Dendro: a FAIR, open-source data sharing platform

Lázaro Costa[0000−0002−0544−4618] and João Rocha da Silva[0000−0001−9659−6256]

INESC TEC / Faculdade de Engenharia da Universidade do Porto[⋆⋆]
lazaroosta@hotmail.com, joaorosilva@gmail.com

Abstract. Dendro, a research data management (RDM) platform developed at FEUP/INESC TEC since 2014, was initially targeted at collaborative data storage and description in preparation for deposit in any data repository (CKAN, Zenodo, ePrints or B2Share). We implemented our own data deposit and dataset search features, consolidating the whole RDM workflow in Dendro: dataset exporting, automatic DOI attribution, and a dataset faceted search, among other features. We discuss the challenges faced when implemented these features and how they make Dendro more FAIR.

Keywords: research data · repositories · data citation · FAIR · Dendro

1 Introduction

Research data management (RDM) is essential for science, enabling the reproducibility of research results[8] and credit attribution[1, 3, 4]. Data citation requires data search, access, sharing and reuse, being a widely recognized practice in many research areas[1].

The FAIR principles[9] assist in RDM, both by humans and machines[5, 6], and help steer the design of adequate RDM workflows and software.

The Dendro platform[7][1] is an RDM platform under development at FEUP's InfoLab since 2014. It supports the collaborative management of research data within research groups, in preparation for their deposit in a data repository.

We made Dendro more FAIR by enabling the deposit and cataloguing of research datasets in Dendro itself, no longer requiring (but still allowing) the export to an external repository for long-term sharing of finished datasets. The new version also integrates with DataCite for automatic DOI attribution.

[⋆⋆] **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.

[1] https://github.com/feup-infolab/dendro

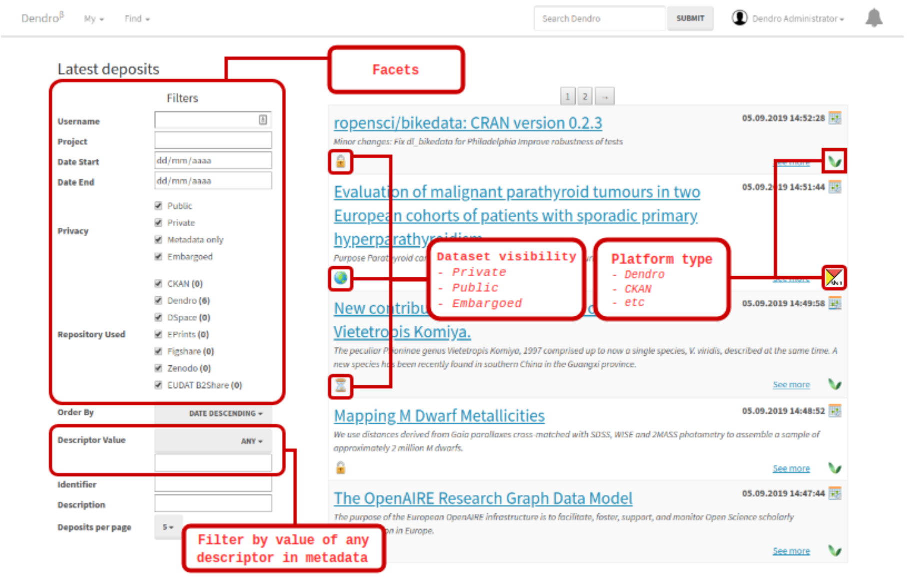
2 New deposit and citation features

Starting from the user requirements, we outline a new workflow for data deposit and citation and present several new features.

- **From researcher requirements to new Dendro features**. The implementation of this new deposit and citation functionality in Dendro derives from several researcher requirements. These features are partially present in some of the platforms (CKAN[2], DataVersef[3], Zenodo[4], DSpace[5] or ePrints[6]), but no single software solution implements them all[2].
- **Flexible, domain-specific metadata** Since not all repository platforms support generic and domain-specific metadata, these would have to be translated via crosswalking or embedding (e.g. specific metadata descriptors such as `Sample Size` from the Social Sciences or `Reactor Type` from Hydrogen Generation would become part of a Dublin Core `Description`). With our own deposit features, all metadata is kept intact in the published dataset.
- **History of metadata changes** Dendro also maintains a complete history of changes made to metadata records. After exporting to an external repository, this important provenance information would be separated from the finished dataset record; it is now maintained if the dataset is published in the Dendro instance.
- **Structured datasets** Since all other platforms represent their datasets as a flat list of files with metadata associated the root dataset record only, Dendro datasets (which are hierarchies of files and folders) had to be exported as BagIt files containing RDF metadata in order to preserve the structure and metadata associated to every node. Now the structure is kept intact.
- **Citation snippets** Zenodo and Dataverse already provide citation snippets. We have added this feature to Dendro, providing a BibTeX citation snippet for every dataset. The snippets are provided by DataCite upon minting of the DOI, which is done automatically by Dendro.
- **More refined access management** We now support embargo, the ability to specify terms of use when accessing a data and an on-demand dataset access workflow for creators to approve data requests.
- **Automatic DOI attribution** With this new implementation, Dendro can now mint persistent identifiers (DOI) via DataCite. These identifiers are now added to the metadata record of each published dataset, while a metadata record is automatically translated into the DataCite Schema 4 at the time of DOI minting. This way, even if that Dendro becomes unavailable, the metadata remains accessible via DataCite.
- **Faceted Search** Figure1 shows our new faceted search, powered by SPARQL and available as an API. Facets include: dataset visibility (public, private or

[2] https://ckan.org/
[3] https://dataverse.org/
[4] https://zenodo.org/
[5] https://duraspace.org/dspace/
[6] https://www.eprints.org/uk/

embargoed), dataset creator, creation date, and a logical combination (OR or AND) of descriptors-value pairs to filter datasets.



Dendro^β My ▾ Find ▾ | Search Dendro | SUBMIT | Dendro Administrator ▾

Latest deposits

Filters

Username	
Project	
Date Start	dd/mm/aaaa
Date End	dd/mm/aaaa
Privacy	☑ Public ☑ Private ☑ Metadata only ☑ Embargoed
Repository Used	☑ CKAN (0) ☑ Dendro (6) ☑ DSpace (0) ☑ EPrints (0) ☑ Figshare (0) ☑ Zenodo (0) ☑ EUDAT B2Share (0)
Order By	DATE DESCENDING ▾
Descriptor Value	ANY ▾
Identifier	
Description	
Deposits per page	5 ▾

1 2 →

ropensci/bikedata: CRAN version 0.2.3 — 05.09.2019 14:52:28
Minor changes: Fix dl_bikedata for Philadelphia Improve robustness of tests

Evaluation of malignant parathyroid tumours in two European cohorts of patients with sporadic primary hyperparathyroidism — 05.09.2019 14:51:44
Purpose Parathyroid ca

New contribu… Vietetropis Komiya. — 05.09.2019 14:49:58
The peculiar P…oninae genus Vietetropis Komiya, 1997 comprised up to now a single species, V. viridis, described at the same time. A new species h…s been recently found in southern China in the Guangxi province.
See more

Mapping M Dwarf Metallicities — 05.09.2019 14:48:52
We use distances derived from Gaia parallaxes cross-matched with SDSS, WISE and 2MASS photometry to assemble a sample of approximately 2 million M dwarfs.
See more

The OpenAIRE Research Graph Data Model — 05.09.2019 14:47:44
The purpose of the European OpenAIRE infrastructure is to facilitate, foster, support, and monitor Open Science scholarly …on in Europe.
See more

Fig. 1. Faceted dataset search in Dendro

2.1 A revised workflow

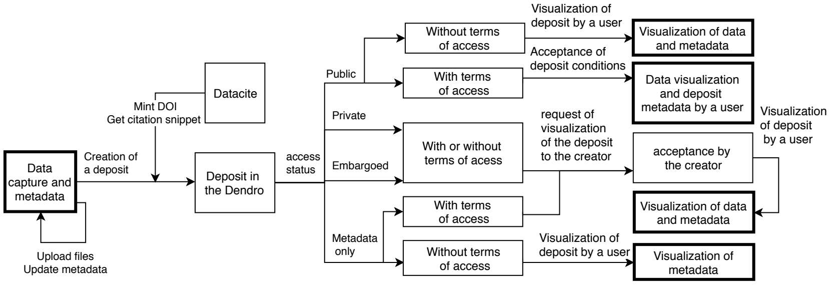

Data capture and metadata
Upload files
Update metadata
Creation of a deposit
Mint DOI
Get citation snippet
Datacite
Deposit in the Dendro
access status
Public
Private
Embargoed
Metadata only
Without terms of access
With terms of access
With or without terms of acess
With terms of access
Without terms of access
Visualization of deposit by a user
Acceptance of deposit conditions
request of visualization of the deposit to the creator
Visualization of deposit by a user
Visualization of data and metadata
Data visualization and deposit metadata by a user
acceptance by the creator
Visualization of data and metadata
Visualization of metadata
Visualization of deposit by a user

Fig. 2. New deposit workflow implemented in Dendro

Figure 2 shows our new workflow. Upon data deposit, the user can pick the visibility of the new dataset and, optionally, add access terms that any user wanting to access the data must accept. Visibility can be private (the dataset appears in the search but will not be accessible without author approval), public (data and metadata will be accessible), metadata only (dataset metadata will be visible when accessing the dataset, but not the data) and embargoed, which is private until after a certain date specified by the creator, and then becomes public. If terms of access are specified, they have to be accepted by users when accessing the data. In the case of private datasets, users can request access to the data creator. In that case, the creator is notified to approve or reject the request, and the requester is then notified of the result. Dataset creators retain the power to revoke access permissions at any time.

3 Conclusions

Dendro now supports storage, description and retrieval of research data over the entire data lifecycle. It also facilitates data citation by both a DOI and a BibTeX citation snippet for datasets, fostering credit attribution.

The platform is now more FAIR, closely the FAIR Guiding principles for scientific data management: persistent identifiers for datasets; rich, domain-specific metadata that complies with FAIR ontologies and faceted search, accessible to both humans and machines. Metadata is still accessible even if the data is not as the DataCite metadata record and DOIs will remain.

References

1. Altman, M., Crosas, M.: The Evolution of Data Citation: From Principles to Implementation. IASSIST quarterly **37**, 62–70 (2013)
2. Amorim, R., Castro, J., Rocha da Silva, J., Ribeiro, C.: A comparison of research data management platforms: architecture, flexible metadata and interoperability. Universal Access in the Information Society **16**(4) (2017). https://doi.org/10.1007/s10209-016-0475-y
3. Costello, M.J.: Motivating Online Publication of Data. BioScience (2009). https://doi.org/10.1525/bio.2009.59.5.9
4. Leonelli, S., Spichtinger, D., Prainsack, B.: Sticks and carrots: encouraging open science at its source. Geo: Geography and Environment (2015). https://doi.org/10.1002/geo2.2
5. Pontika, N., Knoth, P., Cancellieri, M., Pearce, S.: Fostering open science to research using a taxonomy and an eLearning portal. In: Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business - i-KNOW '15 (2015). https://doi.org/10.1145/2809563.2809571
6. Ross-Hellauer, T., Deppe, A., Schmidt, B.: Survey on open peer review: Attitudes and experience amongst editors, authors and reviewers. PLoS ONE (2017). https://doi.org/10.1371/journal.pone.0189311
7. da Silva, J.R., Ribeiro, C., Lopes, J.C.: Ranking Dublin Core descriptor lists from user interactions: a case study with Dublin Core Terms using the Dendro platform (2018). https://doi.org/10.1007/s00799-018-0238-x
8. Silvello, G.: Theory and practice of data display. CoRR **abs/1706.0** (2017). https://doi.org/10.1088/0305-4624/9/3/409, http://arxiv.org/abs/1706.07976
9. Wilkinson, M.D.: The FAIR Guiding Principles for scientific data management and stewardship pp. 1–9 (2016)