiliano Assante, Leonardo Candela, Donatella Castelli, and Alice Tani. Are scientific data repositories coping with research data publishing? *Data Science Journal*, 15, 2016.
4. João Castro, João Rocha, and Cristina Ribeiro. Creating lightweight ontologies for dataset description: Practical applications in a cross-domain research data management workflow. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 2014.
5. P. Brian Heidorn. Shedding light on the dark data in the long tail of science. *Library Trends*, 57(2):280–299, 2008.
6. Robin Rice and J Haywood. Research data management initiatives at University of Edinburgh. *International Journal of Digital Curation*, 6(2), 2011.
7. João Rocha da Silva, Cristina Ribeiro, and João Correia Lopes. The Dendro research data management platform: Applying ontologies to long-term preservation in a collaborative environment. In *iPRES Conference Proceedings*, 2014.

[5] An example of a deposited dataset: `b2share.eudat.eu/record/404`