Domain-specific metadata for publishing research data in B2SHARE

Joana Rodrigues[a], João Aguiar Castro[b] João Rocha da Silva[c], Nelson Pereira[d], Yulia Karimova[e] and Cristina Ribeiro[f]

[a]INESC TEC and FEUP, University of Porto, Portugal, joanasousarodrigues.14@gmail.com
[b]INESC TEC and FEUP, University of Porto, Portugal, joaoaguiarcastro@gmail.com
[c]INESC TEC and FEUP, University of Porto, Portugal, joaorosilva@gmail.com
[d]INESC TEC and FEUP, University of Porto, Portugal, nelsonpereira1991@gmail.com
[e]INESC TEC and FEUP, University of Porto, Portugal, ylaleo@gmail.com
[f]INESC TEC and FEUP, University of Porto, Portugal, mcr@fe.up.pt

Abstract

Research data management (RDM) is a complex problem, involving many stakeholders with different needs that give rise to diverse requirements and technical challenges. Metadata production is a not only a practical issue with a clear impact on data discovery and reuse but also a clear social challenge for RDM. As research funders press for Open Data in the EU and researchers become more aware of the benefits of RDM and data publication, there is an increasing need for institutional support for RDM both at the institutional level and at the EU level. In this exploratory study we report on five data publication case studies, where we involved researchers in the description and sharing of their datasets using our Dendro platform and the EUDAT's B2SHARE. The resulting metadata records combine generic and domain-specific metadata descriptors, highlighting the need to complement Dublin Core descriptions when it comes to metadata for research datasets.

Key-words: research data management, metadata, data publication, Dendro

Proposal

The Dendro platform is an open-source data storage and description platform designed to help users describe their data files and is fully built on Linked Open Data and ontologies.

Moreover, it is a collaborative platform, capturing metadata as soon as researchers start to produce their data. Whenever users want to publish a dataset, they can export the corresponding project or folder to a repository such as CKAN, Zenodo, DSpace, or EUDAT's B2SHARE. Hence, Dendro complements data repositories in the early stage of the research data workflow.

Dendro manages metadata with flexible models. With EUDATLink, the gateway between Dendro and EUDAT built in the UPorto Pilot, data are deposited and Dublin Core descriptive metadata are transformed into B2SHARE metadata. Domain-specific metadata, however, is only present in the associated metadata files, which are not used to index or find the datasets. With a proper infrastructure and tools in place, researchers will be able to share data and exploit it to the full to derive new knowledge.

We report here a set of case studies from the long tail of science. The groups participating in these cases interacted with us at 3 points in time: 1) by discussing their metadata models with a data curator and defining a metadata model; 2) by describing datasets with the metadata model they adopted; 3) by comparing the full metadata records with the ones displayed in B2SHARE.

Each case study resulted from the interaction with a research group from the University of Porto that has described their data using the Dendro platform, using domain-specific metadata. The data packages, with both the datasets and the metadata, were then transferred from Dendro to B2SHARE. Together, the selected cases capture data description requirements from a diversity of research domains and from projects with a different scope, including observational, experimental and computational data.

The research data deposited in B2SHARE include the outcome of social science related studies, such as partial data from a longitudinal psychology project, and data from two projects in innovation management and social media communication analysis. The experimental and computational data were provided by groups in electrical and computer engineering, computational engineering, and materials engineering domains. The results show that researchers can provide metadata elements that best fit the context of their data, thus improving the access and interpretation of the datasets by B2SHARE users. In our case studies researchers have selected metadata elements from the Data Documentation Initiative, Dublin Core and the Friend of a Friend vocabularies, but also from domain-

specific metadata models developed at the University of Porto, previously represented in the Dendro platform.

The results, comparing the B2SHARE metadata with those in the full record captured with Dendro, suggest that the B2SHARE metadata representation may need to evolve to more flexible models. Further work is necessary to expand this study to more domains and contribute to the goal of flexible metadata in B2SHARE.