Data Deposit in a CKAN Repository: A Dublin Core-Based Simplified Workflow

Yulia Karimova(✉), João Aguiar Castro, and Cristina Ribeiro

INESC TEC, Faculty of Engineering, University of Porto,
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
ylaleo@gmail.com, joaoaguiarcastro@gmail.com, mcr@fe.up.pt

Abstract. Researchers are currently encouraged by their institutions and the funding agencies to deposit data resulting from projects. Activities related to research data management, namely organization, description, and deposit, are not obvious for researchers due to the lack of knowledge on metadata and the limited data publication experience. Institutions are looking for solutions to help researchers organize their data and make them ready for publication. We consider here the deposit process for a CKAN-powered data repository managed as part of the IT services of a large research institute. A simplified data deposit process is illustrated here by means of a set of examples where researchers describe their data and complete the publication in the repository. The process is organised around a Dublin Core-based dataset deposit form, filled by the researchers as preparation for data deposit. The contacts with researchers provided the opportunity to gather feedback about the Dublin Core metadata and the overall experience. Reflections on the ongoing process highlight a few difficulties in data description, but also show that researchers are motivated to get involved in data publication activities.

Keywords: Research data management · Metadata · Dublin Core · CKAN · Data publication

1 Introduction

Research data management (RDM) is becoming an important activity for researchers. To promote auditability of results, access, reuse, and transparency, the deposit of research data is required in the grant applications to most funding agencies [10]. Moreover, Data Management Plans (DMP) are also required, to provide detailed information about the project, indicating the context and objectives, method, tools and techniques of data collection, the form of preparation, how data will be described, preserved, and shared, as well as issues related to reuse [3,6]. Therefore, competence in RDM is considered an essential skill for good scientific practice.

It has been observed that researchers show interest in publishing data [22], thus having their work discovered, reused and cited in essential [24]. Yet,

P. Manghi et al. (Eds.): IRCDL 2019, CCIS 988, pp. 222–235, 2019.
https://doi.org/10.1007/978-3-030-11226-4_18

researchers still face several difficulties in RDM activities, such as insufficient experience, lack of knowledge on metadata standards, inadequate tools for deposit and description of the data, lack of time, and lack of perceived rewards for the RDM tasks [18–20].

In this context, many institutions are looking for solutions to help researchers publish their data, supporting them in the RDM activities and providing tools and repositories [13,17,24]. From the researchers' perspective, data description and deposit should be simple, supported by tools that ease the creation of metadata. Metadata are essential for data access, interpretation and preservation. However, metadata standards can be complex and hard to adopt by researchers [7,15].

In this work we introduce researchers to RDM through data description, in a quick and practical fashion, using Dublin Core descriptors. The assumptions are that (1) the domain-neutral nature of Dublin Core is convenient for situations where a specific assessment of metadata requirements is not feasible; and (2) it is easy for researchers to grasp the concepts behind Dublin Core descriptors.

The aim of this paper is to describe a set of data deposit examples performed by researchers, with specific attention to the difficulties that they face and ways to overcome them. The next section is an overview of issues related to the development of the data repository at INESC TEC (Institute for Systems and Computer Engineering, Technology and Science, Portugal)[1], followed by a presentation of the data deposit process on the repository in Sect. 3 and the details of the examples with the identification of difficulties in Sect. 4. Results are presented in Sect. 5, followed by the discussion of feedback by researchers in Sect. 6. Section 7 presents the conclusions and future work.

2 The Data Repository at INESC TEC

Under the TAIL project[2], we are creating RDM workflows based on the integration of different tools according to the requirements of the researchers and their groups. The complete workflow covers important stages of the data lifecycle. The description stage occurs in the Dendro[3] platform, which helps researchers prepare datasets, combining generic and domain-specific metadata elements. In the context of the TAIL project, we elaborated specific metadata models for Material Fracture, Analytical Chemistry, Biodiversity, Simulation of Vehicles, Biological Oceanography, Hydrogen Production [4] and adopted descriptors from the Data Documentation Initiative [23] for several areas in the Social Sciences. These metadata models were based on contributions by researchers concerning the contextual information required to enable data interpretation. When they are ready for publication, data are transferred to a data repository, such as B2SHARE[4] [12].

[1] https://www.inesctec.pt/en.
[2] https://www.inesctec.pt/en/projects/tail.
[3] https://github.com/feup-infolab/dendro.
[4] https://b2share.eudat.eu/.

Although we recognize the need to prepare metadata according to domain-specific requirements, this process may take some time and require an effort that researchers are not able to commit. Moreover, there are many interesting datasets from closed projects that will not reach publication stage without a more agile process for deposit. As an alternative, we designed a workflow where data are directly deposited in the data repository at INESC TEC[5]. The data repository is an instance of CKAN (Comprehensive Knowledge Archive Network)[6], an open-source data platform built as a data management system, popular with open government data around the world (e.g. UK government and European Commission), and widely supported by the developer community [1,5,24].

CKAN provides an intuitive interface and visualization tools, which make data easily accessible. Moreover, it has a flexible architecture that allows for the customization of its features. Metadata fields, for instance, can be customized with key-value pairs, so users can define new ones [1].

Research at INESC TEC covers many domains, and the work with researchers to capture metadata requirements and design metadata models is ongoing. In this simplified workflow we use Dublin Core as a domain-neutral metadata schema. This is considered as a prudent entry plan for researchers lacking RDM skills. Many data repositories are already based on Dublin Core metadata[7], some allowing the addition of new descriptors. Moreover, a standard metadata schema is convenient for search and access, accounting for the diversity of research data from different scientific domains.

CKAN has default descriptors suitable for datasets: Title, Description, Tags, License, Organization, Visibility, Source, Version, Author, Author email, Maintainer, Maintainer Email, and Group. More detail can be added with the descriptors available for each file of the dataset: Name, Description, and Format. Dates for the creation and modification of the dataset are automatically generated upon deposit, and recorded on the Created and Last Updated CKAN descriptors. Each dataset and each of its files get an ID in the process. Most descriptors are easily mapped to Dublin Core, which allows for OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) compliance. The CKAN metadata schema is very simple, yet it can be extended according to researchers' requirements and to assure platform interoperability [24]. In the INESC TEC data repository some key-value pairs were added to enable the use of descriptors beyond the default CKAN ones.

3 Simplified Data Deposit Workflow

In the context of the TAIL project, we collaborated with several researchers from different domains on RDM issues [16]. Some contacts led to data description and then deposit at the INESC TEC repository. The process starts with the decision of researchers to share their data or to cite data in a research paper and a

[5] https://rdm.inesctec.pt/.
[6] https://ckan.org/about.
[7] https://www.re3data.org/metrics/metadataStandards.

contact with the RDM team. The RDM process proceeds with a first meeting with researchers about general RDM issues and an introduction to the INESC TEC data repository. This also serves to assess familiarity of the researchers with respect to data publication and metadata standards.

To simplify the preparation of the data, we created a dataset deposit form based on Dublin Core[8]. This form is a template for the researcher to fill in. The researcher completes the form and returns it to the curator, who validates the metadata and completes the deposit process.

We chose Dublin Core as the core of the dataset deposit form since this standard is understandable by most [9], while the use of descriptors widely used in repositories allows for basic interoperability [8] and interdisciplinary discovery [14].

The dataset deposit form contains several Dublin Core descriptors[9], Dublin Core Qualifiers[10], and CKAN descriptors. CKAN descriptors Organization, Visibility, Version, Maintainer, Maintainer email and Group are not part of the form. They are assigned by the curator in the deposit step.

Table 1. Dataset attributes and corresponding descriptors in the repository

Dataset attributes	Corresponding descriptor and vocabulary
Availability	Visibility (CKAN), DOI
Bibliometric data	-
Coverage	Coverage.Temporal (Dublin Core), Coverage.Spatial (Dublin Core)
Date	Date (Dublin Core)
Format	Format (CKAN), Format (Dublin Core), Format.Extent (Dublin Core)
License	License (CKAN), License (Dublin Core)
Minimal description	Title (CKAN), Name (CKAN), Author (CKAN), Author email (CKAN), Description (CKAN), Maintainer (CKAN), Maintainer email (CKAN), Type (Dublin Core), Language (Dublin Core), Publisher (Dublin Core), Contributor (Dublin Core)
Paper reference	Relation (Dublin Core)
Project	Organization (CKAN), Group (CKAN)
Provenance	Source (CKAN), Version (CKAN)
Subjects	Tags (CKAN)

Descriptors from several vocabularies are being used in research data repositories such as Dryad, Figshare, Zenodo or CSIRO. There is no agreed-upon vocabulary for scientific data, but a set of eleven so-called classes of metadata attributes have been identified by Assante et al. to capture the essential aspects of datasets [2]. Table 1 presents a mapping of these dataset attributes into the descriptors used in the INESC TEC repository.

[8] https://tinyurl.com/ybbwvq57.
[9] http://dublincore.org/documents/dcmi-terms/.
[10] http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/.

The Availability attributes include descriptors that help to get access to the dataset [2]. Descriptor Visibility used in our form defines whether the dataset is publicly available or privately closed, and is therefore classified as an Availability attribute. If the dataset is private only the curator has access to the dataset. This is provided to comply with embargo periods, and the availability status can be altered at any time. The INESC TEC repository also assigns DOI to datasets. The DOI descriptor also contributes to the Availability attribute and its assigned by the curator. The INESC TEC data repository does not provide bibliometric data, such as statistics about data visualization or the number of downloads. These are important features to address in future development. This fact is acknowledged in the table in the line corresponding to the Bibliometric Data attribute.

The rest of the attributes correspond to standard descriptors and can provide information such as author, title, description, format, license, spatial, temporal coverage, data creation and related publication for the dataset. In some cases, we added Dublin Core descriptors even though the corresponding CKAN descriptors are present. One example is the License descriptor from both CKAN and Dublin Core, used since the CKAN descriptor does not include the "*Creative Commons Attribution-NonCommercial-NoDerivs 2.0 Generic (CC BY-NC-ND 2.0)*" in its default license list.

The deposit process is accomplished with researchers filling the form, by themselves or with our support, and with the verification of the metadata before deposit. This "approval" step provides control over the description and enforces some required information, such as the data formats and the size of the dataset and its files. As soon as the researcher sends the first version of the form, we evaluate the metadata record taking into account the knowledge of the nature of the domain, acquired in previous meetings. When a dataset contains GPS data and the metadata record lacks information about the geographic coverage, we assume that this is a possible metadata quality limitation, and ask the researcher to fill in the corresponding descriptor. If necessary, the complementary metadata is added with our help. This approach to metadata quality [21] is based on a human assessment performed by the curator while verifying the form. Moreover, the adoption of Dublin Core as a standard to ensure interoperability and the existence of this second round of description to enrich the metadata records also contribute to metadata quality.

4 Examples of Deposited Datasets

The deposit process resulted in 21 datasets from research groups in the domains of Biomedical engineering (1 dataset), Environmental radioactivity (7 datasets), Biomedicine (3 datasets), Robotics (1 dataset), Information science (1 dataset), Natural language processing (3 datasets), Music streaming (1 dataset), and Information retrieval (4 datasets). In general we had the collaboration of one researcher from each group for the description and deposit of their data. However, in every case all the group elements were aware of the process and validated

the deposits. This is a test to a deposit process for research groups at INESC TEC where RDM tasks and the contact with curators are delegated to some group members.

The size of the files in the datasets is in the range of 2 kB to 4 GB, which is the configured limit on the repository. However, there is no limit on the number of files in a dataset. Deposits took place over a period of one and a half year, with some groups that make systematic data collection contributing with several datasets. Table 2 shows the distribution of deposits in time, and their accessibility status (public—open to all; private—not accessible to all). Sharing of private datasets is not excluded, but it requires a request to the authors and possibly some agreement on the terms of use.

Table 2. Datasets deposited at the INESC TEC data repository up to June 2018

Domain	Datasets deposited in 2017				2018	
	1 trimester	2 trimester	3 trimester	4 trimester	1 trimester	2 trimester
Biomedical engineering	•					
Environmental radioactivity	•	•	•	•	•	•
Biomedicine		•			•	
Robotics					•	
Information science					•	
Natural language processing		•	•	◇		
Music streaming	•					
Information retrieval	•		◇			

• - *public datasets* ◇ - *private datasets*

We kept record of the collaboration with researchers: for each case, we wrote a short description of the dataset, issues, questions, and features of interest mentioned by the researchers. This included datasets that researchers decided not to deposit, with the arguments supporting their decisions. In the following we provide a brief description of the process in each group.

Biomedical Engineering: The goal of this group is the development of products, tools, and methods for prevention and early detection of diseases[11]. The first contact revealed a lack of knowledge in RDM. We explained the meaning of the descriptors in our form. Afterward, the researcher sent us the completed form. We changed the Title, added information about Authors, Contributors, Format, Format.Extent and Relation. The references to papers were standardized and associated to the existing DOI. After checking all this information, we proceeded with the deposit on the repository and sent the link to the researcher to verify the description. Although this group has only one deposited dataset, they promoted the institutional repository with other groups, by sharing our

[11] https://www.inesctec.pt/en/centres/c-ber.

contact, showing others how to access the deposited dataset, and telling about their RDM experience.

Environmental Radioactivity: The Environmental radioactivity research group deals with engineering problems facing the industry, by analyzing, designing, mining and implementing large information systems[12]. They had prior RDM experience, namely on metadata standards and on the use of a domain-specific data repository. We check this repository and obtained an example of a Readme.txt file with all the recommended descriptors. However, in conversation with the group it was decided to use the INESC TEC dataset deposit form because the descriptors recommended in the domain repository were considered too specialized and not suitable for their data. After the first deposit this group made regular deposits with data from ongoing campaigns, with increased description quality. This constitutes a running collaboration that has raised the interest of this group in RDM activities, leading to further collaboration in the design of a data management plan.

Biomedicine: A researcher of this domain (see footnote 11) had interest in data deposit to publicize results related to neuro-technologies. The dataset contained sensitive information, namely images of patients, and although the information was anonymized the researcher decided to keep the data as restricted access. A common requirement from researchers in this domain is to restrict dataset download to registered users under the acceptance of a specific license. Therefore, we recommended the preparation of a password-protected zipped dataset and the specification of the license in the dataset description. The researchers showed interest in having an interface with mandatory fields for prospective users, providing information on how to request access to the dataset. This request has been recorded for future implementation.

Information Science: The researcher in this domain worked in a group related to the development and promotion of innovation management practices, and was familiar with RDM activities and metadata related issues. Therefore, the data description and deposit tasks ran easily and the deposit was completed with little effort.

The ***Natural language processing*** and ***Information retrieval*** research groups are part of the Information Systems and Computer Graphics center (see footnote 12), with research related to programming languages and data processing. In the data preparation phase, we identified, in both cases, that some datasets could be made publicly available while others were private, depending on the permission status granted by the original databases.

The process in the ***Robotics***[13] and ***Music streaming***[14] domains followed the defined approach, without any specific challenge. We expect a productive collaboration with these groups, due to their strong Data Science connection, and the volumes of data generated in their projects. Datasets are expected from

[12] https://www.inesctec.pt/en/centres/csig.
[13] https://www.inesctec.pt/en/centres/liaad.
[14] https://www.inesctec.pt/en/centres/cras.

robotic solutions, autonomous navigation, a variety of sensor measurements, and also complex objective decision models. Researchers in these domains are likely to raise interesting issues for the RDM process.

5 Description Results

In retrospect, most of the researchers that we have worked with never received RDM training of any sort, not knowing, in some cases, anything about data description and data deposit. They agree on the challenging nature of the data description process. We noticed that some concepts were easily understood, but others such as metadata or descriptor required more discussion, and the support provided by the RDM team was valued. Another common difficulty was the knowledge of metadata standards to be used in their domains. Dublin Core was used as an example of a generic standard. Although they felt the need to make their data accessible to others, they did not know how to do it. Table 3 compares the number of descriptor occurrences, by descriptor, filled in by the researchers in the first description round, with the final number of occurrences after the collaboration with the curator. The final number corresponds to the metadata actually included in the deposit of the dataset.

The results show that the Title, Author and Author email descriptors were the most frequently used. The Description element is also among the most used ones, which is natural since it is a descriptor that gives flexibility to capture information that otherwise would make sense in domain-specific elements, e.g. concerning experimental configurations. The descriptor Tags was widely used by the researchers to represent the subject, with the goal of improving data findability in the repository. The descriptors Source, Contributor, and Format.Extent were the least used. The Contributor descriptor was used a few times and only for the description of external parties, while group members were identified as Author. In all cases, when the raw data belonged to other institutions or researchers, the descriptor Source was used, sometimes upon recommendation.

Moreover, the results show that in general the metadata was improved after the feedback provided by the curator, particularly with a more detailed description by increasing the number of descriptors used. The Date descriptor, which can be captured in every domain, was added fifteen times by the researchers. We looked into this and tried to understand why the date information was missing in some cases. We concluded that the descriptor meaning was too ambiguous for the researchers, who often asked if the date was for the creation of the dataset or for the start of the project. Based on this feedback we added a more precise definition and made sure the Date information was present in all metadata records by the time the datasets were deposited.

Although some descriptors have the same number of occurrences, some of them have also been corrected by the curator to improve the quality of the description. The dataset Title was edited in some circumstances, for instance from "*Capsule videos*" to a more accurate "*Red Lesion Endoscopy Dataset*" after the recommendations by the curator.

Table 3. Descriptor occurrences before and after curator mediation

Descriptor	Descriptors used by researchers	Descriptors after curator feedback
Title	21	21
Author	21	21
Author email	21	21
Description	20	21
Format	20	21
Tags	19	21
License	16	21
Coverage (Temporal)	16	17
Type	16	20
Date	15	21
Coverage (Spatial)	14	15
Language	14	16
Publisher	12	21
Relation	12	15
Source	10	10
Contributor	6	6
Format.Extent (File size)	5	21

Some difficulties felt by the researchers had an influence on their decisions. For instance, the Format.Extent descriptor was perceived as ambiguous and in many cases researchers did not provide it. In general, researchers state that they do not want to spend too much time in the description so they prioritize other descriptors. Data organization issues also emerged, as researchers wanted to deposit datasets as one zip file, so the format was described only as zip. Sometimes that made sense, but in most cases we advised them to deposit files separately. This makes the contents of the dataset more explicit and favours detailed metadata at file level, making the data easier to interpret. Thus, we edited the information in descriptor Format by replacing *.zip* with the extensions of the corresponding files, *.tiff* and *.py*, for example.

The difference between descriptors Format and Type is also a known issue. To clarify this, we used examples from the data repository. Examples of format are *.jpg, .txt, .xls* whereas type can be *Measurements*, *Events*, or *Entity Annotated News*. In cases where it was difficult to distinguish between the description of the dataset and the description of individual files in the dataset, we have explained that the dataset includes information about all the contained files, and each file can have its own specific description.

Some questions arised concerning controlled vocabularies. For example, the descriptor Language was filled inconsistently with strings such as "*Português*" or "PT". This is a case where a controlled vocabulary can help to normalize the description and facilitate the introduction of the text. In this example we could

use Dublin Core Encoding Schemes (see footnote 9), namely ISO 639-2: Codes for the representation of names of languages.

Another question related to controlled vocabularies has to do with the License. By default, in CKAN values for this descriptor come from a list of licenses, as a closed controlled vocabulary. In some cases, it was necessary to add the Dublin Core descriptor License to add specific information, not on the list, e.g. "*Creative Commons Attribution-NonCommercial-NoDerivs 2.0 Generic (CC BY-NC-ND 2.0)*".

Controlled vocabularies are useful tools and could be used in specific descriptors as custom fields in INESC TEC repository. In the context of the TAIL project, we developed and implemented them on the Dendro platform in the Hydrogen Production domain. Preliminary results on the use of controlled vocabularies showed that they contribute to simplify the description process and to reduce metadata errors [11].

In general, many questions were raised in each collaboration. Researchers were curious about data citation, the overall data repository functionality, the limits on file upload size, the inclusion of sensitive data. The availability of the repository in the long run was also a concern for them. Furthermore, researchers inquired about data preservation, demonstrating a sense of awareness about this RDM dimension.

6 Feedback from Researchers

The experience of getting researchers to participate in metadata-related activities has shown that it is difficult to anticipate RDM scenarios and researchers' expectations. Although most researchers are not familiar with metadata concepts, let alone metadata standards, they show different levels of awareness, thus requiring us to adapt our approach to each domain. The data deposit process was adapted to researchers' perspectives and agendas. Each collaboration took from one to more than two months. It was important to adopt flexible strategies to gradually involve researchers. To accomplish that we systematically gathered feedback from them, and adjusted or approach to their contexts. We noticed that more experienced researchers showed greater interest, and took advantage of the collaboration to deepen their knowledge in RDM. The collaboration has branched, in some of the groups, to the definition of a data management plan, to the creation of metadata models for their domain, and to data description in the Dendro platform, along with experiences with other tools besides the repository platform.

Even considering some challenges, most researchers were willing to deposit their data, recognizing the advantages. Once the deposit was completed, we asked researchers to comment on the overall experience and its impact on their RDM awareness. To this end, an informal email was sent with several questions about the experience. Their opinions were written in Portuguese, which we freely translate here.

Some researchers confirm that their datasets were shared by others and cited in scientific articles: "*Yes, I have shared my dataset several times with other*

researchers and for project proposals"; "*We have used the dataset link in our papers*", yet they point out the lack of a mechanism to show information about the downloaded datasets: "*I think our data was not reused yet (or, if they were, we are not informed)*", therefore "*It would be useful to have some idea of how many times the data was downloaded*". Currently, the repository does not provide download statistics.

Most people state that with each deposit they have gained more confidence and motivation: "*At first I found it hard, but as we go it becomes much easier and simpler*". Moreover, some of them highlight the importance of having assistance: "*The process itself is a bit time consuming. The choice of descriptors is not something for which we are oriented and the support of the curator is fundamental here*". In other words, they have improved their skills in describing data, each time the description process takes less time and fewer corrections, and researchers become more engaged and independent in RDM activities.

In most cases researchers showed their datasets to others, promoting the INESC TEC data repository at the same time - "*Yes, once or twice I have recommended the deposit of data in the repository*". They acted as intermediaries between our team and other interested parties - "*Yes, I already gave your contact to one of my postdocs, who has seen my data in the repository and showed interest in depositing another kind of data (laboratory measurements)*".

In brief, our experience provides preliminary insight on the diversity of needs, issues, and motivations to publication, but most importantly to the level of awareness researchers have to RDM and metadata at our institution. With this in mind we are adapting the data preparation and deposit phases accordingly. Notwithstanding, Dublin Core proved to be a suitable metadata standard to initiate researchers in RDM activities.

7 Conclusion

The use of the Dublin Core descriptors as the basis for metadata on the INESC TEC data repository was assumed as a first step to involve researchers in data description. This is also regarded as the starting point in RDM training activities, leading researchers to understand metadata terms and standards and familiarizing them with RDM tools. In order to further our approach and the kind of collaboration described here, we still have to address specific data description requirements to improve the overall quality of the metadata. This will probably lead to an extension of the proposed metadata form. Moreover, given the experience gathered with these examples, we have to continue our work and focus on helping researchers make their data fit for reuse.

Although the Dublin Core elements have provided satisfactory results and good feedback, they have also revealed cases where the metadata requirements, or researchers' expectations, are not quite fulfilled. For instance, after the deposit of the Information Science dataset, we kept improving the metadata record, based on more detailed description contributed by the researcher, using some elements from the Data Documentation Initiative standard.

A researcher from the biomedicine domain decided not to deposit the data unless access restrictions were implemented. This example motivated the definition of new configurations in the repository, to address the requirements of sensitive data. It also shows how researchers can provide a critical view of the repository, when considering their priorities and the features of their data.

As the deposit process in the INESC TEC repository proceeds we will gather further insight on domain-specific requirements. Future work will address those requirements, with incremental adjustments to our dataset deposit form, allowing researchers to choose from or to add more descriptors. In addition, work also proceeds on workflows with more substantial involvement of researchers and more work on the definition of metadata models. Flexible metadata tools, like the Dendro platform, are essential to accommodate domain-specific requirements prior to data deposit, leading to richer metadata but also requiring a deeper involvement of the researchers.

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. Yulia Karimova is supported by research grant SFRH/BD/136332/2018, provided by the FCT - Fundação para a Ciência e a Tecnologia.

References

1. Amorim, R., et al.: A comparison of research data management platforms: architecture, flexible metadata and interoperability. Univers. Access Inf. Soc. **16**, 851–862 (2017). https://doi.org/10.1007/s10209-016-0475-y
2. Assante, M., et al.: Are scientific data repositories coping with research data publishing? Data Sci. J. **15**, 6 (2016). https://doi.org/10.5334/dsj-2016-006
3. Bishoff, C., Johnston, L.: Approaches to data sharing: an analysis of NSF data management plans from a Large Research University. J. Libr. Sch. Commun. **3**(2), eP1231 (2015). https://doi.org/10.7710/2162-3309.1231
4. Castro, J.A., et al.: Involving data creators in an ontology-based design process for metadata models. In: Developing Metadata Application Profiles, pp. 181–214. IGI Global (2017). https://doi.org/10.4018/978-1-5225-2221-8.ch008
5. European Commission: Accompanying the document Proposal for a Directive of the European Parliament and of the Council on the re-use of public sector information. SWD/2018/145 final - 2018/0111 (COD). Brussels (2018). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=SWD%3A2018%3A145%3AFIN
6. European Commission: Horizon 2020. Work Programme 2016 - 2017. Annex L. Conditions related to open access to research data (2017). https://ec.europa.eu/research/participants/data/ref/h2020/other/wp/2016-2017/annexes/h2020-wp1617-annex-ga_en.pdf
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. Farnel, S., Shiri, A.: Metadata for research data: current practices and trends. In: International Conference on Dublin Core and Metadata Applications, pp. 74–82 (2014). http://dcpapers.dublincore.org/pubs/article/view/3714
9. Gartner, R.: Metadata becomes digital. Metadata, pp. 27–39. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40893-4_3
10. Hudson Vitale, C.R.: The current state of meta-repositories for data. In: Johnston, L.R. (ed.) Curating Research Data, Volume One: Practical Strategies for Your Digital Repository, pp. 251–261. Association of College and Research Libraries, Chicago (2017)
11. Karimova, Y.: Vocabulários controlados na descrição de dados de investigação no Dendro. Universidade do Porto, Faculdade de Engenharia (2016). http://hdl.handle.net/10216/85221
12. Karimova, Y., Castro, J.A., da Silva, J.R., Pereira, N., Ribeiro, C.: Promoting semantic annotation of research data by their creators: a use case with B2NOTE at the end of the RDM workflow. In: Garoufallou, E., Virkus, S., Siatri, R., Koutsomiha, D. (eds.) MTSR 2017. CCIS, vol. 755, pp. 112–122. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70863-8_11
13. Lee, D.J., Stvilia, B.: Practices of research data curation in institutional repositories: a qualitative view from repository staff. PLoS ONE **12**(3), 1–44 (2017). https://doi.org/10.1371/journal.pone.0173987
14. Qin, J., Ball, A., Greenberg, J.: Functional and architectural requirements for metadata: supporting discovery and management of scientific data. In: Proceedings of the DCIM International Conference on Dublin Core and Metadata Applications, pp. 62–71 (2012). http://dcpapers.dublincore.org/pubs/article/view/3660
15. Qin, J., Li, K.: How portable are the metadata standards for scientific data? A proposal for a metadata infrastructure. In: Proceedings of the International Conference on Dublin Core and Metadata Applications, pp. 25–34 (2013). http://dcpapers.dublincore.org/pubs/article/view/3670
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, juldez, pp. 256–264 (2016). https://www.bad.pt/publicacoes/index.php/cadernos/article/view/1603
17. Rocha, J., Ribeiro, C., Lopes, J.C.: The Dendro research data management platform: applying ontologies to long-term preservation in a collaborative environment. In: Proceedings of the 11th International Conference on Digital Preservation, iPRES (2014)
18. Sayogo, D.S., Pardo, T.A.: Exploring the determinants of scientific data sharing: understanding the motivation to publish research data. Gov. Inf. Q. **30**, 19–31 (2013). https://doi.org/10.1016/j.giq.2012.06.011
19. 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
20. Swan, A., Brown, S.: To share or not to share: publication and quality assurance of research data outputs. A report commissioned by the Research Information Network (2008). https://eprints.soton.ac.uk/266742/
21. Tani, A., Candela, L., Castelli, D.: Dealing with metadata quality: the legacy of digital library efforts. Inf. Process. Manag. **49**(6), 1194–1205 (2013). https://doi.org/10.1016/j.ipm.2013.05.003
22. 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

23. Vardigan, M., Heus, P., Thomas, W.: Data documentation initiative: toward a standard tor the social sciences. Int. J. Digit. Curation **3**(1), 107–113 (2008). https://doi.org/10.2218/ijdc.v3i1.45
24. Winn, J.: Open data and the academy: an evaluation of CKAN for research data management. In: IASSIST 2013, pp. 28–31, May 2013. http://eprints.lincoln.ac.uk/9778