Getting in touch with metadata: a DDI subset for FAIR metadata production in clinical psychology

João Aguiar Castro; Joana Rodrigues; Paula Mena Matos; Célia Sales; Cristina Ribeiro[1]

Abstract

When addressing metadata with researchers, it is important to use models that include familiar domain concepts. In the Social Sciences, the Data Documentation Initiative (DDI) is a well-accepted source of such domain concepts. To create data and metadata, that is Findable, Accessible, Interoperable, and Reusable (FAIR), it is necessary to establish a compact set of DDI elements that meet project requirements and are likely to be adopted by researchers inexperienced with metadata creation. Over time, we have engaged in interviews and data description sessions with research groups in the Social Sciences, identifying a manageable DDI subset. TOGETHER, recent Clinical Psychology project dealing with risk assessment for hereditary cancer, considered the inclusion of a DDI subset for the production of metadata that are timely and interoperable with data publication initiatives in the same domain. Taking the DDI subset identified by the data curators, we present a preliminary assessment of its use as a realistic effort on the part of the researchers, taking into consideration the metadata created in two data description sessions, the effort involved, and the overall metadata quality. A follow-up questionnaire was used to assess the perspectives of the researchers regarding data description.

Keywords

Research data management, metadata, FAIR, Data Documentation Initiative; Clinical Psychology

Introduction

The fast-paced growth of scientific production and the risk of data being permanently lost (Vines, 2014) have prompted funding agencies to define policies to promote data FAIRness and require Data Management Plans (European Commission, 2016a). With the interest in Research Data Management (RDM) globally on the rise (Perrier et al., 2017), researchers are increasingly aware of the need to develop adequate skills to organize their data. In this context, metadata production is an essential activity that involves all the stages of well-managed research data to ultimately enable reuse.

To promote Open Science and the adoption of the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles, the European Commission Expert Group on FAIR data recommended the development of cases to further engage communities and the provision of tools to make metadata production as easy as possible for researchers (European Commission, 2018). Moreover, the Libraries for Research Data Interest Group of the Research Data Alliance recognizes that direct training is an effective way to make people aware of the importance of data management best practices (Clare et al., 2019).

Our experimental work at the University of Porto, under the TAIL project[2], focused on the development of an RDM workflow that integrated a set of different tools depending on the requirements of researchers (Ribeiro et al., 2018). The TAIL project regarded researchers as core RDM stakeholders who needed straightforward workflows. In this sense, the availability of tools to support data organization and metadata creation at the beginning of research projects can be a determinant to improve data management practices. Thus, and in order to motivate small research groups with no time or funding for data curation, the TAIL team developed Dendro[3], an open-source platform designed to help researchers describe their data, fully built on Linked Open Data (Rocha da Silva, Ribeiro and Lopes, 2018). Dendro includes domain-specific metadata models to address disciplinary data description requirements (Castro et al., 2017).

The integration of tools in the research workflow designed during the TAIL project resulted from the assessment of requirements of several researchers from different domains, who were contacted individually over time. This enabled the team to obtain feedback from a diverse panel of researchers. Contact with researchers at the University of Porto had been previously established with a scoping study (Ribeiro and Fernandes, 2011) sent through the deans of its 15 schools in 2011. This scoping study can be regarded as a preliminary effort to poll the availability of researchers. During the TAIL project, some sessions were organized to disseminate the project among researchers. One of these sessions was targeted at a group of researchers affiliated with the Faculty of Psychology and Educational Sciences of the University of Porto (FPCEUP). Further contacts were made with two researchers working in family psychology and another working in clinical psychology, who had shown motivation to adopt measures leading to better practices and were starting a new project.

The work described in this paper results from the collaboration between the TAIL project and members of TOGETHER[4], a project in the psycho-oncology domain. TOGETHER ran from July 2018 to June 2021 and was a partnership between the FPCEUP and the Portuguese Oncology Institute (IPO-Porto), a state-run institute for healthcare and research. The goal of TOGETHER was to study the process of psychosocial adaptation of individuals and families enrolled in genetic counselling to assess and manage the increased risk of hereditary cancer syndromes. Over a period of two years, individuals enrolled for genetic testing and their families were monitored regarding psychological and relational variables, as well as their needs and preferences for care. A longitudinal design with a mixed qualitative and quantitative methodology included semi-structured interviews, self-report questionnaires, and documentary analysis of clinical records. The project followed a participatory approach, with a collaborative panel of users and professionals involved in methodological decisions and the interpretation and dissemination of results. This project was a first step toward a strategy of family-centered psychological care, indispensable in personalized preventive medicine of hereditary cancer.

Table 1: Overview of the TAIL and of the TOGETHER projects

TAIL	The TAIL project focused on providing researchers with adequate tools to organize, describe, and publish their data. The TAIL team developed an RDM workflow based on the integration of different tools taking into account the requirements of a panel of researchers built over time.
TOGETHER	Using a longitudinal mixed methods design, the purpose of the TOGETHER project was to study the process of psychosocial adjustment of both unaffected individuals undergoing genetic testing and their families. More specifically, it intended to analyze the mechanisms by which psychosocial factors and clinical factors shape psychosocial outcomes of Genetic Testing. The ultimate goal is to gather knowledge that informed an integrated family-centered care for inherited cancer syndromes.

In this paper, we focus on the preliminary steps in the development of a DDI-based ontology to support researchers to describe data. Our primary objective is to raise RDM awareness and simplify the description of data to promote sharing of project data. The cooperation between researchers and data curators is the backbone of our approach, so we aim to address the perspective of researchers when they become aware of data management tasks and have to reconcile them with the already demanding research work. This requires providing the support they deem necessary in adopting RDM practices, such as the organization and description of data in a first stage.

Our interactions with researchers showed that most are inexperienced with metadata and therefore, to achieve our objectives, we opted for a DDI subset that, for the sake of interoperability, takes into account the already DDI-based list of metadata fields recommended by the Interuniversity Consortium for Political and Social Research (ICPSR)[5], already DDI-based, since it stands out as the world´s largest social science data repository. Adopting the DDI subset in the early stages of research also meant that, if the group decided to make their data available, FAIR metadata would already be available by the time they proceeded to the deposit stage.

Next, we provide a brief overview of related works and of the Dendro platform used in the description tasks. Then, we proceed with a description of the steps in the development of an ontology based on the DDI subset, which is followed by a description of the activities performed in this work, including the interviews and how we prepared the data description sessions. Finally, we highlight the organization and metadata practices of interviewed researchers, the metadata created during the data description sessions, taking into account the number of descriptors filled in and the amount of time spent in the activity, and the feedback of participants regarding their experience in metadata production.

Support for researchers in metadata production

With their domain expertise and proximity to data collection, researchers have the potential to create highly detailed descriptions about their data, which makes them key stakeholders in FAIR metadata production. Even though data curators are experts at making sure that metadata follows a set of rules to ensure data is findable, accessible and interoperable, their limited ability to capture domain-specific knowledge in the metadata can hinder reusability.

From interviews carried out with 23 quantitative social scientists who failed data reuse experiences (Yoon, 2016), it was found that access and interoperability were chief primary conditions for successful data reuse, whilst understanding data documentation was less of an issue, at least for experienced researchers, though the process was still seen as challenging. The lack of support in reusing data was the most prominent issue in the reported failed data reuse experiences, making it necessary to establish support systems for those willing to reuse data. In another study, 13 social scientists were interviewed to assess which factors influenced the perceptions and experiences of researchers in attempts to reuse data. It was concluded that data documentation was, among other factors, an important enabling factor for data reuse (Curty, 2016). An institutional study conducted to evaluate data management skills and including, both graduate students and postdoctoral researchers, concluded that many researchers were frustrated when former colleagues left without providing annotations of the completed work. Consistent data description and organization were regarded as challenges given the different workflows, practices, and value concepts of individuals. A practical solution to address this limitation was the provision of a short description to enable group members to understand the research workflow (Wiley and Kerby, 2018).

By comparing the metadata created by researchers and information professionals, White (2014) found that researchers were more focused on the details and produced more granular metadata. However, the same study found that there was a difference between what was created for personal use and what was created in a formal deposit setting, as more descriptive metadata was added in the deposit stage. Moreover, a study with researchers from the Center for Embedded Networked Sensing also demonstrated that researchers rarely created documentation that was not directly tied to their own personal use, and therefore data sharing with users from outside of their immediate projects was rare (Mayernik, 2011).

Hence, it seems necessary to combine the skills of data curators and researchers to move the production of FAIR data and metadata forward. There are several initiatives to support metadata creation in a number of disciplines. For instance, the Research Data Alliance, in the context of the Metadata Standards Directory Working Group[6], lists available metadata standards by domain, such as DDI[7] for the Social and Behavioral Sciences and the Darwin Core[8] for Biodiversity. The Directory also includes domain-neutral standards for different functions, namely the Common European Research Information Format (CERIF)[9] for recording research activity and PROV[10] for data quality and reliability.

However, most standards target data description only at the end of the research workflow and their adoption by researchers can be hard (Qin and Li, 2013). According to Qin et al. (2012), it makes more sense to develop specific goal-oriented metadata schemes, as smaller and more specific schemes will likely increase their adoption by researchers. An analysis of several metadata standards corroborated

the idea that these do not follow principles of simplicity and sufficiency, since most of them lacked a minimal set of essential domain elements (Willis, Greenberg and White, 2012). As a result, disciplinary vocabularies for research data description are mostly underused, so there is a need to implement more effective processes for the adoption of vocabularies by research communities (European Commission, 2018).

In this context, we opted for a top-down approach to vocabulary development, with the selection of a set of essential domain elements tailored to describe data from the beginning of the research workflow. The goal was for the metadata model to fit the metadata needs of both the TOGETHER project and of similar projects, and for it to be simple enough to encourage researchers in their first experience describing data. Thus, we decided to develop a minimalist ontology having the DDI as the reference. Working with a DDI subset enabled an incremental build-up of the ontology according to researchers' needs, which were identified by engaging them in metadata production activities. This approach was made possible by the existence of a staging platform for data organization and description, which provided researchers with a training environment for metadata production and with the flexibility to combine the most relevant metadata elements for their data.

Dendro, a staging platform for data description

We sought to simplify and promote FAIR metadata production at the University of Porto by embedding the Dendro platform in the research workflow. Dendro (Rocha da Silva, 2016) was developed within the TAIL project and is an open-source, collaborative data organization platform that promotes the description of data from the moment of its creation. Dendro follows a file management structure that resembles popular cloud storage environments, with additional collaborative capabilities common in semantic wikis. The effort in creating metadata is reduced by the incremental description of data, since project members can add and fill in new metadata elements at different times to enrich the quality of the metadata records.

Figure 1 depicts Dendro´s data description user interface. The folder and file management panel is on the left, while some of the available vocabularies are accessible on the right-hand side. Each vocabulary has its own set of descriptors.

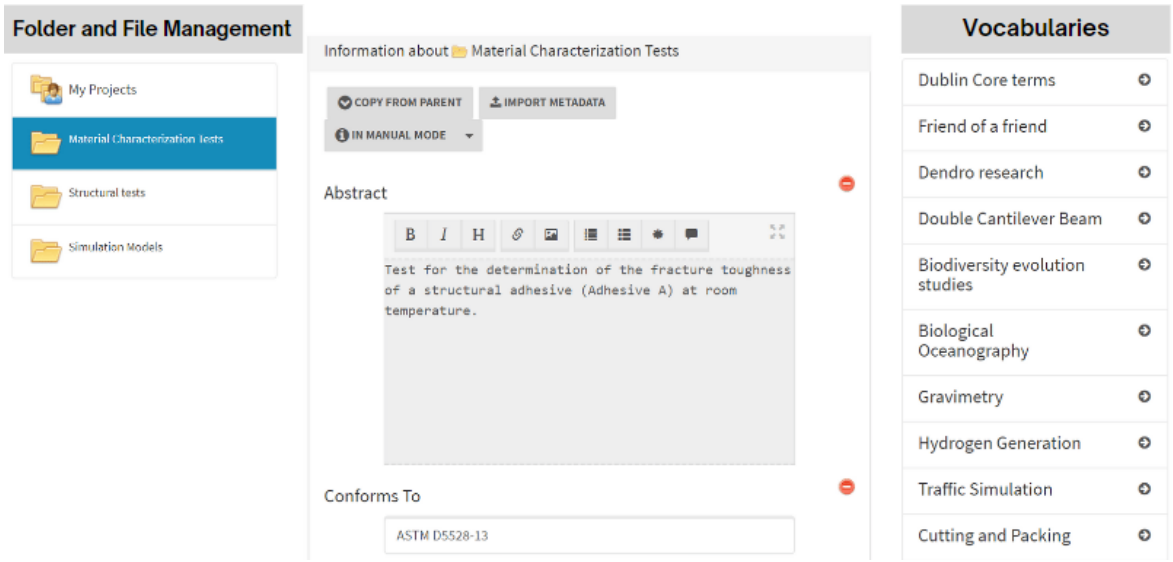

Figure 1: Dendro data description user interface

Dendro favors the compliance with the FAIR guidelines in several aspects. Findability is achieved by assigning persistent identifiers and by allowing the creation of rich metadata records. In Dendro, researchers can produce a versatile description for each dataset and have the flexibility to combine descriptors from multiple vocabularies.

These vocabularies can either be domain-specific or developed with the participation of researchers with whom the TAIL team collaborates. The latter mostly happens when there are no available standards for very particular applications such as the Double Cantilever Beam and the Hydrogen Generation vocabularies represented in Figure 1. Other vocabularies can combine elements from different standards. For example, for the Biodiversity Evolution Studies and Biological Oceanography we selected descriptors from multiple standards which we combined with new descriptors created based on the specificity of each experiment (Castro et al., 2017). Finally, some vocabularies can be based on a single standard if it is comprehensive enough to cover the identified data description requirements. A good example of this is our work with the DDI standard, detailed in the next section.

The data and metadata are Accessible by their identifier, using a standardized protocol that is open, free, and universally implementable. Interoperability relies on the use of a formal, accessible, and broadly applicable language for knowledge representation. Dendro uses Linked Open Data at the core, which encourages data curators to model ontologies that satisfy the needs of each specific domain while maintaining the interoperability characteristics of the ontology itself.

The metadata also meets domain-specific community requirements to make data Reusable. Besides the full representation of the Dublin Core terms[11], we reuse concepts from disciplinary standards, as mentioned above. Moreover, Dendro integrates with several data repositories, e.g., CKAN instances and EUDAT's B2SHARE (Silva et al., 2018). A package containing the dataset and its metadata can be submitted to the intended repository in the final step of the process. The adoption and combination

of multiple descriptors address the data documentation limitations associated with generalist repositories meant to cover several research communities, as identified by Assante et el. (2016).

Development of the DDI subset ontology

In the course of our RDM activities, we established several contacts with experts from a diversity of domains through open data management sessions or word-of-mouth recommendations. Since data reuse depends on the knowledge of variables, data collection methodologies, and experimental parameters, engaging with domain experts is a useful approach to identify relevant concepts for the production of domain-specific metadata, as well as a favorable pretext for researchers to improve their RDM awareness. Once validated by researchers, those concepts can be formalized as *data properties* (using the Protégé editor, a free and open-source ontology software), with specified *rdf:labels* and *rdf:comments,* in a lightweight ontology (Castro, Rocha da Silva and Ribeiro, 2014). These labels define the natural language representation of the descriptor and the comment is used for the definition presented in tool interfaces.

From the moment we started our contacts with social science researchers from the University of Porto, we proceeded to identify a first set of descriptors based on the DDI, namely: *Data Collection Methodology*; *Data Source*; *Sample Size*; *External Aid*; *Kind of Data*; and *Universe*, as depicted in Figure 2 (Amorim et al., 2015). Further feedback from social scientists focused on the importance of describing the methodology and the need for additional finer descriptors in order to enrich the metadata.

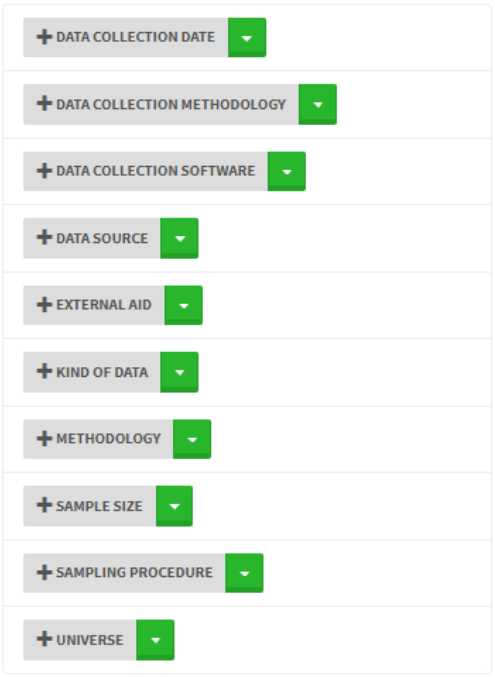

Figure 2: First set of DDI descriptors implemented in the Dendro platform (Amorim, 2015)

As we pursued the contacts in the social sciences, there was the need to extend the number of descriptors in the DDI subset included in Dendro, so we looked at the metadata recommendations for data deposit in the ICPSR repository. We assumed that the metadata fields recommended for the deposit in the ICPSR are intuitive for both domain researchers and data curators with less social sciences expertise. Moreover, combined with Dublin Core metadata they offer the guarantee of interoperability and dissemination of data outside the project. The intention is to have data described from the outset and to be ready for deposit.

Additional descriptors that we added to the DDI subset include *Analysis Unit; Dependent / Independent Dimension; Sampling Procedure; Summary Statistics;* and *Time Method.* We also uploaded the DDI-RDF Discovery Vocabulary (Disco)[12] to Dendro.

Approach

In order to promote RDM awareness and simplify the description of data by the researchers from the TOGETHER project, our approach consisted of a set of interactions, using a combination of techniques.

The first contact with the researchers from the TOGETHER project took place before the launch of the project. More specifically, it happened in the general meeting with a group of researchers from the Psychology and Educational Sciences domain, in the context of the dissemination of the objectives of the TAIL project to the scientific community of the University of Porto. The researchers that were preparing the TOGETHER project expressed their interest in developing knowledge to adopt RDM measures and agreed to collaborate in the proposed activities.

Figure 3 is a timeline of the contacts with the researchers after the initial general meeting. We scheduled interviews to assess practices and domain-specific metadata requirements in November 2017 and December 2017, and researchers were also involved in data description sessions at the beginning and the end of 2018. Between the first interview and the first data description session, we developed the metadata model to make sure that DDI descriptors were available in the Dendro platform.

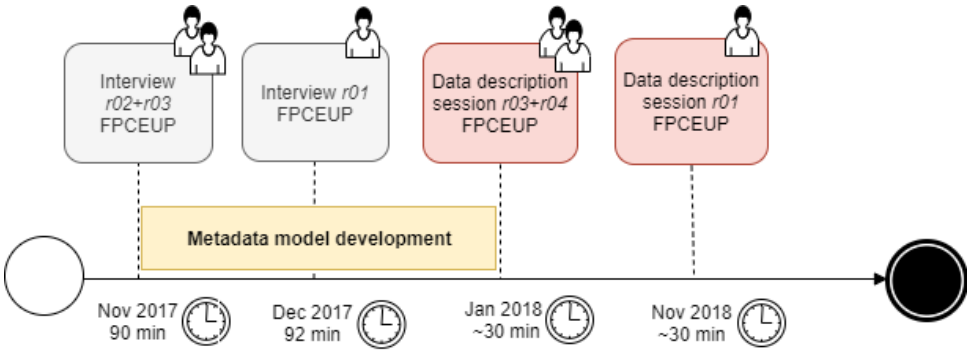

Figure 3: Timeline of the activities and participants in the study.

The members of the TOGETHER project who participated in this work were its Principal Investigator (r01) and the Co-Principal Investigator (r02), both senior researchers with a PhD. In both cases, the frequency of data usage is low given the leadership, monitoring, and management roles assumed in their current projects. Two other postdoctoral researchers, (r03) and (r04), also participated in this work, given their collaboration with r02 in other projects and their frequent activity in the collection and analysis of data.

Diagnostic interview

The method used in the TAIL project to engage researchers in data management started with a diagnostic in the form of a semi-structured interview based on the Data Curation Profile Toolkit, Interview Sheet (Carlson, 2012). This interview sheet is designed to develop the data curation profile of particular projects. The sheet is structured in modules for specific stages in the data life-cycle and takes into account researchers' practices and perspectives to guide the conversation, namely details about the dataset, data sharing, repository usage, and organization and description of data.

The interviews focused on the researchers´experiences in previous projects, as well as their general research interests and background, and were conducted in late 2017. The interview with r01 lasted 92 minutes and the one with r02 took 90 minutes.

The interviews were transcribed and coded, in Portuguese, using the ATLAS.ti software. Six main code categories, detailed below, were defined to markup relevant statements. Other annotations were made freely to highlight important excerpts that did not fit the predefined categories. This task was performed by a single analyst in the context of a doctoral thesis that involved several other interviews.

i. *Demographic information*: Information such as professional title, data usage frequency, and level of metadata expertise;
ii. *Awareness*: Statements that show that the interviewee is aware or unaware of a given data management topic, which can be raised during the interview itself;
iii. *Share:* Statements that indicate interest in data sharing and issues related to data sharing;
iv. *Organization practice*: Statements that describe tasks performed by the researcher to organize data, both for problem-solving activities and perceived issues;
v. *Annotation practice*: Statements that encompass activities to document data, from ad-hoc metadata practices to standard metadata usage;
vi. *Reuse perspective*: General statements concerning data reuse potential and positive or negative experiences with data reuse.

Data description sessions

In our workflow, a data description session is an activity designed to introduce and train researchers in metadata creation. When scheduling the sessions with the researchers, they are asked to select a dataset to describe (if possible, one mentioned during the interview) from an ongoing project or a recent publication. The biggest constraints to scheduling these sessions were related to researchers' busy schedules and lack of interest in participating.

In the data description sessions, we started by introducing researchers to Dendro. We briefly demonstrated features, like creating a new project and filling metadata, adding contributors, making backups and keeping versions. Each researcher was then asked to create a folder and upload a dataset. After this step, we explained in detail the choices that could be made in the descriptors panel, recommending DDI and Dublin Core descriptors according to the characteristics of the datasets generated by the researchers.

During the sessions the selection of metadata elements was mostly up to the researcher - we stepped in when requested or when we realized that the researcher was not progressing in the task. In case there was an opportunity, we advised them on the meaning of descriptors or the use of metadata.

Audio from the sessions was recorded with consent and deleted after the transcription of the relevant events and comments, which were then used to complement the analysis of the produced metadata. The audio was also used to mark the moment when researchers started and finished the description, i.e., to ascertain the session duration. With the experience accumulated at the University of Porto, we found that these sessions tended to last approximately 30 minutes. In general, the shorter sessions took place in experimental domains, where many metadata fields could be represented with a numerical value. In one case, a researcher from an experimental domain completed 28 fields in 16 minutes. At the other extreme, a researcher from a sociology area completed 21 fields in 75 minutes. Here, the metadata had a substantial volume of text and the researcher consulted documents and justified most choices.

A few weeks after the sessions, we asked the researchers to fill out a brief online questionnaire to get additional feedback, namely concerning the perceived usefulness of data description for their research purposes, the degree of interest regarding RDM activities, and the most important factors for RDM engagement. We also showed a document to the researchers with the metadata record created in Dendro and asked them to judge whether the information was sufficient or if more information might be provided.

Results

Data organization and metadata practices

In the interview, r01 addressed a study related to methods of assessing general psychological well-being in mental health contexts, with the goal to understanding how individualized measurements can add value when compared with standardized ones. This study generated self-reported data collected over time in a clinical context.

Metadata was a concept vaguely understood in the form of keywords by r01. To keep track of data, r01 resorted to meaningful file names and the date of recently saved files, although this strategy was recognized as not always effective. For versioning information, a recurring solution was to write a complementary text file that served as an alert when the database was shared or when the information was directly edited.

A specific data sharing challenge identified by r01 related to security and confidentiality, particularly when matching codes to combine datasets from questionnaires that were passed on to people with databases from confidential clinical processes. Data sharing with external parties is a very delicate and rare issue, requiring tight regulation, particularly if the data is to be reused in a different country. Another aspect mentioned by r01 was that the ability to remember important details decreases over time. Although data may not be physically lost, the confidence to reuse it may decrease, reinforcing the need to make the information more explicit.

In the r02 case, the background for the interview was a study to understand how the dynamics between work and family are linked to the exercise of parenting and the children´s socioemotional development. This researcher had no previous experience with metadata – "*that is what we want you to teach us*".
An issue for r02 was how to retrieve data in periods of greater workload, something that is overcome by personal strategies and through years of experience.

When addressing the ability to remember specific details about the data, r02 compared the direct collaborators to hard discs. One of the things required from collaborators are memos, "*which might be metadata*", of things related to the data. In relation to metadata, r02 expressed curiosity towards other methodologies and processes to document data, despite an overall feeling that up to that point the projects the projects had been successful in that respect. Nevertheless, the need to share data was present and r02 mentioned the importance of having data accompanied by an "*instruction book*".

Sharing data with outsiders was something r02 was anticipating on a network with partners from Europe, the USA, Japan, and China, and the expected issues concerned the different cultural contexts in which the databases were created. A specific issue was how to integrate cultural knowledge in the interpretation of data. When asked about the reuse potential of data outside of the original project and domain scope, r02 considered that the question posed an interesting challenge to think about, and was not sure about how data might be reused beyond the psychology and social sciences fields.

Both researchers agreed that data documentation is pivotal for data reuse. For r02, data documentation is directly related to the quality of the data itself. A scenario described by r02 was the need to document occurrences that are out of the ordinary— a teenager responding in jest, for instance. This type of behavior from participant behavior is easily spotted by the researcher in charge of data collection and this information is useful to process the data accordingly. Similarly, r01 mentioned frequently asking collaborators to annotate data if they felt that the person did not understand the questionnaire. Moreover, r01 considers data documentation essential to reuse data in a different context, for instance when clinical data collected by a therapist is aggregated at the service level to evaluate service quality.

Data description by researchers using Dendro

The two data description sessions lasted approximately thirty minutes each, excluding the time dedicated to introducing Dendro or addressing any questions. Session 1 was carried out in January 2018 with two participants, r03 and r04. A dataset with results from descriptive statistics on children's emotion regulation, parents' work-family conflict, and psychological availability was described during this session and published in B2SHARE[13]. Although r02 was the Principal Investigator of the project that generated the data they opted to delegate the description to the two collaborators that worked with the data regularly, r03 and r04.

In this session, r03 and r04 filled in 13 metadata elements in 30 minutes. Both researchers had no data description experience but became familiar with the proposed task quite easily. They talked to each other during the session to discuss the meaning of some descriptors. The researchers were very prudent in the selection of descriptors and in the information provided given the perspective of subsequent data publication. They selected metadata elements for the temporal context and methodological information such as *Sampling Procedure, Time Method,* and *Sample Size*. In this case, they considered the DDI subset convenient for their data, especially because the concepts are close to the terminology they regularly adopt.

Session 2 took place in with r01 in November 2018. The metadata recorded in this session pertained to the validation of three tools for assessing the psychosocial impact of genetic testing for cancer risk. In this session, r01 produced a metadata record with 17 descriptors. The metadata values consisted mostly of short texts. This researcher used the *Description* element to contextualize the project and used it again to identify the type of data that made up the uploaded file. Despite the incomplete information on the characteristics of the dataset, r01 showed awareness of the need to use the same metadata element at different levels of description.

The *Deviation From from Sample Design* was also filled, showing consistency with what had been said during the interview regarding the need to document the contingencies in the research process. The quality of the metadata may have been hindered by two constraints. First, the data description session

took place amidst a busy schedule; second, as the follow-up questionnaire would reveal, the benefits and the objectives of metadata creation were still not totally clear.

Table 2 shows the list of descriptors from Dublin Core and DDI filled in during each session. Additionally, it shows the number of times that these descriptors were used by other social scientists in all contacts we established during the TAIL project, beyond the TOGETHER project. From the beginning of 2018, we carried out 8 data description sessions with social scientists in Dendro, including the two detailed here. In addition to the sessions reported in this work, the other social sciences domains represented in the right column cover studies related to consumption sociology, questionnaires to evaluate fitness trackers, the nutritional status of people with dementia, work psychology, organizational sociology, and fragility assessment. We recognize the interdisciplinary nature of most research projects, but we take into account the typology of the data described in each case and the fact that the DDI is the most suitable vocabulary to represent them.

Metadata elements related to methodological aspects such as *Sampling Procedure*, *Kind of Data*, and *Sample Size,* are generally the first to be filled in because they are common in the research process and researchers can easily understand these concepts. We also found that, due to their scope, these descriptors are chosen by researchers from domains other than the social sciences who are looking for a less general description of their data. A researcher working with nanoparticle synthesis decided to make a high-level description and after browsing Dendro, regarded the concepts represented in the DDI subset as the most appropriate for an immediate registration of metadata.

On the other hand, administrative and descriptive metadata elements are not consistently used, which is in line with the results obtained by White (2014), as researchers tend to focus more on scientific details. In Dendro, we observed a tendency of researchers to choose elements from their domain ontology and to show little interest in exploring complementary vocabularies that would enrich the quality of the metadata. However, this implies that researchers still need to deepen their general understanding of the benefits of metadata.

Table 2: Descriptors filled in during the data description sessions

Descriptor use	Session 1 (*r03* + *r04*)	Session 2 (*r01*)	Total in 8 sessions in social sciences
Abstract	✔	✔	6
Sampling Procedure	✔	✔	6
Kind of Data	✔		6
Temporal Coverage	✔		5
Sample Size	✔	✔	5
Data Collection Methodology		✔	5

Creator	✔	✔	4
Spatial Coverage		✔	4
Language	✔	✔	4
Methodology		✔	4
Universe		✔	4
Date Created		✔	4
Subject	✔		4
Audience		✔	3
Format	✔		3
Analysis Unit		✔	2
Description		✔	2
Relation	✔		2
Summary Statistics	✔		2
Access Rights		✔	2
Variable	✔		2
Time Method	✔		1
Instrument		✔	1
Deviation From Sample Design		✔	1
Coverage		✔	1

Overall, the researchers in the two sessions created good quality metadata records, considering that they were able to produce comprehensive and detailed metadata records in their first experience in this kind of activity. The metadata provided by r03 and r04 can easily be enriched with temporal coverage information and the accuracy of the format information can be refined.
The record created by r01 lacks subject and temporal metadata. In both cases, it would not take much to improve the metadata records in order to promote search and access to the data. The metadata is rich in terms of information regarding the methodologies and context of data production.

Only the *Abstract*, *Sampling Procedure*, *Creator,* and *Language* co-occurred in the two sessions. This suggests that it is necessary to maintain a subset with an adequate number of descriptors, i.e., large enough to fit the expectations of researchers, and also to have flexible tools that enable researchers to combine suitable descriptors according to the metadata requirements of a given dataset. It also suggests that researchers continue to be involved in training related to the production of metadata to increase their awareness of the benefits that can result from detailed and accurate metadata.

Researchers' perspectives

The additional feedback obtained through the online questionnaire showed that the researchers think that the metadata they produced during the session was sufficient. According to r01, there was no need for additional information. Both r03 and r04 stated that more information might have been useful but did not provide details.

After participating in the session, and despite having created a detailed metadata record, r01 did not identify any particular usefulness in the data description – "*I have not yet fully understood the application of the knowledge that resulted from the description of the data. I still considered it as important but it is something abstract*". This researcher thinks that data description is an important subject, but an overly abstract activity. Data description was perceived as a very boring activity, slightly difficult and time-consuming. On the other hand, the interest in RDM is very high since it helps to organize, store, and reuse data.

As for r03, data description is a somewhat easy and practical task, yet slightly time-consuming. The activity was considered useful to facilitate the dissemination of data to other researchers and to the academic community. Moreover, according to this researcher, the reuse of existing databases is also a benefit since it prevents overloading participants with new questionnaires. The degree of interest of r03 regarding data management is moderate, which can be explained by the general feeling that the research group was being successful in data documentation, mentioned by r02 during the interview.

The participants agree that they will have more interest if RDM practices bring more visibility to their work via data citation and enable mid- or long-term data reuse. On top of that, r01 highlights the availability of RDM tools as a top factor for increased motivation. The availability of data description tools was a preference for r03, corroborated by r04. Additionally, r01 stated that the developed activities can be improved by practical examples of data description in the Clinical Psychology domain.

After carefully thinking about the collaboration described in this case study, r01 highlighted that if data management activities are not properly integrated into the research workflow, they may be perceived as time-consuming and as an overload, which may prevent researchers from engaging in them. Moreover, there is also the need to work on the communication between data curators and researchers. In the words of r01: "*On one hand we now have an established communication channel and a work relationship, we are aware of our data management needs, and we wish to be on board in the development of tools that integrate data management in routine research work. However, on the other hand, we found ourselves lost in the transition between the data manager´s and the researchers' world. For us as researchers, data management terminology is still perceived as abstract and technical. On the other hand, data management activities are perceived as essential but logistic, and the aim is to spend as little time as possible planning and implementing them. Essentially, we are aware that there is much to do in order to address RDM routinely in research projects*".

Conclusion

The work carried out during the TAIL project focused on the integration of data management tools early in the research workflow. To do so, we established a set of contacts with researchers at the University of Porto to assess domain requirements and test such tools.

In promising cases, such as the one established with the TOGETHER project, the collaboration can take place throughout the project as an independent and complementary task. Among other aspects, it can make it possible to develop in-depth knowledge that can be applied in recommending best practices for ongoing and future projects with similar characteristics.

We envision the collaboration with researchers as an opportunity to establish data management practices tailored for small projects and supported by recommendations from expert communities. Documenting successful stories due to the adoption of good RDM practices can have a persuasive effect on others since researchers often ask for practical examples from their domain to understand what is expected from them to comply with the current RDM mandates.

This paper described activities to simplify metadata production by researchers, in a context where adequate resources to get them involved in RDM are missing. We implemented a DDI subset in Dendro, a staging platform for data description, that was used by researchers in their first contact with metadata creation.

Good examples can provide valuable insight for the adoption of the DDI subset and lay the foundations for data documentation in projects in the Clinical Psychology domain.

The results showed that participants were able to choose and fill in several descriptors in a reasonable amount of time (half an hour), thus producing a comprehensive metadata record, especially when compared to the metadata that is usually available in generic data repositories. Nevertheless, to meet the data deposit metadata a requirement of a disciplinary repository like ICPSR further investment is required.

The metadata records created were also similar to those we obtained with other social scientists using the DDI representation in the Dendro platform. On average, our data description sessions took 35 minutes and resulted in 13 metadata elements. Still, data description was considered a time-consuming activity by the participants and boring by r01. If on the one hand, the communication between r03 and r04 helped them to understand the meaning of some concepts, r01 acknowledged doubts regarding the meaning behind *Analysis Unit* and *Time Method*. These observations indicate that data description may be perceived more as an additional task than as an activity that saves time and offers other benefits afterward.

It should be noted that this feedback is tightly related to the specifics of Dendro´s design, and users´ experience of metadata production may vary according to the according to the perceived usefulness and usability of the data description platforms.

Our approach to model a domain standard and train researcher in metadata creation was also exploited in other contexts, namely with researchers from the biomedical domains, from a large institute for health sciences and technologies in Porto (Sampaio et al. 2019). The feedback from these biomedical researchers suggests that the use of a restricted vocabulary favors data description but did not prevent them from identifying the limitations of the model and did not prevent them from arguing about the usefulness of descriptors even more specific to the type of experiments they perform.

Further developments would benefit from opening these activities to more researchers from the same domains, particularly to assess the quality of the metadata produced. However, reaching new participants is often a laborious task on its own. In addition to their busy schedules, a general belief that current practices are already good enough may prevent some researchers from participating in this type of study.

Another possibility is to study the attitude of researchers toward data description and the overall quality of metadata over different platforms. To improve the results, it would be useful to have a default list of general descriptors presented to the researchers as they start the description of data, as in most cases they do not know where to begin. It would also avoid browsing the Dendro vocabulary list. Researchers from experimental domains, for instance, have mentioned interest in having a limited number of high-level elements that can be broken-down into more specific elements according to their initial choices. Implementing controlled vocabularies is something that can also improve the experience of researchers.

Researchers pointed out that it would be helpful to see practical examples of the use of data description and to involve others in collaborative training activities, within research units or through an institutional training plan. The engagement of more and more researchers is likely to encourage others to participate.

References

Assante, M., Candela, L., Castelli, D. and Tani, A. (2016). 'Are Scientific Data Repositories Coping with Research Data Publishing?', Data Science Journal, 15:6, pp.1–24. doi: http://dx.doi.org/10.5334/dsj-2016-006.

Castro, J.A.; Amorim, R., Gattelli, R., Karimova, R., Silva J. R. and Ribeiro, C. (2017) 'Involving data creators in an ontology-based design process for metadata models', Developing Metadata Application Profiles, pp. 181-213. doi: 10. 4018/978-1-5225-2221-8.ch008

Carlson, J. (2012) 'Demystifying the data interview. Developing a foundation for research librarians to talk with researchers about their data', Reference Services Review. 40(1). doi: 10.11080090732121120360 3

Clare, C., Cruz, M. Papadopoulou, E., Savage, J., Teperek, M., Wang, Y., Witkowska, I. and Yeomans, J. (2019) 'Engaging Researchers with Data Management: The Cookbook', Open Reports Series. 8. doi:10.11647/OBP.0185

Curty, R. G. (2019) 'Factors Influencing Research Data Reuse in the Social Sciences: An Exploratory Study', International Journal of Digital Curation. 11(1). doi:10.2218/ijdc.v11i1.401

European Commission (2016a) 'Guidelines on FAIR Data Management in Horizon 2020', Technical Report. Available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

European Commission (2016b) 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)', Official Journal of the European Union. Available at: http://publications.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en

European Commission (2018). Turning FAIR into reality. Final Report and Action Plan from the European Commission Expert Group on FAIR Data. doi:10.2777/1524

Mayernik, M.S. (2011) 'Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators', ProQuest Dissertations and Theses. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2042653

Perrier, L., Blondal, E., Ayala, P., Dearborn, D., Kenny, T., Lightfoot, D., Reka, R., Thuna, M., Trimble, L. and MacDonald, H. (2017) 'Research data management in academic institutions: A scoping review', Plos One, 12(5). doi:https://doi.org/10.1371/journal.pone.0178261

Qin, J., Ball, A. and Greenberg, J., (2012) '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.

Qin, J. and LI, K., (2013) '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.

Ribeiro, C. and Fernandes, M. (2011) 'Data Curation at U. Porto: Identifying current practices across disciplinary domains, IASSIST Quarterly, 35(4), doi:10.29173/iq893

Ribeiro, C., Silva, J. R., Castro, J. A., Amorim, R., Lopes, J.C. and David, G. (2018) 'Research Data Management Tools and Workflows: Experimental Work at the University of Porto', IASSIST Quarterly, 42(2), pp.1–16. doi:https://doi.org/10.29173/iq925

Rocha da Silva, J. (2016) 'Usage-driven application profile generation using ontologies'. PhD thesis, Faculdade de Engenharia da Universidade do Porto.

Rocha da Silva, J., Ribeiro, C. and Lopes, J.C. (2018) 'Ranking Dublin Core descriptor lists from user interactions: a case study with Dublin Core Terms using the Dendro platform', International Journal on Digital Libraries. doi:https://doi.org/10.1007/s00799-018-0238-x

Sampaio, M., Ferreira, A. L., Castro, J. A., Ribeiro, C. (2019) 'Training Biomedical Researchers is Metadata with a MIBBI-Based Ontology', In: Garoufallou, E., Fallucchi, F., William De Luca, E. (eds) Metadata and Semantic Research. MTSR 2019. Communications in Computer and Information Science, vol 1057. Springer, Cham. https://doi.org/10.1007/978-3-030-36599-8_3

Vines, T.H., Albert, A.Y.K., Andrew, R.L., Débarre, F., Bock, D.G., Franklin, M.T., Gilbert, K.J., Moore, J.S., Renaut, S., and Rennison, D.J. (2014) 'The availability of research data declines rapidly with article age', Current Biology, 24(1), pp.94–97. doi: 10.1016/j.cub.2013.11.014

Yoon, A. (2016) 'Red flags in data: Learning from failed data reuse experiences', Proceedings of the Association for Information Science and Technology, 53(1). doi: 10.1002/pra2.2016.14505301126

White, H.C. (2014) 'Descriptive Metadata for Scientific Data Repositories: A Comparison of Information Scientist and Scientist Organizing Behaviors', Journal of Library Metadata, 14(1), pp.24–51. doi: https://doi.org/10.1080/19386389.2014.891896

Wiley, C., and Kerby, E. (2018) 'Managing Research Data: Graduate Student and Postdoctoral Researcher Perspectives', Issues in Science and Technology Librarianship. doi: 10.5062/F4FN14FJ

Willis, C., Greenberg, J. & White, H. (2012) 'Analysis and Synthesis of Metadata Goals for Scientific Data', Journal of the American Society for Information Science and Technology, 63(8). doi: 10.1002/asi.22683

End-notes

[1] All authors are affiliated with the University of Porto

[2] https://www.inesctec.pt/en/projects/tail, Supported by European Compete grant (POCI-01-0145-FEDER-0167

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

[4] Supported by European Compete grant (POCI-01-0145-FEDER-030980) and Portuguese National Funds FCT – Fundação para a Ciência e a Tecnologia, I.P. (PTDC/PSI-ESP/30980/2017)

[5] https://www.icpsr.umich.edu/web/pages/datamanagement/lifecycle/metadata.html

[6] https://rd-alliance.org/groups/metadata-standards-directory-working-group.html

[7] http://www.ddialliance.org/Specification/

[8] https://dwc.tdwg.org/

[9] https://www.eurocris.org/cerif/main-features-cerif

[10] https://www.w3.org/TR/prov-overview/

[11] http://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[12] http://www.ddialliance.org/Specification/RDF/Discovery

[13] doi: 10.23728/b2share.7b3c66dfa4df4a7f9ba04fbc30cfb8bc