Promoting Semantic Annotation of Research Data by Their Creators: A Use Case with B2NOTE at the End of the RDM Workflow

Yulia Karimova(✉), João Aguiar Castro, João Rocha da Silva, Nelson Pereira, and Cristina Ribeiro

INESC TEC, Faculdade de Engenharia, Universidade do Porto,
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
ylaleo@gmail.com, joaoaguiarcastro@gmail.com, joaorosilva@gmail.com,
nelsonpereira1991@gmail.com, mcr@fe.up.pt

Abstract. Research data management is promoted at different levels with awareness actions carried out to encourage cooperation between researchers. However, data management requires tools to set the scene for researchers and institutions to disseminate the research data they produce. In this context good quality metadata play an important role by enabling data reuse. EUDAT is an European common data infrastructure, with integrated services for data preservation and dissemination. The TAIL project, at the University of Porto, proposes workflows based on Dendro, a collaborative environment that helps researchers prepare well described datasets and deposit them in a data repository. We propose a data deposit workflow use case for a small research project with emphasis in data annotation. Data is organized and described in Dendro; deposited in B2SHARE; and semantic annotation is performed with the new B2NOTE service from EUDAT.

Keywords: Research data management · Dendro · B2NOTE · Semantic annotation

1 Introduction

Research environments are characterized by a huge amount and wide variety of research data, but many issues are raised with respect to data access and data reuse [21]. Two motivators are prompting researchers to adopt a more active attitude into so-called open science practices: the compliance with funder requirements and the growing recognition of data as first-class research outputs. The fact that the main funding agencies in the US and EU now require researchers to attach Data Management Plans (DMP) to their grant applications is a clear statement of the importance of this topic. DMPs must specify the storage and long-time preservation conditions for the data during their lifecycle, as well as the representation of the data and the context of their creation. Without storage, one cannot recover the data in the long term, but the context is equally

E. Garoufallou et al. (Eds.): MTSR 2017, CCIS 755, pp. 112–122, 2017.
https://doi.org/10.1007/978-3-319-70863-8_11

important to make data findable and to allow others to make sense of them, favouring reuse.

In this context, the wide adoption of research data management (RDM) best practices is an essential step towards data reuse. Yet, despite the increasing interest in making their data available [20] and the existing institutional infrastructures and workflows designed to support researchers in RDM activities [7], researchers still need to deal with several problems related to RDM, such as the inadequacy of existing tools to support metadata creation [17].

Even if research data gets to the publication stage, potential reusers are very likely to disregard them if they are not conveniently described [18], since metadata is a determinant in data reuse [23]. High quality metadata is a positive outcome from the involvement of researchers in data description, as they are expected to generate more specialized descriptions [22]. A promising investment in RDM is to guide researchers in self-publication of research data, both to engage them in data management and to integrate RDM into the research workflow, alleviating data curation costs. In short, data reuse depends on the involvement of researchers in RDM activities, namely on the enrichment of data with quality metadata.

At the University of Porto, under the TAIL project [16], we are exploring the integration of different tools to build RDM workflows that are suitable for research scenarios with the typical requirements of the long tail of science. The proposed workflows anticipate the description requirements prior to the deposit stage, supporting them via the Dendro platform[1], but as an alternative we are also proposing researchers to directly deposit and describe their data in a CKAN-powered data repository at our institution, INESC TEC[2].

We explore here the definition of a workflow that integrates our tools with the services from the EUDAT platform, illustrating it with the use case of an MSc. researcher from the University of Porto who generated a dataset as a result of academic work and explored the publication of the data before the final thesis delivery. In this workflow, data are first prepared and described in Dendro and then transferred to the B2SHARE repository, where they can be further annotated with B2NOTE. This enables data to be reused and cited, considering that the nature of data from this project is appealing to others [15]. The next section is an overview of the main issues regarding RDM workflows, including a brief presentation of the Dendro + B2SHARE workflow, followed by a more detailed description of the EUDAT B2NOTE service.

2 RDM Workflows

An RDM workflow is a "sequence of repeatable processes (steps) through which research data passes during their lifecycle, including the steps involved in its creation, curation, preservation and possible disposal" [1]. To improve the value of data in the long term, researchers should systematically perform management

[1] Link: https://github.com/feup-infolab/dendro.
[2] Link: https://rdm.inesctec.pt/dataset/cs-2017-005.

tasks throughout the data lifecycle, meaning that, among other tasks, they need to describe their data on a regular basis. However, more often than not, they find themselves without adequate RDM tools, leaving them to resort to ad-hoc RDM practices supported by any tools that they have at their disposal [22], often addressing personal and immediate needs [13].

If a researcher promptly addresses data description during the initial stages of the data lifecycle, most of the work will be done when data gets to the deposit stage. The advantages are that early descriptions are probably richer than those made long after data production and are also more likely to ensure compliance with an existing DMP.

When data get to the deposit stage, researchers need access to trusted data repositories. Moreover, to improve data findability, accessibility and reusability, RDM workflows have to ensure that metadata is interoperable and has comprehensive information, is open and complies with legal and ethical rules for encouraging reproducible science [11].

However, RDM workflows are often built around multiple RDM systems that are not fully integrated, and any communication gaps between these systems may erode the willingness of the researchers to deposit their data—this is especially true if their dedication in early stages leads to redundant RDM tasks later in the data lifecycle. To safeguard more data, it is therefore crucial to provide effective and well integrated tools to researchers, in order to simplify and streamline the whole RDM workflow, making the processes clearer to the researchers [1].

The EUDAT Collaborative Data Infrastructure[3] currently proposes a suite of services to address the full lifecycle of research data. The services used in our workflow are B2SHARE—a trusted repository to support sharing of long tail data, B2FIND—a multidisciplinary joint metadata catalogue to find and access data in EUDAT, and B2NOTE—a semantic annotation service. These services are evolving, while EUDAT aims to establish a common model and lead the development of an infrastructure of data management services to cover European research data centers and community data repositories [11].

Complete RDM Workflow with Dendro and B2SHARE

At the University of Porto, with the TAIL project, we are proposing workflows that integrate tools to support RDM during the research data lifecycle, with particular attention to the data description requirements from different research domains [16].

Figure 1 depicts a workflow consisting of the Dendro platform and the EUDAT B2SHARE service, which interact through an API. This connection is part of a Data Pilot established between the TAIL team and EUDAT to allow researchers to describe their data using generic and domain-specific vocabularies through Dendro, and to import the resulting data and metadata to B2SHARE [6].

In Dendro, description ideally occurs when the data is captured (Steps 1 and 2), considering that pertinent information about research data may be

[3] Link: https://www.eudat.eu/.

forgotten if not recorded right away. The purpose of data description in Dendro is to capture metadata for research datasets, based on ontologies [19], combining description elements from widely adopted metadata standards such as Dublin Core, for the sake of interoperability, with domain-specific elements for specificity. The latter can either be sourced from domain metadata standards, or otherwise defined in collaboration with the researchers after analysing the terms they already associate to their data [3,4]. For some of the domains tested with Dendro, controlled vocabularies were created to restrict the possible values for some fields to facilitate description and improve its quality [10]. Since Dendro is a collaborative platform, researchers can improve their metadata from feedback provided by others. This is implemented in Social Dendro [14], where researchers are notified when others *like, share or comment* their metadata. When researchers decide that their data are ready for deposit they can send them to a data repository that complies with their requirements. Dendro currently interfaces with CKAN, Zenodo, Figshare and EUDAT's B2SHARE, among others. Figure 1 shows a deposit in B2SHARE (Step 3). After the deposit, users can proceed to data annotation, this time with the B2NOTE service, using tags derived from controlled vocabularies, or free-text keywords and comments (Step 4).

Fig. 1. Complete RDM workflow with Dendro and B2SHARE

This workflow illustrates that, while Dendro is intended for the organization and description of data, EUDAT B2SHARE is tasked with publishing and sharing data. B2NOTE complements the annotation of datasets at a post-deposit stage.

3 B2NOTE

B2SHARE, like most multidisciplinary data repositories, has no specific community in mind, assuming a generalist approach to data publication [2] reflected in its domain-agnostic deposit form.

B2NOTE is a standalone research data annotation service based on the W3C Web Annotation Data Model. It currently integrates with B2SHARE, and will integrate with other EUDAT services [5]. With a flexible approach to data annotation, B2NOTE appears as a post-deposit tool. In most systems, the metadata is not supposed to change after publication. Data annotation can be regarded as a source of community metadata, providing evidence of data usage and comments on their limitations, since it is available to users without mediation.

When used for specific metadata elements, controlled vocabularies can provide lists of terms that promote uniform descriptive cataloging, labeling, or indexing [8]. Controlled vocabularies are also expected to improve the quality of the descriptions added to research datasets by restricting the valid values of specific metadata elements. However, it has been observed that, while researchers are interested in using them, they are not widely implemented in data repositories [25].

Using B2NOTE, researchers can complement the information available in the metadata generated by the authors, using semantic tags, without changing the original data file and its description. These tags are filled by means of auto-completion boxes where terms from specific controlled vocabularies appear. These additional tags can help other users find, organize and aggregate files, datasets and documents. The goal is to improve retrieval, helping users find specific files by the annotated subject. In the current version, semantic tags are drawn from controlled vocabularies in the Bioportal repository, but more vocabularies will be considered by EUDAT, based on the analysis of controlled vocabularies already in use in research data repositories. Since there is risk of vocabulary fragmentation, the choice of the right vocabulary for multidisciplinary data annotation can be addressed through a social marketplace where users share their discipline-specific experiences [11,12].

Besides using the semantic tags from controlled vocabularies, B2NOTE users can also annotate data with free-text keywords that identify the subject of a resource if no semantic tag is appropriate, or include a free-text comment, open to any kind of additional information. Free-text keywords are a good complement and a more flexible approach to annotation than the controlled vocabularies, allowing users to classify and retrieve resources based on *folksonomies*. This can result in the expansion of the formal structured vocabularies with new terms [24]. However this approach has its own limitations, mostly related to issues with vague meaning, term variations, homonyms and polysemy, and may result in tags that only make sense to an individual user, making it difficult to build a hierarchy of concepts [9].

Free text comments capture non-structured, informal information, desirable for expressing opinion and recommendations. Free comments are open to all users and may be used to enable the collaboration between researchers.

4 Use Case

This use case illustrates a complete RDM workflow, from data description in the Dendro platform to data annotation using B2NOTE after publication.

Step 1. Data Production

Different types of data call for different management practices and data description requirements vary depending on the data type or discipline. In this case the data were produced as part of an M.Sc. dissertation, focused in entity extraction from Portuguese news articles. These entities can be names of persons, organizations, or places. The goal of the work was to select a tool for entity extraction that can be adopted in projects with similar entity-related challenges, namely ANT[4]. The dataset contains models trained with a dataset created in the HAREM evaluation initiative[5]. The dataset that results from the trained models, *HAREM NER Models*, is a valuable contribution since this kind of data is rare in the Portuguese language.

Step 2. Deposit and Description

Deposit of the *HAREM NER Models* dataset in Dendro started with the creation of a project to organize the data and associated metadata. From there the researcher managed the data exploring the four main sections (see Fig. 2): user area (area **1**), the file manager (area **2**), the description zone (area **3**) and the descriptor selection zone (area **4**). In the user area the researcher has created a folder named *HAREM NER MODELS*, where four files, corresponding to the models, were deposited. Then, the researcher selected the vocabulary with the concepts that better fit the data. In this case Dublin Core descriptors were selected to describe the folder, namely *Title*, *Subject*, *Description*, *License*, *Format* and *Language*. Since each individual file has specific properties, a *Description* was added to each one. After the data were organized and described in Dendro, the researcher sent a package, containing both data and metadata files, to the B2SHARE data repository. This makes the data exposed to a larger community [18] and allows for data citation.

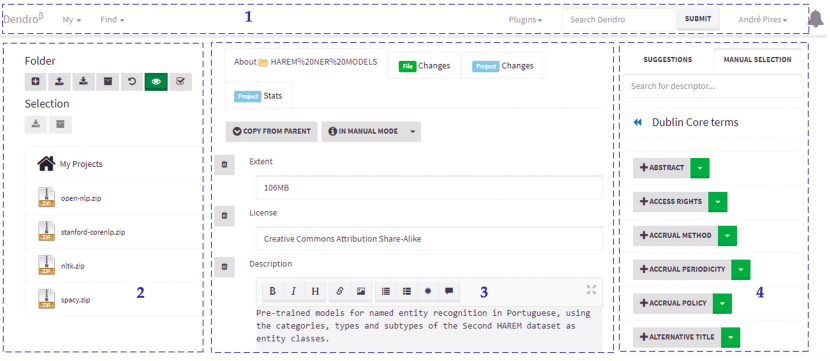

Fig. 2. Data deposit and description in Dendro

[4] Link: http://ant.fe.up.pt/.
[5] Link: http://www.linguateca.pt/aval_conjunta/HAREM/harem_ing.html.

Step 3. Publication

When a data package is transferred from Dendro to B2SHARE, the former automatically fills the metadata fields available in B2SHARE at the deposit stage. The remaining fields, present in Dendro but not ingestable by B2SHARE, are exported as an RDF file (see Fig. 3, area **1**), that can be consulted by the users to see more information about the data (area **2**) [18].

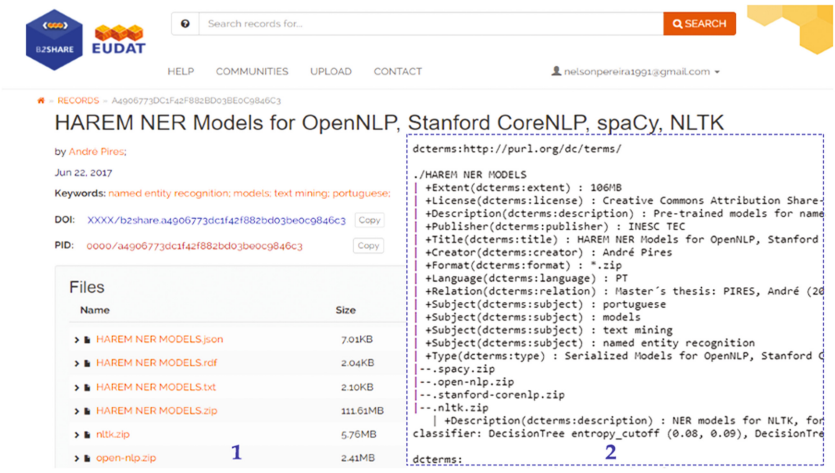

Fig. 3. Data package deposited in B2SHARE and additional metadata

This transfer takes advantage of the data description work the researcher has already performed in Dendro to fill in the metadata required in B2SHARE, while keeping the full metadata record from Dendro. However, this approach has its limitations since the information contained in the RDF file is not actually used for retrieval purposes in B2SHARE, and its format is not user-friendly. The ability to add annotations after the deposit stage, using B2NOTE, may alleviate some of this inconvenience.

Step 4. Annotation

After the data are published in B2SHARE, annotations in B2NOTE can add information to the metadata previously captured, and link resources within the EUDAT CDI or with external resources (see Fig. 4, area **1**). Annotations are saved in a machine-readable format, according to the W3C Web Annotation model, in order to be findable and viewable [5].

There are three types of annotations in B2NOTE: semantic tags, free-text keywords and free text comments (area **2**). In this case the researcher was aware that some information to be shared with potential users had not been captured during the preparation stage in the Dendro platform. Thus, the researcher used B2NOTE to add a reference to the open-source tool *OpenNLP* using a free-text keyword; OpenNLP is a tool that supports natural language processing,

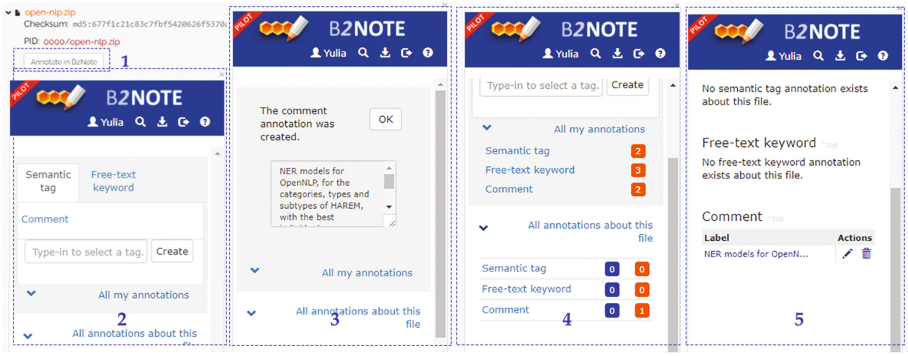

Fig. 4. Data annotation in B2NOTE

particularly entity extraction, used to analyze the Portuguese news articles. The free comment option was also used to comment about missing information in one of the files (area **3**). At this time there are no recommendations on how to write these comments; they can be regarded as personal notes used to provide more insight, or updates, to help others explore the data in a meaningful way.

The annotations made by the researcher are then displayed for all users, "*All annotations about this file*", and all registered users can make additional annotations to the data file. B2NOTE users can choose to visualize all comments or can choose to show only their own annotations, by clicking "*All my annotations*" (areas **4, 5**).

Registered users can also search the annotated file using "Search" (see Fig. 5, area **6**) and export results to JSON-LD or RDF files for their own purposes (areas **7, 8**). Searching is performed over semantic-tag and free-text keywords.

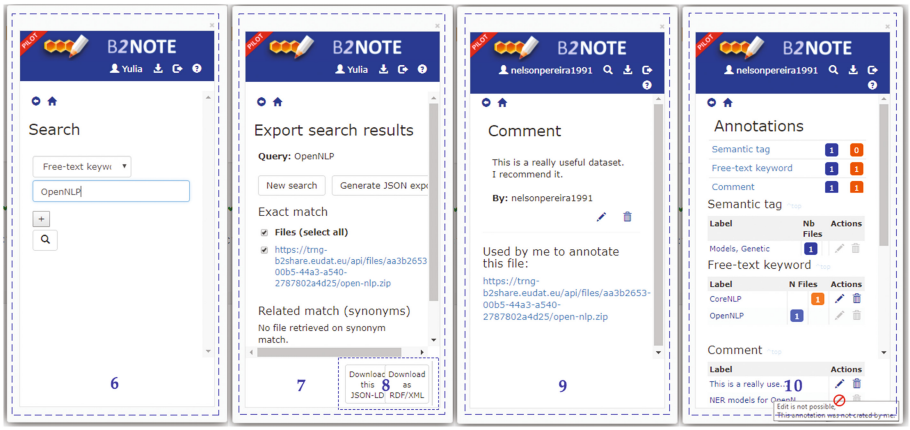

Fig. 5. Search, export results and annotation visualization in B2NOTE

Although users can add as much information as they see fit to any file (area **9**), they cannot edit annotations made by other users (area **10**). In our case, one of the users added a free-text comment about the reuse and utility of this dataset. The M.Sc. researcher may later access the B2NOTE platform and read the comments regarding the dataset and even reply to them.

5 Conclusions

Data reuse is strongly influenced by the information that data creators convey to others about the context of data they intend to share. Usually, research data deposit workflows resemble traditional publication ones, with the risk of essential metadata being lost, if captured at all.

The Dendro + B2SHARE + B2NOTE workflow presented here addresses this issue by covering important stages of the data lifecycle, reinforcing the notion that data description should appear on time in the research process to render good quality metadata. Furthermore, data annotation at the end of the workflow adds new pathways to the data, while also encouraging the exchange of ideas between researchers using more casual notes.

Tools and guidelines for better and clearer RDM encourage researchers to share their data on a broader scale. The overall workflow, and in particular the Dendro and the EUDAT services, are currently under development and are more likely to succeed if they evolve in close collaboration with researchers. For instance, at the time of writing, annotations made in B2NOTE are publicly available, yet in a future release researchers will have the option to keep them private, a requirement gathered from user feedback. It would also be useful for researchers if B2NOTE notifies them when new annotations occur. This kind of behaviour is explored in Social Dendro, a social extension for description in Dendro.

The use case in this paper is as close as possible to a real-world scenario, taking into consideration that the B2NOTE service is only available in a training instance of the B2SHARE platform. Therefore, there are aspects that will be interesting to assess as B2NOTE evolves. An evaluation of the use of annotations, for instance, and how they help users find data, can justify the effort of creating richer metadata. The need to update data and metadata, and their impact in the final stages of the data publication workflow, can also result from the observation of the annotation tool.

From the researcher perspective this case study was an opportunity to explore a set of RDM tools, according to users needs, rather than the execution of a designated set of tasks to evaluate tool performance. The researcher had the primary goal of publishing the project data. This led to a natural and low-effort exploration of the tools, using Dublin Core according to their needs. The obvious way to expand this work is to handle use cases that demand more specific metadata elements, clearly demonstrating the role that staging platforms, like Dendro, or annotation tools, as BNOTE, may have to alleviate the difficulties with metadata in generic data repositories.

Future work will be informed by further use cases resulting from data deposit needs derived from funder requirements, as a way to train researchers in RDM activities. This will make it possible to amass domain metadata and requirements stemming from the domain data types that are likely to improve RDM workflows.

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 TAIL, POCI-01-0145-FEDER-016736. João Aguiar Castro is supported by research grant PD/BD/114143/2015, provided by the FCT - Fundação para a Ciência e a Tecnologia. We thank Yann Le Frank and the B2NOTE team for the availability of the beta version of B2NOTE and the helpful remarks.

References

1. Addis, M.: RDM workflows and integrations for higher education institutions using hosted services (2015). Arkivum white paper https://www.digital-science.com/resources/reports/
2. Assante, M., et al.: Are scientific data repositories coping with research data publishing? Data Sci. J. **15**(6), 1–24 (2016). https://doi.org/10.5334/dsj-2016-006
3. Castro, J.A., da Silva, J.R., Ribeiro, C.: Creating lightweight ontologies for dataset description. Practical applications in a cross-domain research data management workflow. In: IEEE/ACM Joint Conference on Digital Libraries (JCDL) (2014). https://doi.org/10.1109/JCDL.2014.6970185
4. Castro, J.A., et al.: Involving data creators in an ontology-based design process for metadata models. In: Malta, M.C., Baptista, A.A., Walk, P. (eds.) Developing Metadata Application Profiles, pp. 181–214. IGI Global (2017). https://doi.org/10.4018/978-1-5225-2221-8.ch008
5. EUDAT. Annotate your research data with B2NOTE (2017). EUDAT news https://eudat.eu/news/annotate-your-research-data-with-b2note
6. EUDAT. EUDAT as a long-term repository for the University of Porto (2017). EUDAT Data Pilot https://eudat.eu/communities/eudat-as-a-long-term-repository-for-the-university-of-porto
7. Van den Eynden, V., et al.: Managing and sharing data - best practice for researchers. UK Data Archive, pp. 1–40 (2011). ISBN: 1904059783
8. Hedden, H.: Taxonomies and controlled vocabularies best practices for metadata. J. Digit. Asset Manag. **6**(5), 279–284 (2010). https://doi.org/10.1057/dam.2010.29
9. Huang, S.-L., Lin, S.-C., Chan, Y.-C.: Investigating effectiveness and user acceptance of semantic social tagging for knowledge sharing. Inf. Process. Manage. **48**(4), 599–617 (2012). https://doi.org/10.1016/j.ipm.2011.07.004
10. Karimova, Y., Castro, J.A.: Vocabulários controlados na descrição de dados de investigação no Dendro. In: Cadernos BAD N.2, jul-dez, pp. 241–255 (2016)
11. Latif, A.: EUDAT: Research data infrastructure and European Open Science Cloud vision (2017). Team ZBW Mediatalk https://www.zbw-mediatalk.eu/en/2017/05/eudat-research-data-infrastructure-and-european-open-science-cloud-vision/
12. Le Franc, Y.: Organise, retrieve and aggregate data using annotations with B2NOTE (2017). EUDAT webinar https://eudat.eu/events/webinar/eudat-webinar-organise-retrieve-and-aggregate-data-using-annotations-with-b2note

13. Mayernik, M.S.: Metadata realities for cyberinfrastructure: data authors as metadata creators (2011). SSRN https://ssrn.com/abstract=2042653. https://doi.org/10.2139/ssrn.2042653
14. Pereira, N., da Silva, J.R., Ribeiro, C.: Social Dendro: social network techniques applied to research data description. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) TPDL 2017. LNCS, vol. 10450, pp. 566–571. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67008-9_47
15. Pires, A.: Named entity recognition on Portuguese web text. Master thesis, Faculdade de Engenharia da Universidade do Porto (2017)
16. Ribeiro, C., et al.: Projeto TAIL - Gestão de dados de investigação da produção ao depósito e à partilha (resultados preliminares). In: Cadernos BAD N.2, jul-dez, pp. 256–264 (2016)
17. Shearer, K., Furtado, F.: COAR survey of research data management: results. Confederation of OpenAccess Repositories (2017). https://www.coar-repositories.org/files/COAR-RDM-Survey-Jan-2017.pdf
18. Silva, F., Amorim, R.C., Castro, J.A., da Silva, J.R., Ribeiro, C.: End-to-end research data management workflows: a case study with Dendro and EUDAT. In: Garoufallou, E., Subirats, C.I., Stellato, A., Greenberg, J. (eds.) MTSR 2016. CCIS, vol. 672, pp. 369–375. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49157-8_32
19. da Silva, J.R., et al.: The Dendro research data management platform: applying ontologies to long-term preservation in a collaborative environment. In: iPres 2014 Conference Proceedings (2014)
20. Tenopir, C., et al.: Changes in data sharing and data reuse practices and perceptions among scientists worldwide. Plos One **10**(8) (2015). https://doi.org/10.1371/journal.pone.0134826
21. Vines, T.H., et al.: The availability of research data declines rapidly with article age. Curr. Biol. **24**(1), 94–97 (2014). https://doi.org/10.1016/j.cub.2013.11.014
22. White, H.C.: Descriptive metadata for scientific data repositories: a comparison of information scientist and scientist organizing behaviors. J. Libr. Metadata **14**(1), 24–51 (2014). https://doi.org/10.1080/19386389.2014.891896
23. Willis, C., Greenberg, J., White, H.: Analysis and synthesis of metadata goals for scientific data. J. Assoc. Inf. Sci. Technol. **63**(8), 1505–1520 (2012). https://doi.org/10.1002/asi.22683
24. Zervas, P., Sampson, D.G.: The effect of users' tagging motivation on the enlargement of digital educational resources metadata. Comput. Hum. Behav. **32**, 292–300 (2014). https://doi.org/10.1016/j.chb.2013.06.026
25. Zhang, Y., et al.: Controlled vocabularies for scientific data: users and desired functionalities. Proc. Assoc. Inf. Sci. Technol. **52**(1), 1–8 (2015). https://doi.org/10.1002/pra2.2015.145052010054