







# Hands-On Data Publishing with Researchers: Five Experiments with Metadata in Multiple Domains

Joana Rodrigues<sup>(✉)</sup> , João Aguiar Castro<sup></sup>, João Rocha da Silva<sup></sup>,  
and Cristina Ribeiro<sup></sup>

Faculty of Engineering of the University of Porto, INESC TEC,  
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal  
joanasousarodrigues.14@gmail.com, joaoaguiarcastro@gmail.com,  
joacrosilva@gmail.com, mcr@fe.up.pt

**Abstract.** The current requirements for open data in the EU are increasing the awareness of researchers with respect to data management and data publication. Metadata is essential in research data management, namely on data discovery and reuse. Current practices tend to either leave metadata definition to researchers, or to assign their creation to curators. The former typically results in ad-hoc descriptors, while the latter follows standards but lacks specificity. In this exploratory study, we adopt a researcher-curator collaborative approach in five data publication cases, involving researchers in data description and discussing the use of both generic and domain-oriented metadata. The study shows that researchers working on familiar datasets can contribute effectively to the definition of metadata models, in addition to the actual metadata creation. The cases also provide preliminary evidence of cross-disciplinary descriptor use. Moreover, the interaction with curators highlights the advantages of data management, making researchers more open to participate in the corresponding tasks.

**Keywords:** Research data management · Data publication · Metadata · Dendro

## 1 Introduction

Current research is characterized by an unprecedented growth in the volume of data being produced, as powerful computational capabilities are available to even small research groups—the so-called long-tail of science [6]. Usually, small groups or individual researchers have very limited resources to ensure long-term availability of their data. As such, they need adequate research data management (RDM) practices supported by practical tools, so that the datasets they produce can be made available to others. This is especially important as more research funding agencies adhere to the European Commission’s Guidelines on FAIR Data

Management in Horizon 2020, which advocate for a set of principles to make data Findable, Accessible, Interoperable and Reusable [10].

Data publishing involves peer review, unique identification and semantic enrichment of datasets. Published data raises awareness of new research claims or findings, promotes reuse, and brings scholarly credit to data authors [15]. Research data are slowly becoming first-class research objects on par with traditional publications, but a solid data citation culture is required for this to become the norm [13].

Data reuse depends on the quality of the metadata associated to a dataset, because that is often the only information available for researchers to interpret the data and decide on their quality and relevance to their work. Also, the production of metadata for research datasets requires the involvement of data creators, as domains are diverse and specific knowledge of the domain is required.

Researchers in the long-tail may commit to data organisation and description, but will probably lack the expertise to create FAIR data and metadata and to publish them in the most suitable repository. On the other hand, data curators working on their own will probably not be able to provide the rich contextual information that enables reuse. This exploratory study focuses on practical aspects of research data publication, namely the importance of the collaboration between researchers and data curators. The process used here assumes that data creators have the assistance of a curator to formalize the knowledge of the data production context and to assist in data publishing.

The study follows five research groups in the description and publication of datasets, and considers in more depth the choice of metadata elements. It is supported on Dendro<sup>1</sup>, a data organisation and description platform that enables the combination of generic descriptors with those tailored for the research domains. Data are then published in a data repository—in this experiment, the B2SHARE service of the EUDAT infrastructure.

The five cases correspond to different domains. Social sciences are present with a case in Family Psychology and another in Information and Innovation Management. For science and technology there is a case of named entity recognition in Portuguese, one of automatic detection of hate speech in text, and of one machine vision with a multi-camera system.

The paper is organized as follows. Section 2 provides an overview of the requirements and issues concerning the adoption of metadata standards for research data, while Sect. 3 presents related work. The study configuration is described in Sect. 4, continued in Sect. 5 with the details of the five cases. The results of the data description experiments are explored in Sect. 6 and discussed in Sect. 7.

---

<sup>1</sup> <https://github.com/feup-infolab/dendro>.

## 2 Requirements for Research Metadata

The specialization of researchers makes them natural providers of domain-specific metadata, even more so in the long-tail of science [11]. As data creators, they have unique knowledge of the data production context, including domain-specific concepts and configurations used in the production of the data. Conversely, while the skills of data curators may ensure the correctness of the metadata from a formal point of view, a metadata record produced solely by an institutional data curator may not be comprehensive enough. Curators are not domain experts, and thus may not anticipate what a researcher needs to know about the dataset to reuse it. Thus, the collaboration between the data creator and the curator has the potential to generate both correct and comprehensive metadata records.

Under the Research Data Alliance initiative<sup>2</sup>, the Metadata Standards Catalog Working Group is working in a Metadata Directory<sup>3</sup> listing available metadata standards by domain, such as the Data Documentation Initiative (DDI)<sup>4</sup> for data description in the Social and Behavioral Sciences and the Darwin Core<sup>5</sup> for Biodiversity data. Moreover, the Directory also includes general standards that can be broadly applied to the scientific context, such as Dublin Core<sup>6</sup> for generic descriptive metadata, CERIF<sup>7</sup> for recording research activity and PROV<sup>8</sup> for data quality and reliability purposes.

Metadata elements are building blocks for a comprehensive data record. Together with identity and subject descriptors, information captured in metadata elements for aspects such as the temporal, geospatial and scientific context is essential to promote data reuse [9]. An analysis of nine scientific standards, considering domain, objective and architecture, showed that the ability to add new elements or modules to address domain-specific needs is a common requirement [18]. However, features such as simplicity and sufficiency are likely to be appreciated by researchers when describing their data, and should also be considered in the development of metadata models.

## 3 Related Work

The relevance of RDM is visible in the growing body of studies in this area, some with a focus on researchers' perspectives and practices regarding data organization and sharing, while others have a closer look at researchers' metadata practices.

<sup>2</sup> <https://www.rd-alliance.org/>.

<sup>3</sup> <http://rd-alliance.github.io/metadata-directory/standards/>.

<sup>4</sup> <http://www.ddialliance.org/>.

<sup>5</sup> <http://rs.tdwg.org/dwc/index.htm/>.

<sup>6</sup> <http://dublincore.org/>.

<sup>7</sup> <http://www.eurocris.org/cerif/main-features-cerif/>.

<sup>8</sup> <http://www.w3.org/2001/sw/wiki/PROV/>.

A multinational survey of data sharing and reuse found that researchers are willing to share data and reuse data created by others, despite barriers slowing data sharing that may be overcome with user-friendly tools [14]. Interviews with researchers also show that journal requirements and normative pressure at the discipline level, together with the perceived benefits at the individual level, have positive effects in data sharing behaviors. On the other hand, realizing the effort involved in data sharing has a significant negative impact [7].

A study regarding barriers to data reuse suggests that the ease of access and the interoperability are important initial conditions for successful reuse. Also, while some lack of data documentation can be overcome by more experienced researchers, it is still an obstacle to data reuse [19]. Another study in the field of evolutionary biology looked into data organization and concluded that all participants used some kind of metadata or a personally created organization scheme [16], but are mostly unfamiliar with data documentation with reuse in mind [2].

Data descriptions produced by researchers were found to be more focused on the details rather than in general features when compared to those made by information specialists. This highlights the difference between descriptions meant for personal archives and those meant for data repositories, which tend to have reuse in mind [17].

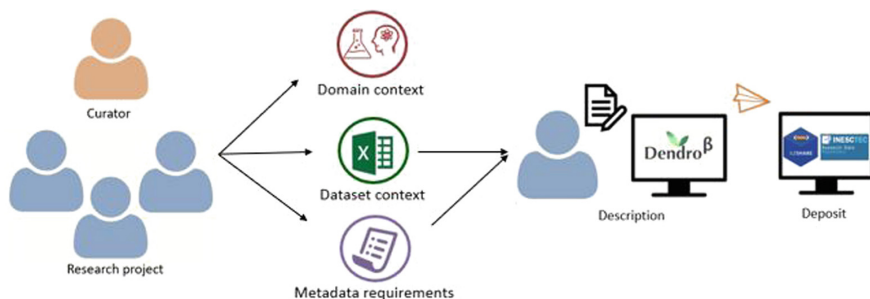
Studies on the relation of researchers with RDM and metadata provide evidence of the need to bridge the gap between researchers and data curators, while guidelines and tools to support metadata and RDM are being created. A good example of a metadata tool developed with researchers in mind is a framework to record minimal information for data reuse in the geobiology domain, resulting from stakeholder engagement [8].

## 4 Methodology

For this exploratory study we collected five cases by challenging researchers at the University of Porto (U.Porto) to publish datasets from their projects, either recently finished or soon to be completed. These datasets were likely to lose their publication opportunity soon, due to the re-assignment of researchers to other projects. After an introduction to some of the benefits of having their data published, such as the association of a DOI, researchers demonstrated motivation to describe and make their data available to others. Our approach explores a set of techniques consistent with participatory research methods [4], namely by combining casual meetings, interviews, content analysis and data description. The flexibility to combine these techniques is essential in this case not only due to the busy schedules of the participants, but also because there are many domain-specific cultures and different points of intervention to account for in RDM [5].

Figure 1 shows the data publication workflow, which includes the description process. Although intended as a systematic one, the techniques applied in the cases varied according to the availability of researchers and also their level of

awareness of RDM. The process includes contacts with researchers to select key concepts and identify metadata requirements, a description phase where the selected metadata elements are used, and ends with data deposit in a repository.



**Fig. 1.** Data publication process

In each case, we have conducted casual meetings with elements from the project teams to reach an agreement on the activities required to prepare data for publication. To assess the domain, the dataset and the metadata requirements, we conducted interviews or content analysis sessions—at least one in each case. The choice of method was determined by the availability of the researchers: while interviews take time from the researchers’ agendas, content analysis can be performed by the curator without the researcher. In this case papers and other technical documents are used to provide the context, and the resulting metadata model requires the validation of the researchers.

After assessing the metadata requirements in the five cases, we decided to include a selection of DDI elements in the set of descriptors available in Dendro, the ontology-based platform used for the data description activities [12]. In Dendro, descriptors are drawn from generic standards such as Dublin Core, existing domain-specific ontologies, or other ontologies that built in collaboration with researchers, such as Hydrogen Generation, Vehicle Simulation, and Gravimetry, introduced as a result of previous partnership [1].

Most of the DDI elements used were already part of the DDI-RDF Discovery vocabulary<sup>9</sup>, so we could load that ontology directly into Dendro. Those that were not in DDI-RDF Discovery were captured as a separate ontology, also loaded into the platform.

Dendro integrates with several data repositories and platforms (e.g. DSpace, figshare, CKAN, EUDAT’s B2SHARE) [12]. This is used in the final step of the process, where a package containing the dataset and its metadata is submitted to the B2SHARE repository<sup>10</sup> and in some cases also to the institutional CKAN-powered research data repository at INESC TEC<sup>11</sup>.

<sup>9</sup> <http://rdf-vocabulary.ddialliance.org/discovery.html>.

<sup>10</sup> <https://b2share.eudat.eu/>.

<sup>11</sup> <https://rdm.inesctec.pt>.

This process allows researchers to create domain-specific metadata records that will be used in the data repositories to find and access the datasets, and also by other researchers to estimate their value. The use of multiple ontologies addresses metadata limitations often associated with generalist repositories that are meant to cover several research communities [3].

All data description sessions were performed in Dendro except one, where the researcher could not participate in person. In this case the curator team described the dataset, which was then validated by the researcher.

## 5 The Five Data Publication Cases

The workflow described above was applied to five projects in different research domains: two in social sciences (Psychology and Innovation), three in science and technology (Hate Speech, Multi-Cam, and NER).

### 5.1 Family Psychology (Psychology)

The goals of the underlying project are to understand how the dynamics between work and family are linked to the exercise of parenting, the parent-child relationship and the child's socioemotional development, and to examine how the relationship between teachers and children intersects with the parental roles for the socioemotional development of children in dual-earner families<sup>12</sup>.

The research team is collecting observational data from families with preschool children and explores them combining a cross-sectional study design with a longitudinal design. The raw data is organized in a database where subsets are then selected by researchers to work on a specific perspective or a certain reality.

A dataset containing processed data concerning children's emotions regulation, parents' work-family conflict and psychological availability, represented as descriptive statistics and organised in a table, was selected by the researchers to be published in B2SHARE<sup>13</sup>.

### 5.2 Automatic Detection of Hate Speech in Text (Hate Speech)

The goal of the Hate Speech project is to study the Internet and social networks, in particular for detecting hate speech posted online. The researcher annotated a dataset in Portuguese and built a classification system for types of hate speech using a hierarchical structure. The project aims to deliver tools for automatic detection of aggressive communications. Contributions include the definition of hate concepts, namely hate speech subtypes<sup>14</sup>.

The data was collected from Twitter, and manually annotated. The dataset contains 5,668 messages from 1,156 distinct users, and handles 85 classes of hate

<sup>12</sup> [https://www.fpce.up.pt/reconciliar/index\\_eng.html](https://www.fpce.up.pt/reconciliar/index_eng.html).

<sup>13</sup> DOI: <https://doi.org/10.23728/b2share.7b3c66dfa4df4a7f9ba04fbc30cfb8bc>.

<sup>14</sup> <http://hdl.handle.net/10216/106028>.

speech. The data types in the dataset are tweets and class taxonomies. Data were published in B2SHARE<sup>15</sup> and also in the institutional INESC TEC data repository<sup>16</sup>.

### 5.3 Multi-camera System for Automatic Positioning (Multi-Cam)

Taking into account the rapid evolution of multimedia platforms, this project explored the use of technological resources and applications in sports. The research focuses on the location of a ball in a 3D space field, as part of the application of computer vision to indoor team sports<sup>17</sup>. Several techniques for camera calibration and 3D reconstruction methods were studied and tested, resulting in the use of a limited number of conventional cameras. The published dataset<sup>18</sup> contains camera calibration data resulting from an acquisition protocol that considered all stages of camera calibration and different static and dynamic ball scenarios.

### 5.4 Named Entity Recognition (NER)

This project addresses the task of named entity recognition (NER), in the field of Natural Language Processing, and explores entity detection as an enabler for more complex tasks<sup>19</sup>, such as relation extraction or entity-oriented search, as applied for instance in the ANT search engine<sup>20</sup>.

The project evaluated existing NER tools to select the best approach and configuration for the Portuguese language, more specifically for institutional content. Results include a richer entity-oriented search experience with new information, as well as a better ranking scheme based on the additional context available to the search engine. The project also created several detailed manuals with systematic analyses of available tools. The dataset was published in the B2SHARE repository<sup>21</sup> and in the INESC TEC repository<sup>22</sup>.

### 5.5 Information and Innovation Management (Innovation)

This project takes an information management perspective on research and development, innovation and entrepreneurship, applying it to the knowledge transfer and the innovation process in the University of Porto. The data from an exploratory study allow the identification of internal and external agents, resources, the relations between actors and institutions, processes and flows,

<sup>15</sup> DOI: <https://doi.org/10.23728/b2share.9005efe2d6be4293b63c3cffd4cf193e>.

<sup>16</sup> <https://rdm.inesctec.pt/dataset/cs-2017-008>.

<sup>17</sup> <http://hdl.handle.net/10216/88168>.

<sup>18</sup> DOI: <https://doi.org/10.23728/b2share.b89c998e26674e8eba8b263c8b4f3a2e>.

<sup>19</sup> [https://www.linguateca.pt/aval\\_conjunta/HAREM/harem\\_ing.html](https://www.linguateca.pt/aval_conjunta/HAREM/harem_ing.html).

<sup>20</sup> <http://ant.fe.up.pt/>.

<sup>21</sup> DOI: <https://doi.org/10.23728/b2share.93f011314ce24391a4c317779ccf8068>.

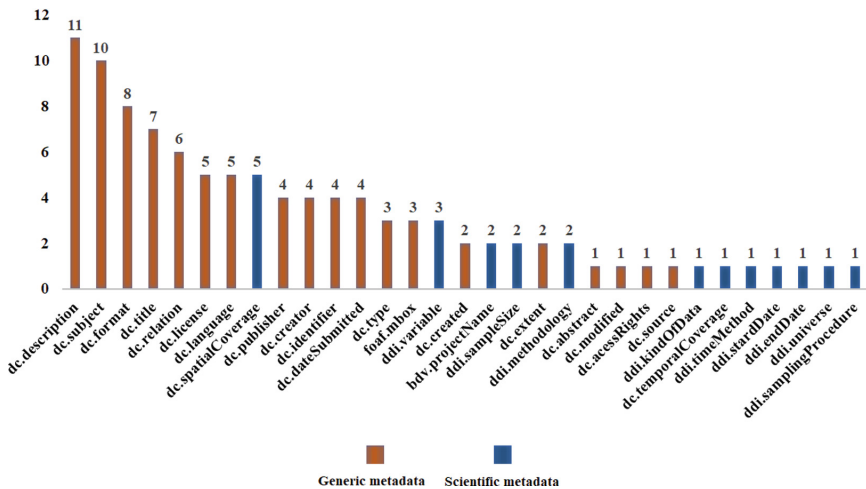
<sup>22</sup> <https://rdm.inesctec.pt/dataset/cs-2017-005>.

and the main inputs and outputs. The project created a model of innovation indicators in an academic context and fitted it to the University of Porto<sup>23</sup>.

The published dataset contains innovation indicators of several Portuguese institutions, used to support the development of the model. Data from this case is published in B2SHARE<sup>24</sup> and in INESC TEC<sup>25</sup>.

## 6 Data Description Experiment Results

Overall, researchers were satisfied with the selection of generic metadata elements. Some went beyond that and produced a more comprehensive metadata record using descriptors from their scientific context, such as temporal and spatial coverage and data collection procedures.



**Fig. 2.** Descriptors used by researchers

In the Psychology case, two researchers participated in the data description session with a data curator, who made recommendations on the use of DDI elements, specific for the Social Sciences. The researchers were autonomous in the selection of descriptors in Dendro, talking to each other and progressing without the data curator intervention. They have selected metadata elements for temporal context, namely `timeMethod` and `temporalCoverage`, and methodological information, e.g. `kindOfData`. In addition to these descriptors they used more generic descriptors such as the point of contact (`mailbox`).

<sup>23</sup> <http://hdl.handle.net/10216/113262>.

<sup>24</sup> DOI: <https://doi.org/10.23728/b2share.7d60f3262d2e49e7a118911077d99eff>.

<sup>25</sup> <https://rdm.inesctec.pt/dataset/ii-2018-001>.



The Hate Speech case was the only one in which the researcher was not directly involved in the description, communicating to the curators a preference for a generic and clear one. The result, based on content analysis, was proposed by the data curator and validated by the researcher. Selected descriptors were **description**, **subject** and **title**, as well as **format**, **dateCreated**, **spatialCoverage** and **relation**. The **mailbox** descriptor was used for the email contact of the researcher.

The Multi-Cam case has the largest data volume, and there was more concern with their organization and structure. The researcher who collected the data was no longer associated to the group, and two members of the team described the data in separate sessions, providing a complementary validation. They chose mostly generic descriptors, namely **creator**, **description**, **title**, **language**, **subject**, **format**, **relation**, and **type**. Spatial and temporal coverage elements were also filled in. As in other cases, a descriptor for the contact with the project researcher (**mailbox**) was included.

The NER case has concentrated on Dublin Core descriptors: 19 of the 21 selected descriptors were from this vocabulary. The researcher considered that they were sufficient to give a correct description of their dataset. They include temporal context descriptors such as **temporalCoverage**, **dateCreated**, **dateModified** and **dateSubmitted**. In addition, administrative descriptors were selected, such as **creator** and **publisher** and content descriptors such as **title**, **description** and **subject**. It is interesting to note that the researcher selected the DDI descriptor **methodology**, considering it a basic descriptor. The same happened with descriptor **mailbox**.

Finally, the researcher in the Innovation case has training in Information Science and some previous knowledge of RDM, so the mediation of the data curator was limited to a brief presentation of the activity goals and of the RDM tools to support data publication. The researcher selected generic descriptors (Dublin Core) to contextualize the dataset, including **creator**, **publisher**, **title**, **subject**, **format**. Dublin Core elements for scientific context were also used, such as **spatialCoverage**, and **temporalCoverage**. In addition to these, the researcher also selected DDI descriptors for a more specialized description: **sampleSize**, **universe**, **methodology**, **startDate** and **endDate**. The name of the project (**projectName**) and the contact of the researcher (**mailbox**) were also chosen in this case.

Figure 2 has the distribution of descriptors used in the five cases. Overall, we noticed a repeated use of the **format** descriptor (8 times). This shows that researchers consider it important to refer to the format in which the data are made available, as it can condition their visualization and reuse by others and even by themselves. Note that descriptors can be repeated; 4 occurrences of **format** are from the NER case.

After the datasets were deposited in B2SHARE, we kept track of user interactions. Table 1 shows the number of downloads and annotations made for each of the datasets exported to B2SHARE, from the date of the deposit (between December 2017 and January 2018) until October 4, 2018. As each dataset

Table 1. User interactions with datasets in B2SHARE

Case	Downloads	Annotations
<b>Psychology</b>	<b>14</b>	<b>1 comment</b>
dataset - table 1.docx	4	"Very interesting dataset. The description in the initial interface caused me interest, however when I opened the file I was in need of seeing more information ... maybe it was a short dataset. However the description we see in the file "(Re) conciliar.txt" informs us about extremely important details and also gives us the indication of the name of the project, which allows to research more about it. It is good to have tools like B2NOTE to share opinions between data producers and users."
(Re)conciliar.json	1	
(Re)conciliar.rdf	5	
(Re)conciliar.txt	3	
(Re)conciliar.zip	1	
<b>Hate Speech</b>	<b>10</b>	<b>0</b>
annotator_classes.csv	2	
dataset_dummy_classes.csv	1	
graph_hierarchical_classes.csv	1	
Portuguese Hate Speech Twitter Dataset.json	2	
Portuguese Hate Speech Twitter Dataset.rdf	2	
Portuguese Hate Speech Twitter Dataset.txt	2	
Portuguese Hate Speech Twitter Dataset.zip	0	
README.txt	0	
<b>Multi-Cam</b>	<b>8</b>	<b>0</b>
Multicamera System for Automatic Positioning of Objects in Game Sports.json	3	
Multicamera System for Automatic Positioning of Objects in Game Sports.rdf	0	
Multicamera System for Automatic Positioning of Objects in Game Sports.txt	2	
Multicamera System for Automatic Positioning of Objects in Game Sports.zip	3	
<b>NER</b>	<b>3</b>	<b>2 comments</b>
HAREM NER Models for OpenNLP, Stanford CoreNLP, spaCy, NLTK.json	0	"The information associated with each file is quite limited. there are many descriptors that could be shown for all datasets, for example the license, and for all files, for example the format."
HAREM NER Models for OpenNLP, Stanford CoreNLP, spaCy, NLTK.rdf	0	
HAREM NER Models for OpenNLP, Stanford CoreNLP, spaCy, NLTK.zip	3	
nlTK.zip	0	
open-nlp.zip	0	
spacy.zip	0	"I would like to find usage examples for the pre-trained models for each of the NER tools."
stanford-corenlp_Copy_created_1510592294091.zip	0	
stanford-corenlp.zip	0	
<b>Innovation</b>	<b>4</b>	<b>2 comments</b>
dataset.xlsx	2	User 1) "Dataset of enormous importance, I would like to see more work developed in this area. Congratulations on the investment in these issues, surely these data will serve many other researchers who want to invest in this area. Certainly I will reuse your work for future investigations. I will recommend this work to colleagues in the field of information management, in particular innovation indicators in the academic context. I'm happy to be able to talk to Fabio more "informally" through tools like B2Note."
U.InovAcelerator.json	0	
U.InovAcelerator.rdf	0	
U.InovAcelerator.txt	1	
U.InovAcelerator.zip	1	User 2) "Some pages, used as drafts, could be removed as they don't contain much. Also, referencing cells between pages (opposed to copy/paste) could help with automated parsing."

consists of several files (\*.zip, \*.txt, \*.rdf, \*.json), the interactions are also provided at the file level.

A total of 39 downloads are recorded for the five cases. Only Psychology had downloads in all the files associated with the dataset, and was also the case with the highest number of downloads. The opposite happened with the NER case, with only 3 downloads of a single file. The remaining cases had a similar number of downloads, which were distributed by the various files.

As for the annotations, 5 were recorded, all in the comment format. These comments were made through B2NOTE, a EUDAT service. This service provides an environment where any interested party (authors, researchers and others) can improve the description using comments, keywords or tags.

A user remarked that the Psychology data were interesting, but noted that they expected to have more data in the dataset. The usefulness of the complementary metadata in the \*.txt file was also noted, namely the name of the project.

In the NER case, two comments were made by the same user. The first is about the limited information associated with each file. The user added that descriptors such as “format” and “license” should be part of the metadata of all datasets. The second comment was more specific and had to do with the user expectations regarding the dataset.

Finally, the Innovation case has two comments by different users. The first highlights the importance of that dataset, as an incentive for more studies in this area, along with the intention to reuse and promote it. The other comment is a recommendation on how to improve the organization of the data.

## 7 Discussion

Regardless of their fields of expertise, researchers involved in this study have developed a sense of awareness and motivation towards RDM. After the first interactions they revealed curiosity in the data publishing process, and noticed that data management skills could contribute to improve their work environments and career opportunities. Our contacts in the five projects in the experiment were junior researchers. This is not surprising, given that they are closer to data production and also, as they are still developing their research routines, more open to introduce RDM practices in their schedules. We expect senior researchers to be attracted by the advantages of data publication as well. In further work, we look forward to include them in activities related to metadata quality and data reuse.

The Psychology researchers had no data description experience but, after some conversations, became familiar with the proposed task quite easily. In this case the researchers considered DDI convenient for their data. The importance of metadata elements that use terminology used in their routines was visible. The Innovation case was the one in which the researcher worked more autonomously, due to previous training in the area of metadata and data management. The case of NER generated the most debate between data curators and researchers.

Researchers in this project were already aware of the importance of the management of research data and, therefore, they had clear ideas on the descriptors they wanted to use prior to the data description session. Furthermore, they anticipated the potential for data reuse in their community, motivating them to make a detailed description.

Overall, the cases that seem to produce metadata records with more scientific elements are those that either have relied on the effective participation of data curators or already had a domain-specific vocabulary such as DDI.

A common feeling with researchers is that the RDM activity within their projects can have positive effects on data quality, allowing them to record details about the data collection process even prior to considering data reuse. Nevertheless, discussing metadata with researchers is never a trivial task, primarily because the boundaries between data and metadata are not always clear cut. This reinforces the importance of communicating through practical metaphors they can relate to their practice, namely those from the traditional publication workflow. Since most repositories use Dublin Core elements, those are also easier to understand by average repository users. This may have resonated in the systematic selection of Dublin Core metadata elements in the descriptions.

The systematic use of temporal and geospatial elements, like `timeMethod`, `endDate`, `startDate`, `temporalCoverage` and `spatialCoverage`, as a whole, reveals that researchers are inclined to register the temporal and spatial dimensions of their projects. Such elements locate the data in space and time, which is important for discovery and for reuse.

We observed that the `methodology` descriptor is easy to understand by researchers and suitable to describe the data collection approach independently of the research domain, if more specific descriptors are not available. Yet, only two researchers have filled in the `methodology` element. A possible solution is to consider this descriptor as a core element for generic RDM data description platforms and repositories. This will change the common practice of including methodological aspects in the more generic `description`.

## 8 Conclusions and Future Work

The main goal of this line of work is to raise awareness of the importance of RDM among researchers, across research domains and data types. The hypothesis here is that concrete tasks are required to illustrate the difficulties and the rewards in RDM. In the five cases presented, researchers participated in the whole process, from the selection of the metadata models, through the description up to deposit and publication. The interactions summarized in Table 1—especially the comments—show the importance of the domain-specific metadata. In the Psychology case, for instance, the end user mentioned the importance of the metadata to interpret a dataset that seemed limited at first, with respect to their expectations. Also, in the case of NER, one user said that they hoped to find more metadata, a clear indication that the metadata is relevant to the interpretation of the data.

A selection of complete cases such as those explored in this study can act as a motivator for other researchers, at our institution and elsewhere. We often observe that people ask for practical examples of data publication, especially in related domains, and also take pride in having their case included as an example. There is also a pressing need for more data to reach the publication stage so RDM can become an established practice at institutional level.

With this study we aim to provide more insight on the collaborative approach between researchers and curators. However, such an approach requires an evaluation, which depends on more data publishing cases and more evidence of end user feedback regarding the use of metadata in dataset dissemination and reuse.

Some researchers are already quite aware of the importance of good RDM practices. In the case of Family Psychology, the researchers included the publication of their data in B2SHARE in the project reports. They consider that this publication is a valuable result of the project.

Besides the 5 cases described here, more data publication scenarios are being explored using this method. With a larger number of cases, more generic observations are expected. But not all people are easily engaged: some of our tentative contacts declare that there is no need to describe data, given that colleagues from the same area will have no difficulty in understanding them.

Data description experiments are a learning process for both data curators and researchers, and are valuable to build trust between them as RDM stakeholders. As data curators, lacking in disciplinary expertise, we had to rely on the descriptions that were provided by domain experts. Even so, the question remains whether there is potential for the selection of more domain-specific descriptors.

The combination of efforts between researchers and data curators is expected to result in higher-quality metadata, but more work and considerably more time are required to evaluate this approach to data description. Testing for reuse is particularly challenging. According to participants in the Psychology case, to anticipate reuse scenarios is a difficult exercise. Moreover, reuse experiments depend on observations over an extended period of time.

Although RDM is increasingly present in the daily work of researchers, it is still necessary to raise awareness of these issues and to promote initiatives to make data repositories grow. Training actions are a good strategy, provided that researchers are involved as active participants, dealing with RDM in their own domains, solving their problems and contributing to the data management process. These initiatives are essential for researchers to become proactive in data management and to take more advantage of the collaboration with data curators.

**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.

## References

1. Castro, J.A., et al.: Involving data creators in an ontology-based design process for metadata models. In: *Developing Metadata Application Profiles*, pp. 181–213 (2017). <https://doi.org/10.4018/978-1-5225-2221-8.ch008>
2. Akers, K.G., Doty, J.: Disciplinary differences in faculty research data management practices and perspectives. *Int. J. Digit. Curation* **8**(2), 5–26 (2013). <https://doi.org/10.2218/ijdc.v8i2.263>. ISSN 1746–8256
3. Assante, M., et al.: Are scientific data repositories coping with research data publishing? *Data Sci. J.* **15** (2016). <https://doi.org/10.5334/dsj-2016-006>
4. Bergold, J., Thomas, S.: Participatory research methods: a methodological approach in motion. *Forum Qual. Soc. Res.* **13**(1) (2012). <https://doi.org/10.17169/fqs-13.1.1801>
5. Cox, A.M., Pinfield, S., Smith, J.: Moving a brick building: UK libraries coping with research data management as a “wicked” problem. *J. Librarianship Inf. Sci.* **48**(1), 3–17 (2016). <https://doi.org/10.1177/0961000614533717>
6. Heidorn, P.B.: Shedding light on the dark data in the long tail of science. *Libr. Trends* **57**(2), 280–299 (2008). <https://doi.org/10.1353/lib.0.0036>. ISSN 1559–0682
7. Kim, Y.: Institutional and individual influences on scientists’ data sharing behaviors. *The School of Information Studies - Dissertations Paper 85 3.1*, p. 304 (2013). <https://doi.org/10.1002/meet.14505001093>
8. Palmer, C.L., et al.: Site-based data curation based on hot spring geobiology. *Plos One* **12**(3), e0172090 (2017). <https://doi.org/10.1371/journal.pone.0172090>
9. Qin, J., Ball, A., Greenberg, J.: Functional and architectural requirements for metadata: supporting discovery and management of scientific data. In: *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pp. 62–71 (2012)
10. European Commission: Directorate-General for Research and Innovation. *Guidelines on FAIR Data Management in Horizon 2020* (2016)
11. Rice, R., Haywood, J.: Research data management initiatives at University of Edinburgh. *Int. J. Digit. Curation* **6**(2), 232–244 (2011)
12. 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. *Int. J. Digit. Libr.* (2018). <https://doi.org/10.1007/s00799-018-0238-x>. ISSN 1432–300
13. Silvello, G.: Theory and practice of data citation. *J. Assoc. Inf. Sci. Technol.* **69**(1), 6–20 (2018). <https://doi.org/10.1002/asi.23917>
14. Tenopir, C., et al.: Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE* **10**(8), 15 (2015). <https://doi.org/10.5061/dryad.1ph92>
15. Thanos, C.: Research data reusability: conceptual foundations, barriers and enabling technologies. *Publications* **5**(1), 16 (2017). <https://doi.org/10.3390/publications5010002>
16. White, H.C.: Considering personal organization: metadata practices of scientists. *J. Libr. Metadata* **10**(2–3), 156–172 (2010). <https://doi.org/10.1080/19386389.2010.506396>
17. 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>

18. Willis, C., Greenberg, J., White, H.: Analysis and synthesis of metadata goals for scientific data. *J. Am. Soc. Inf. Sci. Technol.* **63**(8), 1505–1520 (2012). <https://doi.org/10.1002/asi.22683>
19. Yoon, A.: Red flags in data: learning from failed data reuse experiences. *Proc. Assoc. Inf. Sci. Technol.* **53**(1), 1–6 (2016). <https://doi.org/10.1002/pa2.2016.14505301126>. ISSN 23739231