JOÃO DANIEL AGUIAR DE CASTRO

ENGAGING RESEARCHERS IN RESEARCH DATA MANAGEMENT: CREATING METADATA MODELS FOR MULTI-DOMAIN DATASET DESCRIPTION

SUPERVISOR: MARIA CRISTINA DE CARVALHO ALVES RIBEIRO (PH.D.)

PH.D. THESIS

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Faculdade de Engenharia da Universidade do Porto /
INESC TEC
http://www.fe.up.pt
2020, January

João Daniel Aguiar de Castro: *Engaging researchers in research data management: Creating metadata models for multi-domain dataset description*, Ph.D. Thesis, 2020, January.

E-MAIL:
joaoaguiarcastro@gmail.com

ABSTRACT

The rapid rate at which research data are produced is not matched by generalised data sharing and reuse. Several initiatives and policies have taken shape in recent years to encourage the widespread adoption of Research Data Management best practices. In this context, metadata emerge as an essential component for sustained dissemination of research data. Ideally, the production of metadata should take place as early as possible in the data lifecycle, making researchers the main stakeholders in this activity. However, most researchers do not have established data description practices, and the creation of metadata is generally considered a burdensome task. Moreover, the complexity associated with scientific standards frequently make these tools impractical for researchers to adopt.

In this work the development of domain-specific metadata models is explored as a means to engage researchers in the description of their data, on the assumption that a collaborative approach between data curator and researchers helps to identify familiar concepts for the researchers, facilitating along the way metadata creation. In order to foster this collaboration, a data curator's workflow is proposed. This workflow includes meetings and interviews with the researchers, to understand domain metadata requirements, the development of lightweight ontologies, followed by data description in Dendro, a staging platform for data description, developed at the University of Porto. This approach was first instantiated in several research domains. To overcome possible communication shortcomings with the researchers, content analysis of domain publications was later introduced, and evaluated, as a complementary task in this workflow.

In order to assess whether the data curator's workflow allows for the creation of quality metadata, 13 data description sessions were carried out with researchers from different domains. After the data description sessions, participants completed a questionnaire to measure their attitude towards data description. The results show that researchers have produced satisfactory or good quality in most sessions. The data description activity was characterized as slightly demotivating and slightly time-consuming, yet somewhat interesting, moderately easy and moderately practical. Researchers also considered data description a useful activity. This work allows the conclusion that metadata creation is a realistic activity to be performed by the researchers as long as adequate tools are provided to them. The proposed data curator's workflow is regarded as a promising approach to engage researchers in research data management, through data description.

RESUMO

O ritmo acelerado a que os dados de investigação são produzidos não é acompanhado pela generalização da partilha e reutilização dos mesmos. Assim sendo, várias iniciativas e políticas surgiram recentemente para encorajar a adoção das melhores práticas de Gestão de Dados de Investigação. Neste contexto, os metadados surgem como uma componente essencial para a disseminação sustentada dos dados. Idealmente, a produção de metadados deve ter lugar o mais cedo possível no ciclo de vida dos dados, o que faz dos investigadores os principais stakeholders nesta atividade. No entanto, a maioria dos investigadores não têm práticas de descrição de dados estabelecidas e a criação de metadados é geralmente considerada uma tarefa morosa. Além disso, a complexidade associada aos standards de metadados torna estas ferramentas pouco viáveis à adoção por parte dos investigadores.

Neste trabalho é explorado o desenvolvimento de modelos de metadados específicos a domínios para envolver os investigadores na descrição de dados, no pressuposto de que uma abordagem colaborativa entre o curador de dados e os investigadores ajuda a identificar conceitos familiares para os investigadores, o que provavelmente facilita a criação de metadados. A fim de promover essa colaboração, é proposto um workflow para o curador de dados. Este workflow inclui reuniões e entrevistas com os investigadores, para compreender os requisitos de metadados de domínio, o desenvolvimento de ontologias leves, bem como a descrição de dados no Dendro, uma plataforma desenvolvida na Universidade do Porto. Esta abordagem foi instanciada pela primeira vez em vários domínios de investigação. Para superar possíveis limitações na comunicação com os investigadores, a análise de conteúdo em publicações científicas, foi posteriormente recomendada, e avaliada, como uma abordagem complementar neste workflow.

De forma a avaliar se o workflow do curador de dados permite a criação de metadados de qualidade, foram realizadas 13 sessões de descrição de dados com investigadores de diferentes domínios. Após as sessões de descrição de dados, os participantes preencheram um questionário para medir a sua atitude em relação à descrição dos dados. Os resultados mostram que os investigadores produziram resultados satisfatórios ou de boa qualidade na maioria das sessões. A atividade de descrição de dados foi caracterizada como ligeiramente desmotivante e levemente demorada, porém algo interessante, moderadamente fácil e moderadamente prática. Os investigadores também consideraram a descrição dos dados uma atividade útil. Este trabalho permite concluir que a criação de metadados é uma atividade realista para os investigadores, desde que lhes sejam fornecidas ferramentas adequadas para o efeito. O workflow do curador de dados proposto é considerado uma abordagem promissora para envolver os investigadores na Gestão de Dados de Investigação, através da descrição dos dados.

ACKNOWLEDGEMENTS

I am very grateful to many people who shared their experience, time, help and encouragement while I was working on this thesis. In particular to:

Professor Cristina Ribeiro, my supervisor, for the clarity of ideas and challenges proposed, and above all, for the trust placed in my work;

João Rocha da Silva, who accompanied me from the moment I arrived at Infolab, for his commitment to solving various challenges, which made it possible to carry out this work;

José Luís Devezas for all his help and motivation, especially throughout the final sprint;

All the colleagues with whom I had the opportunity to get along at Infolab, for all the good moments, namely: Ricardo Amorim, Joana Rodrigues, Yulia Karimova, Nélson Pereira and many others;

Cristiana Landeira, Marcelo Sampaio and Ana Luís Ferreira, for the motivation with which they brilliantly developed the activities I proposed to them;

The researchers who kindly dedicated their precious time to collaborate with me;

Luís Miguel Costa for every word of encouragement whenever we meet;

The Scientific Committee of the PhD Programme in Digital Media for granted me the scholarship that allowed me to advance with this work;

My family for the life they provided me;

My daughter, Matilde, born during this trip, for all the nights she allowed me to sleep;

Alexandra, my wife, the person who makes my life easy every day and who brought me here.

SUPPORT FUNDING ACKNOWLEDGMENTS

The funding for this work given by the PhD grant PD/BD/114143/2015, provided by the FCT - Fundação para a Ciência e a Tecnologia.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Today, research environments are highly shaped by the fast-pace at which data are generated. Driven by new methods and advanced computing capacity, the importance and complexity of data in the research activity is growing [54]. However, data availability tends to decline after publications as a consequence of data mismanagement issues [120], or by a generalized lack of awareness regarding data sharing and publication opportunities.

Although not exclusive, these issues are ubiquitous to small projects in the long-tail of science, which usually struggle with lack of funding, infrastructures and personnel [53]. To prevent data from being lost, or not fully available, research funders, including the European Commission (EU), are asking grant applicants to elaborate data management strategies as a funding condition [42]. In an effort to make sure research data is soundly managed, thereby findable, accessible, interoperable and reusable (FAIR), the EU, is running in its own terms "*a flexible pilot under Horizon 2020, called the Open Research Data Pilot (ORD pilot), to improve and maximize access to and reuse of research data generated by Horizon 2020 projets*" [41]. In order to attain this goal the EU is promoting the adoption of the Data Management Plan (DMP), which is required for all the projects that participate in the ORD pilot.

To comply with the DMP, research projects should contain information about how the research data will be handled during their lifetime and beyond. During the development of a DMP researchers should answer, among others, the following questions:

Which data will be collected during the research project and in which conditions will the data be shared? In Chapter 2.4 I overview the RDM landscape with attention to the types of research data reviewed by Willis et al. [128] and to the data lifecycle. In this chapter I also seek to identify relevant initiatives having as their mission to support researchers in implementing best RDM practices, such as the Research Data Alliance (RDA)[1], a community-driven initiative that enables its members to develop and adopt measures to promote data sharing and data-driven research[2]. Moreover, I look into researchers' data sharing and reuse perspectives.

How will the data be made available (e.g. by deposit in a repository)? Recent years have seen great technological developments in the form of institutional or online data repository services. For instance, DSpace instances[3], which traditionally are used for publications, are being customized to support dataset deposit in institutions with short resources. Online services

[1] https://www.rd-alliance.org/
[2] https://www.rd-alliance.org/about-rda
[3] https://www.dspace.com/en/pub/home.cfm

like Figshare[4] and Zenodo[5] are also suitable data management solutions for these institutions. While data repositories are not the focus of this dissertation, in Chapter 3 I summarize existing RDM platforms in which researchers can disseminate their work, thus enabling to answer this question. An interesting perspective on how generalist data repositories are coping with data publishing was surveyed by Assante et al. [8]. I also had the opportunity to collaborate in a comparative study of data platforms, with respect to their architecture, flexible metadata and interoperability [5].

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what kind of metadata will be created and how? In order to answer these questions in Chapter 4.4 I cover the available metadata standards listed in the Digital Curation Centre's Disciplinary Metadata Catalogue[6], which is part of an initiative by the RDA Metadata Standards Directory Working Group [11]. These metadata standards can either be generic or developed by disciplinary communities. Yet, most standards target data description only at the end of the research workflow and their adoption by researchers can be hard [89].

This work explores the hypothesis that the development of domain-specific metadata models is an adequate approach to involve researchers in RDM through data description.

1.1 MOTIVATION

The investment in RDM creates good opportunities for the dissemination of data but there is a need for a broader commitment of researchers. During my interactions with researchers I became aware that talking about mandates is not a factor of motivation for them. To achieve better results RDM must be perceived as a complex ecosystem of stakeholders and technological infrastructures, both raising conceptual issues.

Several initiatives have taken position to better address RDM practical issues by taking advantage of the available infrastructures. However, some of these issues are of a more convoluted nature, as they depend on researcher's motivation to be active participants in RDM workflows. A very concrete challenge is metadata creation. The dissemination of research data heavily relies in metadata [118], and, if it is made dependent on the few data curators, metadata creation will inevitably cause a bottleneck in the RDM workflow [130]. Researchers need to be perceived as key players in metadata creation, especially when institutions cannot allocate a data curator per department or research group. Moreover, the absence of timely metadata from the start of data production can yield lackluster descriptions. Complex issues may also arise when researchers leave their teams after publishing and have not described their datasets.

By being directly involved in research activities, researchers know best the conditions under which the data were created and can yield more detailed information that enable others to find, interpret and reuse the data. However, metadata creation can be a wearing task with no tangible short-term benefits for the researchers [70]. Difficulties to deliver metadata records con-

[4] https://figshare.com/
[5] https://zenodo.org/
[6] http://www.dcc.ac.uk/resources/metadata-standards

taining contextual information for the datasets were associated with the low rates of data that are actually shared and reused [43]. Small research groups may struggle to keep up with the description demands posed by the existing datasets, and it is not expected that researchers will spend much time in data description activities. A possible compromise scenario is to support researchers in the description of their data as they produce them, postponing the data curator intervention until later in the workflow. To find out more about how researchers are being engaged in metadata related studies, I conducted a survey whose results are presented in Chapter 5.

At the University of Porto, the TAIL project[7], which took place from 2016 to 2019, focused in providing researchers with adequate tools to organize, describe and publish their data [95]. The TAIL project envisioned researchers as the core RDM stakeholders who needed straightforward workflows. During the three years of the project, the TAIL team developed a RDM workflow based on the integration of different tools taking into account the requirements of a panel of researchers built over time. Contacts with researchers at the University of Porto started with a scoping study [94] , 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 the researchers to RDM endeavours.

The activities described here were performed in the TAIL context. My participation in this project was inherently related to bridging the gap between researchers and the tools to support metadata creation as early as possible in the research data lifecycle.

1.2 GOALS AND CONTRIBUTIONS

The application of the FAIR principles depends critically on rich metadata, yet domain vocabularies are still mostly underused. Therefore, the Action Plan to turning FAIR principles into reality [36] recommended, in 2018, the development of use cases to further engage communities and the provision of tools to make data description as easy as possible and promote its adoption by researchers. According to the designated Expert Group on FAIR DATA, "*(f)or some disciplines, the current situation may be satisfactory at present, but it is likely also that opportunities for wider use, greater analysis at scale and reuse across domains are being missed. It would be useful to define use cases to demonstrate the benefits and convince such communities to engage more fully with a FAIR ecosystems*".

Thereby, data curators must play an active role in strengthening the RDM practices of researchers within the limited possibilities that they may have to commit to such practices. My objective with this work was to foster the collaboration between data creators and data curators. This work promotes the engagement:

[7] PTDC/EEIESS/1672/2014

The engagement of researchers in a data curator's workflow for the development of domain-specific metadata models. These metadata models capture familiar concepts for the researchers that they can use in more casual descriptions, which likely mitigate existing barriers to metadata creation.

The proposed data curator's workflow, depicted in Figure 1, includes meetings and interviews with researchers so the data curator can understand the domain and the context of data production. Moreover, by talking with researchers it is possible to infer their practices and metadata requirements.

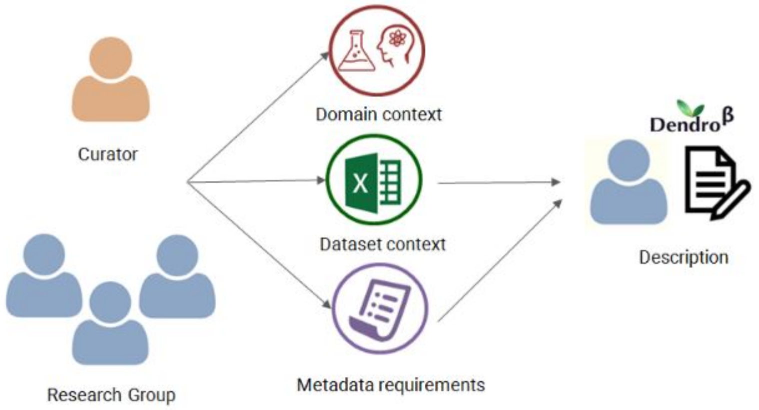


Figure 1: An overview of the data curator's workflow

From there the data curator can suggest potential descriptors to the researchers, which upon validation, are formalized through a lightweight ontology. The final step in this workflow are data description sessions with the researchers. The data description stage is supported by Dendro (see Section 3.3), a platform developed in the TAIL project as well, as part of a PhD dissertation [98].

The proposed data curator's worflow is detailed in Chapter 6. The approach to devise the domain-specific metadata models was first explored in the Fracture Mechanics and Pollutant Analysis domains [24], in the work presented in the DCMI International Conference on Dublin Core and Metadata Applications in 2013[8]. This approach was further instantiated through the modeling of lightweight ontologies and presented in the Joint Conference of Digital Libraries[9] [23] in 2014. The ontologies design process, illustrated with a Vehicle Simulation case study, was communicated in the Metadata and Semantics Research Conference (MTSR)[10] [22] in 2015.

As the number of researchers with whom I collaborated with grew, I realised that my communication with researchers was sometimes hampered both by my lack of domain expertise and their difficulty to master concepts

[8] https://www.dublincore.org/conferences/
[9] https://www.jcdl.org/
[10] http://www.mtsr-conf.org/home

such as metadata. In order to streamline the data curator's workflow I introduce manual content analysis of domain publications, as a way for the data curators to gather domain-specific knowledge, to ease the communication and also to suggest to the researchers meaningful descriptors. This approach is described in Chapter 7, which also includes an evaluation with three researchers from experimental domains. The role of the content analysis approach was disseminated in the International Digital Curation Conference in 2020 [21].

Finally, the data curator's workflow also implies the reuse, whenever possible, of existing metadata standards. The first step is to check the descriptors validated with the researchers in these standards. In Chapter 8 I report an exploratory study where the goal was to train researchers, from the biomedical domain, by introducing them to a previously curated subset of a domain standard. This approach was described in a MTSR publication in 2019 [102].

Several MSc dissertations were developed in close collaboration with researchers, according to the proposed workflow, namely:

- "Gattelli, R, (2015). Gestão de dados de investigação no domínio da Oceanografia Biológica: criação e avaliação de um perfil de aplicação baseado em ontologia [48]"
- "Karimova, Y. (2016). Vocabulários controlados na descrição de dados de investigação no Dendro [59]"
- "Landeira, C. (2018). Gestão de dados de investigação do tipo experimental: casos de uso e contribuições para a melhoria da qualidade dos metadados [61]"
- "Sampaio, M. (2019). Metadados para o uso de ferramentas de gestão com investigadores do I3S [101]"
- "Ferreira, A. (2019). Application of the LabTablet app in a laboratory environement: Case study I3S [45]"

1.3 EVALUATION OF THE DATA CURATOR'S WORKFLOW

To ascertain the merits of the proposed data curator´s workflow I carried out 13 data descriptions sessions, between January 2018 and September 2019, with 13 researchers from a diversity of domains.

The methods and procedures used to set up these sessions are further detailed in Chapter 9. The activities in which each participant was involved include interviews, a data description session in Dendro, and a follow-up questionnaire to obtain their feedback regarding their experience in creating metadata.

In Chapter 10 I explain the sampling techniques used to recruit participants and portray their demographics. Based on the interviews I introduce the participants domain and they are characterized according to their professional title, frequency of data and repository usage, as well as their metadata experience. Moreover, I outline their data sharing and reuse perspectives, along with the data organization and metadata practices. The scientific domains represented in this study are: Family Psychology; Sustainable Chemistry; Clinical Psychology; Magnetic Materials; Services; Consumption Sociology; Organizations Sociology; Nutrition; Magnetic Dynamics; Cultural Studies Work Psychology; Structural Adhesive Joints and Health.

The data description sessions results are showcased in Chapter 11. First, I present individual results per session (Section 11.1), and then the overall sessions results (Section 11.2).

Through the results of the data description sessions I intend to answer two fundamental questions to assess the general applicability of the data curator's workflow to support metadata creation by researchers from different domains.

Research Question 1. *How do researchers assess the data description activity, taking into account the collaborative process between researcher and curator and the domain-specific metadata available?*

In order to verify whether the proposed workflow has culminated in a positive experience for the researchers, it is necessary to probe their general attitude towards the data description activity (Section 11.3). Do they think metadata creation is an interesting and motivating activity? Do they consider it fast and easy? Is data description assessed as a practical activity? Moreover, where did they stand with respect to the perceived degree of data description usefulness?

However, a more positive attitude towards data description, in itself, does not say much about the success of the approach. To this end, the researcher's perceptions has to be compared with the quality of the metadata produced. This brings me to the second question.

Research Question 2. *Do the metadata models available in Dendro enable researchers to create quality metadata?*

I answer this question by assessing the metadata records created by researchers according to the number and type of descriptors researchers have used in describing the data, together with the overall accuracy of the metadata records created. According to the criteria I have established the quality of the metadata is classified as poor, satisfactory or good (Section 11.2.3).

Altogether, the proposed data curator´s workflow can be assessed as a promising approach to engage researchers if the data description sessions yield satisfactory to good quality metadata records, and is generally perceived as a useful activity and a positive experience by the researchers.

The data description sessions are also convenient to gather insight on the choices researchers tend to make in terms of descriptor selection and the amount of time spent in this activity. These are tangible indicators that can inform future decision making in the development of services and tools to support researchers in daily RDM activities.

Part I

An overview of the Research Data Management landscape

2 RESEARCH DATA AND RDM INITIATIVES

Digital technologies, and the way they quickly evolve, raise as many challenges for RDM, as they bring opportunities for the research community, that are still to be fully grasped. The landscape of contemporary science is highly impacted by the availability of research data. This imposes many challenges, as researchers now have to deal with the correct management of these assets.

Major research funding providers are demanding DMPs with recent calls for projects. Examples include the EU under Horizon2020 [41], and the National Science Foundation [78]. Likewise, some publishers have also started to request data as supplementary materials to the submitted articles, under the assumption that their readers should be able to validate or replicate the presented results. Nature, for instance, requests authors to disclose research materials as a condition for the publishing of research papers[1]. Another example is an Open Access publisher, PLOS, that demands a full, unrestricted access to the original data for each of the submitted manuscripts[2]. Following these trends RDM is an increasingly concern for the scientific community. RDM, as a concept, is gaining traction in recent years, and encompasses a wide range of services, tools and practices to help researchers in organizing, documenting and preserving their data, from the moment they start to producing it and beyond the end of research projects.

Several stakeholders are involved in the RDM process [67]. These stakeholders are researchers themselves, the institutions, data curators, data centres, third party users, funders and publishers. Every stakeholder has a role to play, which entails rights and responsibilities. It is the responsibility of the researchers, as well as of the data curators, to ensure that research data can be used by others. Researchers, are key RDM stakeholders, as they are domain experts and should be able to document data to allow others to interpret the data. Data curators, as experts in RDM, can complement their work, and provide them the necessary tools for them to improve their RDM practices.

2.1 RESEARCH DATA

Research data are valuable resources, produced or used in the context of scientific research. It is on the basis of data analysis that researchers support their decisions and draw conclusions that encourage innovation and scientific progress.

[1] https://www.nature.com/srep/author-instructions/submission-guidelines#competing-interests

[2] https://journals.plos.org/plosone/s/data-availability

In its Principles and Guidelines for Access to Research Data from Public Funding, the Organisation for Economic Co-operation and Development [39], defined research data as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. Moreover, a dataset constitutes a systematic, partial representation of the subject being investigated.

The value of data is influenced by their structure and organisation, which is enhanced when available in the digital environment. When this is the case, it can encourage the generation of new scientific knowledge and foster the collaboration between scientific communities.

Research data can be classified as raw data, usually unprocessed, or as processed data, which concerns data that have already went through a process of manipulation. Processed data is more refined and more easily interpreted by third parties, such as tabular data [64]. Research data can also be classified according to the discipline that produces it, and ultimately uses them.

Willis, Greenberg and White [128] argued that the classification of data is helpful for understanding data similarities and differences, as well as its potential use overtime. Therefore, these authors summarized the different types of scientific data classified by the US National Science Board[3], UNESCO[4] and CODATA[5], as either observational, experimental or computational data.

Observational data cannot be recollected and are archived indefinitely, since its measurements are mostly time or space dependent and much of its value is in secondary analysis. Experimental data may be associated with a particular methodology or instruments, its usually produced in disciplines like physics and chemistry, and in principle can be subject to verification by repeating the measurements. Computational, statistical, or data resulting from simulations can also be recreated and verified if information about the original process is captured. Other types of data are those created in governmental, business, public and private life contexts that may also be useful for scientific applications [15].

2.2 DATA CURATION

Data curation is a practice to meet the challenge of the exponential grow of research data, and encompasses the activities undertaken for organising, describing, cleaning, enhancing and preserving data for sharing and reuse.

Data curation is defined by the Digital Curation Centre (DCC), a leading research centre of expertise in digital curation with a focus on building capability and skills for RDM[6], as a process that "involves maintaining, preserving and adding value to digital research data throughout its lifecycle"[7]. For the DCC, active RDM reduces threats to their long-term research value and mitigates the risk of digital obsolescence. On the other hand, data curation is geared towards reducing data duplication, aiming at a higher quality research.

[3] https://www.nsf.gov/nsb/
[4] http://www.unesco.org/
[5] http://www.codata.org/
[6] http://www.dcc.ac.uk/about-us
[7] http://www.dcc.ac.uk/digital-curation/what-digital-curation

Hence, it is important to pay particular attention to preservation and interoperability issues from the beginning of the data lifecycle, in order to counteract a general tendency of the data creators to be concerned with the preservation of data only at the end of the cycle. Therefore, the most relevant RDM decisions should be considered in a timely manner [91]. Framing RDM processes in the data lifecycle allows researchers and data curators to evaluate the actions that should be addressed during the cycle. If planned in time, either before or at the time of the creation of the research data, the completion of the various stages of the data lifecycle are facilitated [10]. Regardless of whether decisions made, even those without a specific goal or strategic purpose for RDM, all research projects have an associated data lifecycle [124].

The DCC curation lifecycle model[8], in Figure 2, provides an overview of all the necessary steps for successful data curation. This model is adaptable to various research domains and can be used to plan activities within organizations to ensure that all stages of the data curation are covered.

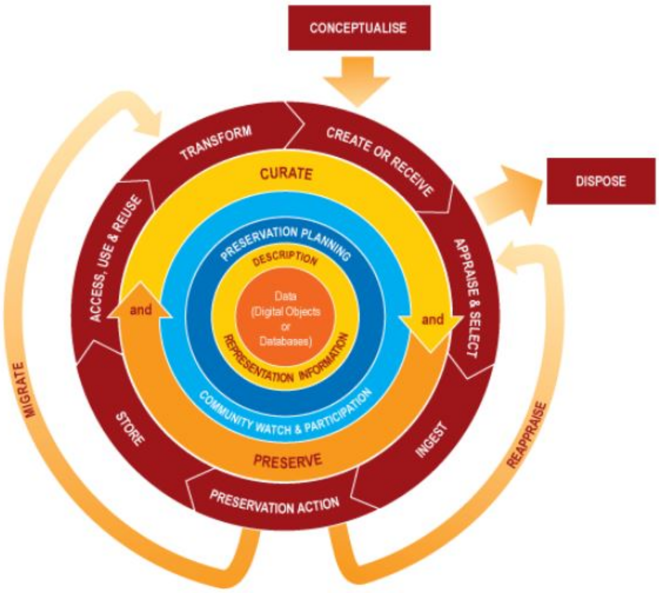


Figure 2: The DCC Curation Lifecycle Model

The DCC curation lifecycle model is divided in full, sequential and occasional actions. In a glimpse, full lifecycle actions correspond to the description of research data, to the awareness to community activities and the participation in the development of shared standards, tools and adequate software. Sequential actions involves planning the creation of data and the production of metadata, the selection of data for long-term preservation, data deposit and storage in a secure manner, as well as transforming data to create new data from the original. Occasional actions includes the disposal of data not selected for long-term preservation, return data which fails validation procedures for further appraisal and re-selection, as well as migrate data to different formats.

Liz Lyon [26], a researcher associated with the DCC, made a comparative

[8] http://www.dcc.ac.uk/resources/curation-lifecycle-model

review of disciplinary data curation differences, based on 16 case studies, and has concluded that it is vital to develop alternative strategies to aggregate the different disciplines, since traditional approaches will not be able to meet the needs of researchers from several domains. This observation reflects, to a large extent, one of the challenges pointed out to the universities seeking to establish data curation infrastructures. According to Martinez-Uribe and Macdonald [69], data curation requires trusted relationships achieved by working and talking with the researchers early on in their research process. To these authors, the failure to engage with the specific need of researchers may lead to the development of data repository services that are under-exploited or indeed may not even be used.

At an institutional level, the people responsible in delivering services to researchers must be aware of the researchers' attitude towards data curation, so that proper RDM can derive from knowing the education and training needs of researchers [105]. Hence, it was also suggested that library staff should have data curation skills [4], while for others, data curation courses should be included in the curriculum of Library and Information Science schools [51]. Toups and Hughes [116], have recommended that small institutions, with discrepancies in funding and man-power, should focused in interdepartmental collaboration to provide robust services for the researchers.

The results from a survey, conducted in 2016, on the Data Asset Framework (DAF)[9], a methodology data provides organizations means to assess their RDM performance, depicted in Figure 3, revealed that 46 percent of researchers at Cambridge were not aware of the services available to them [56]. The survey was self-selecting, which means that the respondents were already engaged to participate in RDM.

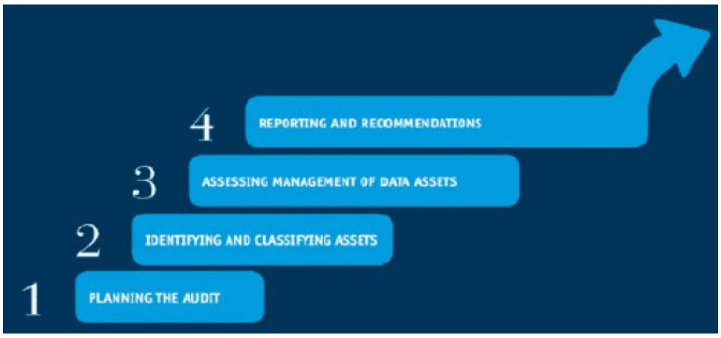

4
REPORTING AND RECOMMENDATIONS
3
ASSESSING MANAGEMENT OF DATA ASSETS
2
IDENTIFYING AND CLASSIFYING ASSETS
1
PLANNING THE AUDIT

Figure 3: Data Asset Framework methodology

2.3 RDM INITIATIVES AND INFRASTRUCTURES

There are many initiatives promoting the management and sharing of research data. For data generated in projects in well-funded research areas there are available mature data curation infrastructures. The NCBI[10], in the life sciences, and the ICPSR[11], in the social sciences, are two good examples of infrastructures used by communities to get base data and to contribute with new research outputs. Other initiatives are also improving the

[9] https://www.data-audit.eu/index.html
[10] https://www.ncbi.nlm.nih.gov/
[11] https://www.icpsr.umich.edu/icpsrweb/

RDM practices of their communities by implementing FAIR principles [36], namely the ESFRI[12] infrastructures in the humanities; DARIAH[13], in the arts and humanities domain; CLARIN[14], which offers digital language resources, particularly targeted for the humanities and social sciences disciplines; and ELIXIR[15], a distributed European research infrastructure for life science information.

JISC[16] is another example of an initiative providing services for researchers, being a non-profit organization that develops resources for education and research in the United Kingdom. JISC has funded a free RDM online course, MANTRA[17], which is organized in 9 units, including a unit dedicated to supporting the development of DMPs.

Stall et al.[109] argued that three changes have to occur across all research fields to change the research culture: make depositing open and FAIR data a priority for all; recognize and incentivize FAIR data practices; fund a global infrastructure to support FAIR data and tools.

2.3.1 FAIR Principles

The FAIR principles were conceived at the Lorenz conference, in 2014, and were published by the FORCE11 Group [36], following the adopted criteria by the European Commission's first set of data guidelines for the Horizon2020 framework programme [41]. Although independent from each other, these principles are interlinked, aiming to make data searchable, accessible, interoperable and reusable. It is a priority to define properties that contemporary data resources, tools, vocabularies and infrastructures should display to enable data reuse by third-parties [127]. Therefore, the FAIR principles suggest the necessary requirements for the data to be in line with the four principles [36]:

Findable: data are findable when described with sufficiently rich metadata and registered or indexed in a searchable resource that is known and accessible to potential users;

Accessible: it means that both humans and machines are provided with the precise conditions by which the data are accessible. Anyone with a computer and an Internet connection should be able to have access, at least, to the metadata. In order for the data to be accessed and used appropriate authorisations and well-defined universal protocols have to be implemented.

Interoperable: data and metadata must use a formal, accessible, shared, and a broadly applicable language for knowledge representation.

Reusable: to make data more usable, understandable or "science-ready" there is a need for rich metadata and documentation that meet relevant community standards and provide provenance information.

These principles do not prescribe, however, a specific line of implementation. Instead, they can be implemented incrementally and through combination.

[12] https://www.esfri.eu/
[13] https://www.dariah.eu/
[14] https://www.clarin.eu/
[15] https://elixir-europe.org/
[16] https://www.jisc.ac.uk/
[17] https://mantra.edina.ac.uk/

This flexible approach makes it easier for more RDM stakeholders to join the FAIR ecosystem[18].

Several policies have worked on these principles and can be seen as precursors to FAIR, such as the OECD's 2007 Principles and Guidelines for Access to Publicly Funded Research Data [39], the Royal Society report in 2012, Science as an Open Enterprise[19], the G8 Meeting of Ministers of Science in 2013[20], and the European Commission Guidelines for Horizon 2020, in 2016 [41].

2.3.2 Research Data Alliance

The Research Data Alliance (RDA)[21] is an initiative launched, in 2013, by the European Commission, the US Government's National Science Foundation and the Australian Government's National Institute of Standards and Technology and Department of Innovation, with the aim of building a social and technical infrastructure to enable open sharing and reuse of data. The participation in RDA is open, and experts from several international communities come together, in Working (WG) and Interest Groups (IG), to exchange ideas and agree on RDM topics like data sharing, education and training challenges, development of DMPs, certification of data repositories, along many other technological aspects.

RDA WKs, last from 12 to 18 months, and are the main vehicle to develop and implement the data infrastructure. On the other hand, IGs, are open-ended in terms of longevity and are focused on solving specific data sharing problems and identifying what kind of infrastructures needs to be built. As an international volunteer member based organization, RDA also encourages the establishment of national nodes, for raising awareness and widespread adoption of recommendations across local communities.

The number of RDA members grew from 8,800 members, in August 2019, to 9,614 in January, 2020, with a total of 88 WK and IGs groups up to this date.

2.3.3 European Open Science Cloud

The European Commission presented, in April 2016, its vision for the European Open Science Cloud (EOSC) as part of the single digital market strategy. The goal of the EOSC[22] is to give the EU global leadership in research data, ensuring that European scientists reap the full benefits of data-based science, through a virtual environment, across borders and disciplines, with free, ready-to-use, open and uninterrupted services for storing, managing, analysing and reusing research data.

The EOSC also envisages the creation of an European data infrastructure supported by high-capacity and supercomputing cloud solutions and its expansion through its gradual opening to the public sector and industry. EOSC is a clear political priority for European research and innovation. This is evidenced by the initial investment of 600 millions euros, between 2016 ad 2020, as well as the likely adoption of the EOSC as a European partnership for FP9 Horizon Europe[23], which will succeed Horizon 2020.

[18] https://www.force11.org/fairprinciples
[19] https://royalsociety.org/topics-policy/projects/science-public-enterprise/Report/
[20] https://www.gov.uk/government/news/g8-science-ministers-statement
[21] https://www.rd-alliance.org/
[22] https://www.eosc-portal.eu/
[23] https://ec.europa.eu/programmes/horizon2020/en/tags/fp9

2.3.4 OpenAIRE

OpenAIRE,[24] is an infrastructure in the FAIR ecosystem, which mission is to "shift scholarly communication towards openness and transparency and facilitate innovative ways to communicate and monitor research." OpenAIRE aims to provide a scientific environment that brings societal benefits.

This infrastructure is characterized by involving RDM stakeholders for an effective implementation of Open Science. Its diverse range of RDM workflow services, includes the provision of training content and the development of common global standards for linking research outcomes to several stakeholders. Thereby, OpenAIRE is connecting open research environments in Europe and promoting data culture changes.

2.4 RESEARCHERS' RDM PERSPECTIVES AND ATTITUDES

Data sharing is one of the validation components of the scientific method, allowing the verification of results and an extension of research on the basis of previous results. The conversation around this topic has mainly discussed the motivation to share data in contrast with the researchers intention to keep control over their data.

In this context, Tenopir et al. [113], ran an international survey in 2011, with a total of 1329 researchers, and concluded that there were effective barriers, rooted in the scientific community culture, which limited data sharing. Most respondents were satisfied with their practices in the initial stages of the data lifecycle, at a time when many organizations did not provide RDM support to their researchers. The main reasons identified for why the data was not made available to others was insufficient time and lack of funding.

Curiously, most respondents (85 percent) were interested in using other researchers' data if the data were easily accessible, yet only half reported making their data available, and 36 percent reported to made their data easily accessible. This illustrated a clear discrepancy between interest and actual data sharing behaviors. Most of the respondents were willing to share data if they could obtain credit through citation, obtain copies of the publications that use their data and knowledge of any outputs developed from their data.

A more recent survey by Tenopir et al. [114], has revealed improvements in the perception researchers have about data sharing and reuse. The authors were able to identify a more positive perception about the value of data sharing and a greater concern about the risks associated with the reuse of third party data. Most importantly, this survey showed progresses in data sharing behaviors, as most researchers reported making at least part of their data available to others.

Several other studies have assessed the main motivations to prevent data sharing. Arzberger et al. [7] also verified that the lack of time and of institutional support for RDM were among the main reasons for researchers to retain data. But commercial interests were also associated to a culture of low data sharing [2]. Savage and Vickers [104] concluded that confidentiality in relation to the subjects, i.e., privacy (clearly evident in the case of medicine), future opportunities for publication and retention of exclusivity about the

[24] https://www.openaire.eu/

data that took a long time to produce, were the main deterrents for data sharing [104]. According to a study by Wicherts et al. [123], there was a more perverse motivation to explain the reluctance of some researchers to share their data; the fear that making data prone to peer scrutiny expose errors or produce conclusions that contradict their own. This authors argued that the detection of errors and data sharing after the publication of results can reflect differences in the rigour with which researchers manage their data, since researchers who apply greater diligence in the archiving and management of their data tend to commit fewer mistakes.

An institutional study about disciplinary differences on RDM attitudes, by Akers and Doty [1], in 2013, found significant disciplinary differences in RDM actions, attitudes and interest in support services, in most surveyed items. One conclusion was that researchers from the basic sciences were most likelier to share data with people outside their groups and deposit their data in data repositories. These researchers were also the ones most familiar with funding agency requirements for DMPs. On the other hand, Kim and Stanton [60], in a 2016 study, with 1317 researchers from 43 disciplines, verified that there was no positive correlation between the funding agencies' policies for data sharing and data sharing attitudes. Yet, journals' policies for data sharing influenced researchers' data sharing practices. However, when Alawi, et al. [4] reviewed 500 publications, from 50 journals, observed that the authors of papers published in high impact journals did not follow the publishers' policies and did not made their data available in open access repositories.

Wiley and Kerby [125], carried out an institutional study in 2018, with graduate students and postdoctoral researchers, to evaluate RDM skills, and concluded that many researchers expressed frustration when former colleagues leave without providing annotations of the completed work. Consistent data description and organization was regarded as a challenge given the different workflows, practices and value concepts of individuals. A practical solution to address this challenge was the provision of short descriptions to enable group members to understand the research workflow. In another study, conducted in 2016 [35], which consisted in 13 interviews with social scientists to assess factors of influence on researchers' perceptions and experiences in attempts to reuse data, it was concluded that data documentation was, among others, an important enabling factor for data reuse.

From interviews, carried out in 2016, with 23 quantitative social science researchers who have failed data reuse experiences [93], it was found that access and interoperability are chief primary conditions for a successful data reuse experience, 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 of reported failed data reuse experiences, making it necessary to establish support systems for those willing to reuse data.

3 RDM PLATFORMS

Following the recent requirements in RDM, several platforms are being developed and integrated into diverse research workflows. Dealing with research data from several domains can be an ambitious task when different requirements are in place [128]. These requirements are often related, at a lower level, with metadata and preservation capabilities; these are, however, very likely to affect their longevity and reuse chances.

On the other hand, at a higher level, user experience and integration capabilities are also crucial to ensure, among other achievements, that RDM platforms are adopted by their end users. An important factor that is currently influencing the way these tools are being developed is the fact that researchers are being increasingly asked to collaborate in the management of their own data. This makes sense, as researchers are the ones who can ideally provide better insight about the data production context and meaning.

The absence of institutional support for data deposit or a suitable data repository for data generated by a particular discipline should not inhibit researchers from making their data available and get credit from their work, or even from seeking opportunities to reuse data produced by others.

3.1 REPOSITORY DIRECTORIES

Data repositories play a key role in the dissemination of research data and results. Due to the increasing production of research data there are several solutions, developed both by open source communities and by organizations, available. Consequently, choosing a platform suited to the needs and requirements of each community can be a difficult task due to the large variety of existing alternatives. In this scenario, repository directories are portals that display available data repositories, as well as metadata about the indexed repositories. These repository directories are convenient tools to help users find solutions that contain data that fits their needs. The Registry of Research Data Repositories[1](re3data) and the Directory of Open Access Repositories (OpenDOAR)[2] are two fine examples of such directories.

3.1.1 re3data

re3data is the result of the combination of two existing initiatives: DataBib and re3data.org. The project partners for re3data are the GFZ German Research Centre for Geosciences[3], the Computer and Media Service at the

[1] https://www.re3data.org/
[2] https://v2.sherpa.ac.uk/opendoar/
[3] https://www.gfz-potsdam.de/en/home/

Humboldt-Universitat zu Berlin[4], the Purdue University Libraries[5] and the KIT Library at the Karlsruhe Institute of Technology[6]. The re3data is steered by the RDA, which decided on the fusion of the two existing portals.

As of January, 2020, re3data offered detailed information on more than 2,000 research data repositories, from a wide variety of domains, such as natural sciences, humanities, engineering and many others. Searching for a suitable data repository can be performed taking advantage of its 27 filters, regarding subject; type of content; country; policies, among other features.

3.1.2 OpenDoar

The OpenDOAR (Open Directory of Open Access Repositories) was put in production in 2005. At the time, there were many lists of repositories and open archives in place. However, there was not a single comprehensive or authority list that could compile all of them.

This directory has a comprehensive listing of repositories from around the world allowing access to their content, simultaneously providing tools to support the activity of repository managers. According to its statistics[7], the number of entries grew from a total of 3701 data repositories listed in September, 2018, to 5390 in August, 2020. To date, most of the repositories listed are multidisciplinary (3276), followed by data repositories targeting the health and medicine disciplines (526).

3.2 MAINSTREAM DATA REPOSITORIES

Depending on the requirements of institutions when considering the implementation of a RDM platform, hiring such services may be costly in the long run, and often poses serious limitations to the implementation of local requirements or platform customization. In this regard, several open source communities are actively developing solutions that can be locally installed and deeply customized to meet such requirements, while keeping a set of features that enable data dissemination through the established protocols. An example of these platforms is CKAN[8], which, although being used mainly by governments to disclose government-related data and contribute to administration transparency, also shows value when applied to the needs of research institutions. There are also online data repository services available to support the publishing of long-tail datasets, from which Zenodo[9], Figshare[10], Dryad[11] and B2Share[12], given their growing popularity, are highlighted.

Since the main goal of having data published is to enable sharing and reuse across a vast community of users, metadata is an essential asset to fulfill this goal. However, when data repositories are implemented with no specific community in mind the metadata is often too generic and does not promote the reuse potential of the datasets [8]. Therefore, it is usefull to

[4] https://www.cms.hu-berlin.de/en
[5] https://www.lib.purdue.edu/
[6] https://www.bibliothek.kit.edu/cms/english/
[7] https://v2.sherpa.ac.uk/view/repository_visualisations/1.html
[8] https://ckan.org/
[9] https://zenodo.org/
[10] https://figshare.com/
[11] https://datadryad.org/stash
[12] https://b2share.eudat.eu/

identify what kind of metadata is captured by these services.

CKAN

CKAN, is an open source, free software, developed by the Open Knowledge Foundation as a data management system. CKAN instances have been adopted by many open government data initiatives worldwide. Although not developed to meet RDM requirements, it has a flexible architecture that allow for the customization of its features to support RDM workflows, even at academic institutions[13]. Each resource is associated with a set of metadata, and its default configuration encompasses responsible party, title, resource description, license and date of creation information. It is important to notice, however, that metadata fields can be customized, so users can define new ones.

Zenodo

Zenodo is an multidisciplinary repository, based on the Invenio framework[14], aimed at the long-tail of science. Zenodo is supported by the European Organization for Nuclear Research (CERN), in cooperation with the EU FP7 project OpenAIRE. This online repository service enables users to set their own communities and accepts data in any format and size. Another useful feature for small project is that Zenodo provides a DOI for every public dataset, so researchers can obtain credit for making their data available. The metadata provided includes title, author, publication date, community, license and subject.

Figshare

Figshare was developed by Mark Hahnel, an advocate for Open Data, and aims to provide an environment where authors gain scientific visibility through citation, by allowing the identification of each researcher. The data deposit process in Figshare is very agile, making data accessible without much delay, while also providing a DOI. It also provides portals and curation workflow services for institutions and publishers. The information associated with research outputs in Figshare includes the title, author, description, date, keywords, license and subject.

Dryad

The Dryad is a non-profit repository resulting from the joint initiative between a group of journals and scientific communities (evolutionary biology and ecology), aimed at developing a sustainable solution for data processing. This repository established a partnership with DataONE[15], in order to guarantee indefinitely access to its contents. Dryad allows the sharing of any kind of data format, the storage of publications, and enables embargo periods as well. Dryad is raising data reuse awareness by encouraging authors to provide additional documentation to efficiently describe their data. Is also grants a DOI to published resources. To describe their resources users can provide, among others, temporal and spatial coverage, the full name of

[13] https://rdm.inesctec.pt/
[14] https://invenio.readthedocs.io/en/latest/
[15] https://www.dataone.org/

the lowest taxon level to which the organism have been identified (scientific name), and provenance information[16].

B2Share

B2Share is a platform available under an open-source license for the dissemination of data from diverse research contexts. It is one of the modules from the EUDAT[17] range of services, an european-wide project that provides a collaborative and interoperable environment for RDM stakeholders. B2Share is a free service that assigns a DOI to published datasets. Some of the metadata used are the author, publication date, description and keywords. Yet, different communities have emerged in B2Share[18] and they were granted with the means to set their own metadata requirements.

Assante et al. [8] have analysed the metadata requested by generalist data repositories at submission time, as well as the metadata exposed to end users, among other features, and concluded that the support these repositories offer resemble the practices in literature publishing. However, metadata requirements for research data are not the same as those for publications, since the content of the datasets, often only numerical, makes their recovery and interpretation difficult. Therefore, these authors noticed that generalist data repositories do not pay sufficient attention to relevant aspects to contextualize research data.

Another aspect to consider is that, data repositories mainly promote description at the deposit stage, when it is already late to capture all the desirable information to make sense of the data. Therefore, it is convenient to adopt flexible solutions that address the needs of researchers from different domains, and encourage data description early in the data lifecycle.

The solution may involve the use of tools that make it possible to organise and describe the data upstream, while also ensuring a flexible means of combining metadata, to address the particularities of each dataset.

3.3 DENDRO, A STAGING PLATFORM FOR DATA DESCRIPTION AND ORGANIZATION

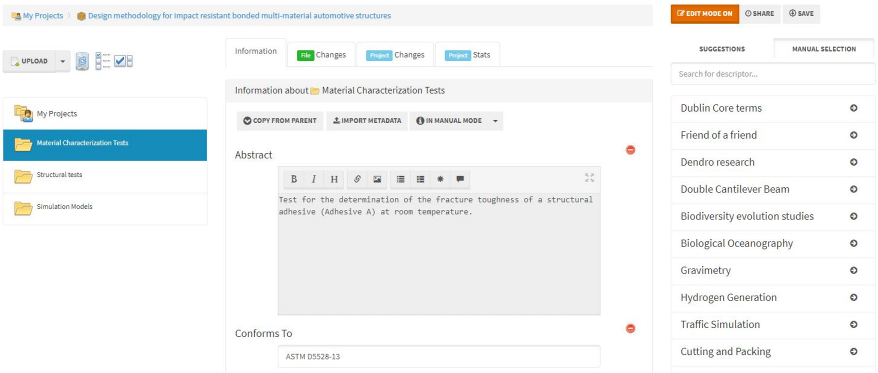
Data repositories are being created to preserve research data, but are in many cases targeted at "finished data", which is only available near the end of the research workflow. Dendro [98], has been developed at the University of Porto, within the TAIL project, as a collaborative RDM platform for small research groups. It is designed to support data description from the moment that data is created and uses Linked Open Data at the core. Its data model encourages data curators to model ontologies that can satisfy the description needs of each specific domain while retaining the interoperability characteristics of the ontology itself.

Data description should start as soon as possible in the research workflow, ideally as a complement task to be performed during the production of the datasets. This is the moment where researchers are still fully involved in the research process and are more aware of the conditions that originate their

[16] http://wiki.datadryad.org/images/8/8b/Dryad3.0.pdf
[17] https://www.eudat.eu/
[18] https://b2share.eudat.eu/communities

data. For instance, recording an event temperature long after its occurrence will surely impact metadata reliability. Dendro, aims to tackle such issues as it offers researchers the opportunity to promptly register the underlying values of metadata for datasets, which is a desirable condition to obtain good quality metadata.

In order to make researchers active stakeholders in the overall description process, data management solutions need to be easily adopted, and Dendro follows a file management structure similar to a popular collaborative environment in the cloud – that most researchers are acquainted to work with – along with collaborative capabilities, thus allowing individuals from the same research team to incrementally create metadata records. This collaborative facet, common in semantic wikis, is handy to ease the data description effort, while also allowing a research team to keep track of the resource´s change log. Furthermore, Dendro is a fully open-source solution for multiple domain dataset description through an extensible, triple-store based data model.



Information about Material Characterization Tests

Abstract

Test for the determination of the fracture toughness of a structural adhesive (Adhesive A) at room temperature.

Conforms To

ASTM D5528-13

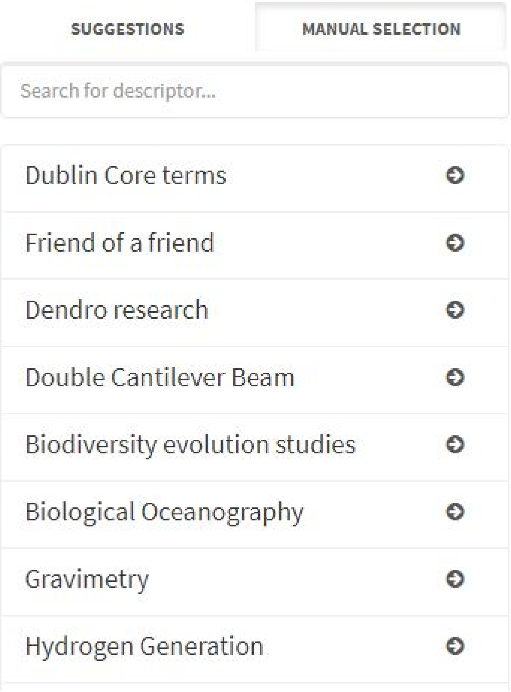
Figure 4: Data description in Dendro

Data curators are also key stakeholders in the Dendro platform workflow. This platform is founded in a flexible data model that data curators can manipulate and expand. The process to evolve the data model is fairly simple, I believe, as data curators with limited programming skills can create ontologies using tools like Protégé[19], and load them into Dendro. These ontologies are then materialized as descriptors in the Dendro user interface, to be combined with other ontologies previously loaded. As shown in Figure 5, descriptors from generic vocabularies are ingested in Dendro, namely the Friend of a Friend vocabulary and an ontology version of the Dublin Core metadata standard (see Section 4.1.2).

Although the creation of ontologies does not require programming skills, it does require conceptual modelling expertise and a thorough analysis of the concepts of the domain being modelled. Furthermore, the ability to add new descriptors into Dendro is fully controlled by the data curator, thus avoiding concept redundancy. This is one of the benefits of working with ontologies, since they enable the incremental approach of adding new descriptors as a data curator sees fit. All the generic and domain-specific ontologies, and correspondent descriptors, are available and searchable to all researchers while using Dendro.

Dendro is expected to act as a platform for data organization and description, based on the principle that researchers are creating documented

[19] https://protege.stanford.edu/



SUGGESTIONS | MANUAL SELECTION

Search for descriptor...

- Dublin Core terms
- Friend of a friend
- Dendro research
- Double Cantilever Beam
- Biodiversity evolution studies
- Biological Oceanography
- Gravimetry
- Hydrogen Generation

Figure 5: Example of available vocabularies in Dendro

versions of their datasets early in the research workflow. By the time data reach a final stage in the research lifecycle, and a final publication is written, core data can be packaged and sent to an external data repository and follow regular deposit procedures. Moreover, if the process is executed fast and is successful, the researcher can still cite the datasets in the publication itself, satisfying the requests of funding agencies and publishers. In short, Dendro aims to make the deposit process as simple as possible for users, while integrating with external platforms for the sake of interoperability. LabTablet, an electronic laboratory notebook solution, also developed within TAIL is one example of the Dendro integration capabilities [6]

Dendro also integrates with several data repositories, allowing researchers to export their described data to an external repository. This approach was tested in a pilot with Eudat. The gathered metadata in Dendro is preprocessed to filter descriptors recognized by B2Share through the existing API, while the complete metadata record is exported as an RDF file [107].

4 METADATA TO SUPPORT RDM

Defined as ”data about data” [77], metadata stands out as an essential resource for data discovery, contextualization, detailed processing, and reuse in the long run. Metadata has become an indispensable component for the management of research data and for scientific communication in general [128]. According to the DCC, metadata is the backbone of data curation and without it, it becomes impossible to recover, interpret and reuse data[1]. Metadata can be seen as a subset of data documentation, in the form of structured information to explain the purpose, origin, spatial and temporal coverage, authorship, access conditions and terms of data use. Likewise, metadata consists in structured searchable information that help users to find existing data resources [32].

Depending on their function, metadata can be categorized into several types [97]. Descriptive metadata describes the content of the resource for search and understanding purposes. Administrative metadata comprehends all the information that enables resource management or that are related to its creation. Within the administrative metadata scope, technical metadata describes the tools needed to manipulate or store files, preservation metadata provides information to support long-term management and preservation strategies, while rights metadata relates to the intellectual properties rights associated to the resources. Moreover, structural metadata provides information with respect to their physical and logical structure, illustrating the relationship between resources.

However, to produce adequate metadata for research data is often a cumbersome task. The documentation of data calls for the participation of researchers, who have to consider the trade-off between performing a time-consuming task with the not so obvious short-term rewards for their effort. Hence, it is normal that researchers might feel discouraged to address data description consistently [89], even though it has been demonstrated that datasets that are linked to detailed metadata records have improved citation rates [86].

Data description should happen, ideally, as soon as researchers start to collect data. The creation of comprehensive data documentation is easier if performed from the outset and continued during the research project [32]. When researchers are actively producing their data they have greater knowledge of the production, and therefore it is the most opportune time to collect metadata. Therefore, postponing data description to the end of the research workflow impairs metadata accuracy, either by loss of detail or, when data producers take on new projects, the description of data from previous projects no longer matters.

[1] http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards

Metadata requirements can be very different across research domains, or even in the same discipline, depending on the research groups culture, funding agency requirements, among others aspects. By taking so many formats, from statistics to interviews, research data demands contextual expertise so that domain-specific requirements are satisfied [33]. Therefore, the classification of data types (see Section 2.1), is helpful to identify high level metadata requirements depending on the type of data: Observational data is often associated with specific locations and time; Experimental data are embedded in laboratorial experiments wherein setup parameters and environmental conditions have influence in data interpretation; Simulational, or computational data requires that all configuration settings are provided, such as the software involved in the capture of data along with the values for the corresponding input variables.

4.1 METADATA STANDARDS

The limited number of research datasets that are actually shared by researchers can be attributed to the difficulties to deliver the metadata records containing information regarding the conditions on which the datasets were created. A possible approach to obtain quality metadata is to balance generic and scientific metadata. By using generic elements, from available metadata standards, in their descriptions, researchers are contributing to a more uniform and interoperable representation of their datasets, albeit only partially addressing data reuse issues. A more exhaustive approach requires the use of scientific metadata for an in-depth and accurate description.

4.1.1 Metadata Directory

The RDA Metadata Standard Directory Working Group[2] set out a metadata standards directory [11]. This directory consists in a community-maintained page of disciplinary metadata, featured as a resource in the DCC website[3]. The directory is organized in five areas: Biology, Earth Science, Physical Science, Social Sciences and Humanities and General Research Data. The latter linking to generic metadata standards that were developed with no scientific communities in mind, but that have been adapted to suit the needs of research data.

4.1.2 Generic metadata standards

Dublin Core

Developed in the mid-1990s, by the Dublin Core Metadata Initiative (DCMI)[4], Dublin Core (DC) is one of the most widely used standards, if not the most, for the description of digital resources. *Dublin* stands for its origin in Dublin, Ohio, and *Core* to the fact that its elements can be broadly applied to describe a wide range of resources. The DC metadata element set was published by ISO 15836:2009[5] for cross-domain resource description, and

[2] https://www.rd-alliance.org/groups/metadata-standards-directory-working-group.html
[3] http://www.dcc.ac.uk/resources/metadata-standards
[4] https://dublincore.org/
[5] https://www.iso.org/standard/52142.html

later revised by ISO 15836-1:2017[6], which does not provide implementation guidelines, rather the elements are used in the context of an application profile, which constrains or specifies their use in accordance with local or community-based requirements and policies. In its basic form, the Dublin Core Metadata Element Set (DCMES)[7] strength lies in its simplicity and ease of implementation. The DCMES vocabulary features 15 elements, namely: Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title and Type.

CERIF - Common European Research Information Format

CERIF[8] is the recommended standard by the EU to its members states for recording information about the research activity. This standard is provided by the International Organisation for Research Information, euroCRIS[9], a not-for-profit association, bringing together experts on research information in general and research information systems. The strenghts of CERIF lies in its broad coverage of all aspects related to research information, such as projects, persons, organisations, funding, publication and datasets.

PROV

Sponsored by the World Wide Web Consortium (W3C), PROV[10] specifies a vocabulary to interchange provenance information. The PROV Family of documents, 12 in total, enables to gather information about entities, activities and people involved in producing data or other resources. The main objective of this vocabulary is to assess the quality, reliability and trustworthiness of data.

4.1.3 Disciplinary metadata standards

Data Documentation Initiative

The Data Documentation Initiative (DDI)[11] is an international standard for the description of data resulting from surveys and other observational methods in the social, behavioral, economic and health sciences. This standard was developed by the DDI Alliance[12], in order to document and manage different stages in the research data lifecycle, from data conceptualization to data discovery and archiving. Among its elements can be found the data collection mode, unit of analysis, kind of data, data source, sampling procedure, instrument and temporal coverage reference. DDI is implemented by the CESSDA catalogue[13], which provides a European-wide interface for social science data, by the UK Data Archive[14], the largest organization responsible for the curation of social science and humanities data in the United

[6] https://www.iso.org/standard/71339.html
[7] https://www.dublincore.org/specifications/dublin-core/dces/
[8] https://www.eurocris.org/cerif/main-features-cerif
[9] https://www.eurocris.org/
[10] https://www.w3.org/2001/sw/wiki/PROV
[11] https://ddialliance.org/explore-documentation
[12] https://ddialliance.org/about/about-the-alliance
[13] https://www.cessda.eu/
[14] https://www.data-archive.ac.uk/

Kingdom, and by the ICSPR data repository[15].

Ecological Metadata Language

The Ecological Metadata Language (EML)[16], was developed by the Ecological Society of America[17] in collaboration with the Knowledge Network for Biocomplexity, to meet the needs of describing research data in the ecology domain. EML also serves to describe data from the biodiversity, ecosystems, meteorology and earth sciences, among others domains. Thus, it is often used to describe experimental and observational data. EML is implemented through a set of XML documents that allow the data to be described at various levels, in a modular fashion, through the use of defined structures, but can also be extended with the introduction of new metadata. This standard modules enable the description of spatial, temporal, taxonomic, research methods and protocols, as well as the structure and content of data. One of the many applications of EML is the Global Biodiversity Information Facility[18], an intergovernmental organization that facilitates access to biodiversity data.

Crystallographic Information Framework

The Crystallographic Information Framework (CIF)[19], is a well-established metadata standard for the treatment, distribution and archiving of resources resulting from crystallographic research. It was developed by the Working Party on Crystallographic Information group of IUCr (International Union of Crystallography)[20] and was adopted in 1990. The CIF is composed of a broad set of descriptors, which fit the description needs of complex resources produced in crystallography. It is essentially applied to the description of experimental data. The CIF divides its descriptors into several sets, like the Core dictionary (coreCIF); the Powder dictionary (pdCIF); the Modulated and composite structures dictionary (msCIF). An example of the use of the CIF data model is the Cambridge Structural Database[21], a worldwide repository of organic, metallic and small molecule crystal structures.

Observations and Measurements

The ISO Observations and Measurements[22] defines XML schemas for observations and for features involved in sampling when making observations. Therefore, it enable the exchange of information describing observation acts and their results. This standard defines a set of core properties for recording observations, for instance, the feature of interest, observed property, the instrument, algorithm or process used, the real-world time associated to the observation and the time when the result was generated. Moreover, it also defines the period of time during which the result may be used.

Other examples of disciplinary metadata standards featured in the Metadata Directory are the *Content Standard for Digital Geospatial Metadata* (FGCCS-

[15] https://www.icpsr.umich.edu/icpsrweb/
[16] https://eml.ecoinformatics.org/
[17] https://www.esa.org/
[18] https://www.gbif.org/
[19] http://www.iucr.org/resources/cif
[20] https://www.iucr.org/
[21] https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/
[22] https://www.opengeospatial.org/standards/om

DGM)[23], which provides a common set of terminology and definitions for the documentation of geospatial data, and the *Darwin Core*[24], which is composed by a set of standards to facilitate data sharing across the biological diversity community.

To another extent it is worth to highlight the role of the European Commission, under the INSPIRE directive 2007/2/EC, in proposing an infrastructure for spatial information data sharing across public sector organizations [13]. The INSPIRE metadata recommendations include elements for the identification of resources, their geographic and temporal references and for the conformity with implementing rules on the interoperability of spatial datasets and services. Nevertheless, the INSPIRE directive considers the possibility for users and systems to combine elements from other metadata standards, if these are prescribed by international standards, to achieve more detailed descriptions.

4.2 REQUIREMENTS FOR SCIENTIFIC METADATA

Metadata standards tailored for scientific applications, are not without limitations, and their complexity and specificity makes their compatibility with the fast-paced growth of research data vulnerable. Furthermore, no single metadata standard is able to encompass the needs of every domain without compromising description accuracy, and many initiatives are seeking to modify or extend existing ones, in conformance with their local needs [89].

Realizing these limitations, some authors argue that application profiles are suitable solutions to deal with the diversity of domains and their metadata requirements [52]. In this context the DCMI proposed the Singapore framework[25], depicted in Figure 6, which comprehends a set of guidelines that application profiles designers must follow to ensure maximum interoperability and reusability of digital resources. Included in these guidelines are the definition of the functional requirements of an application profile, the encoding syntax, and most importantly the domain model. This model includes the identification of the conceptual entities for the domain, which must be described to match metadata functional requirements.

According to Qin, Ball and Greenberg [88], it makes sense to develop specific goal-oriented metadata models for user tasks, as a smaller and more specific scheme will likely increase their adoption. In their functional and architectural requirements for scientific metadata, these authors purposed metadata attributes as building blocks that form a comprehensive representation of data or information objects, as can be seen in Figure 7.

The identity metadata at the top, includes identifiers for several entities linked to the research process or studies, such as researchers, organizations and publications, which may be identified by a standard identifier system. The second layer represents semantic metadata, mainly used for subject identifiers or as a mechanism for linking data with similar subject content. For this purpose, semantic tools like taxonomies, thesaurus, universal classifications or ontologies allow the desirable flexibility in data representation. Scientific context, geospatial, and temporal metadata ensure the requirements for data verifiability, replicability and reproducibility. Despite their

[23] https://www.fgdc.gov/metadata/csdgm/
[24] https://dwc.tdwg.org/
[25] https://www.dublincore.org/specifications/dublin-core/singapore-framework/

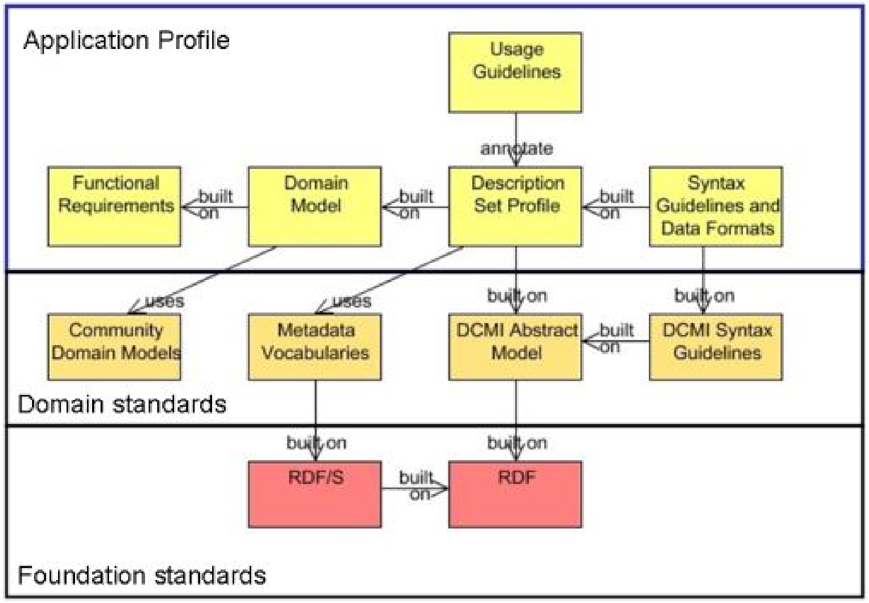


Figure 6: The Singapore Framework: Components of a DC Application profile

association with identity metadata, the scientific, geospatial, and temporal metadata, can be separate units. Finally, miscellany metadata includes elements that do not fit into any other block.

Metadata standards are important tools for normalizing the description of research data, to foster interoperability and to enable the discovery and use of data. However, large, complex metadata standards, can make it hard the creation, sharing, and might incur substantial costs for scientific metadata operations [89]. An analysis of several metadata standards, by Willis, Greenberg and White [128], corroborated the idea that metadata standards are not developed with principles of simplicity and sufficiency, and with a minimal set of essential domain elements in mind. The authors have analysed 9 metadata standards for the description of research data that were used in active data repositories, and verified the frequency of metadata goals and objectives across these standards. Most standards, fulfil requirements associated with scheme extensibility, data interchange, data documentation and retrieval. However, only 4 standards identify a minimal set of essential domain elements for data documentation, and only 2 standards take into account the levels of technical expertise of their community and support those with minimal tools and resources.

Given the fast-pace growing of scientific data, it has been argued that the conventional approach in developing metadata standards are out of date. Metadata standards result in a high number of elements, with complicated linguistics and syntantic forms that makes metadata standards complex to adopt and required highly trained professionals to create standard-compliance metadata records [89]. A possible approach to solve issues related to large, complex metadata standards, according to Qin et al., [90] is to break metadata standards into independent modules to allow for reuse of elements and maximal possibility of automation. Moreover, is proposed that this strategy can be implemented with a metadata infrastructure containing elements, vocabularies and other easy to use metadata artifacts. Considering that the same elements co-occur across different standards, often with

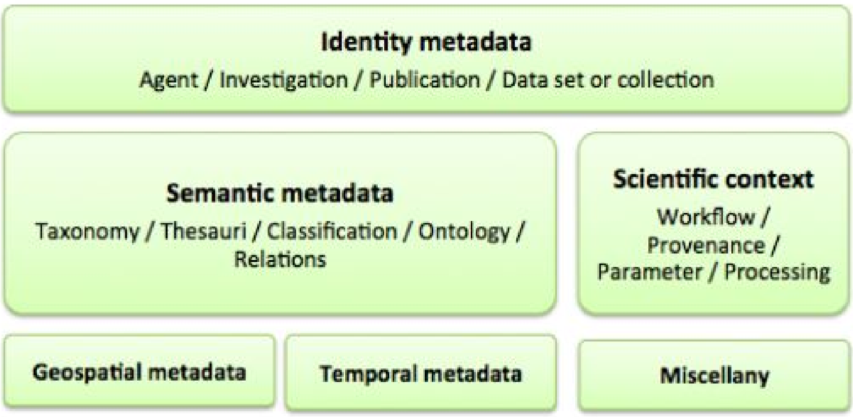

Identity metadata
Agent / Investigation / Publication / Data set or collection
Semantic metadata
Taxonomy / Thesauri / Classification / Ontology / Relations
Scientific context
Workflow / Provenance / Parameter / Processing
Geospatial metadata
Temporal metadata
Miscellany

Figure 7: An architectural view of metadata requirements [88]

inconsistencies and varying name conventions [89], ontologies can be a suitable approach to mitigate this issue.

4.3 ONTOLOGIES FOR RESEARCH DATA

Ontologies have been recognized as essential tools for the description of resources on the Semantic Web [14], as they are knowledge representation structures that have the ability to capture the meaning of each descriptor used in a metadata record in a machine-processable way. A dimension of the convenience of adopting ontologies for RDM is that they can promote the necessary vocabulary agreement [79], thus establishing the common semantics of concepts shared by distinct entities.

Ontology representation serves as a means to obtain the expressive, accurate and unambiguous syntaxes, desirable in today's research data production contexts [65]. Being aware of the advantages of adopting ontologies for the description of research assets, many scientific communities are working to deliver research-oriented ontologies. A good example is the CERIF [57], which is being adopted by numerous organizations and was first intended as a data exchange format for records describing projects. Despite its initial rigid format, CERIF has evolved to provide richer semantics, while its core set ensures interoperability between records. Another example is EXPO[26], an ontology that formalizes generic knowledge about scientific experimental design, methodology and results representation. According to the developers of EXPO, its main advantage is the fact that generic knowledge about experiments can be organized consistently in only one place, to ensure clear updating and non-redundancy [108]. For a more specific application, the Earth System Grid (ESG)[27] and the Extensible Observation Ontology (OBOE)[28] have been developed for the earth sciences and ecology disciplines, respectively. The ESG ontology has search and retrieval of datasets as its primary function [87], while OBOE is a formal ontology for generic scientific observation and measurement semantic representation [68]. Despite the fitness of these ontologies to model the knowledge structure of their do-

[26] http://expo.sourceforge.net/
[27] https://www.earthsystemgrid.org/
[28] http://agroportal.lirmm.fr/ontologies/OBOE

mains, one may argue that they are too fine-grained, particularly ESG and OBOE, to be of practical use in an operational data management workflow targeting data description.

Linked Open Vocabulaires

With a large number of available, and scattered, ontologies on the web, there is purpose for services that facilitate their discovery and retrieval. Linked Open Vocabularies (LOV)[29], is a portal intended to support the sharing of vocabularies in the web. It provides access to several hundred vocabularies, based on quality requirements including URI stability and availability on the Web, use of standard formats and publication best practices, quality metadata and documentation, identifiable and trustable publication body, as well as proper versioning policy. The deposited vocabularies have to match the LOV catalog quality requirements, namely: they should be written in RDF; be parseable without errors; all terms should specify an *rdf:label*; contain basic metadata; and should reuse relevant vocabularies.

4.4 SUMMARY

The commitment to RDM is important for many reasons: not only does it improve the chances of reproducibility and verifiability of the research results, but the advantage of promoting data reuse also decreases data duplication and the inherent efforts to produce new research data. This allows researchers to directly focus their work in the project's specific goals, leaving more time to pursue an extensive validation or other research activities.

RDM workflows involve both practical and technical issues. Sound technological solutions have been presented to reduce the technical issues, and we have seen great progress in that regard; solving the practical issues, however, depends on fostering the interest of researchers to be active stakeholders in the RDM workflow, more precisely in the description of their data. Data description assumes a critical nature in this workflow as it enables researchers with interest in a dataset to find and reuse it. Thus, the dissemination and preservation of research data strictly relies on metadata [118]. Practical and technical issues are, therefore, related, and the need for high-quality metadata drives the technical developments often seen in RDM infrastructures.

However, data description is demanding and time-consuming, let alone the research process itself, so researchers have to progressively invest in metadata creation, dealing with it during their daily activities. This involves producing data from diverse sources and extracting their production context, which is often kept in laboratory notebooks. By nature, such records follow an unstructured approach, strongly dependent on the researchers' perspective. When this is the case, researchers generate metadata that may lose their value upon the project's closure, as their interpretation can be problematic for external parties.

It is becoming clear that researchers must be perceived as central players in data description tasks, as long as they are motivated to assume such a role. Nevertheless, the short-term benefits of this activity are not always tangible for them. Postponing data description to the end of a project's cycle - when

[29] https://lov.linkeddata.es/dataset/lov/

researchers naturally become focused on other projects - is very likely to yield poor metadata.

Metadata has been described as an important instrument for RDM, particularly by allowing data description in the first place, and its later retrieval and interpretation. Metadata production comes at a cost, and technologies have created the conditions for the upsurge of instruments that can be of great value to create metadata records. The scientific community is increasingly conscious of the potential of these instruments, hence efforts have been made to deliver the vocabularies that improve the chances of data reuse.

A possible solution would be to have data curators accompany the research workflow—however, small research groups may struggle to keep up with the description demands posed by the existing datasets, and it is not expected researchers to spend much time in data description activities. To describe their datasets, researchers need to know what metadata to include in the descriptions, something usually prescribed by metadata standards. However, these are often too complex in order to fulfil different metadata requirements [89], making the description process too time-consuming and diverting researchers from their main activities.

It is therefore essential to address this challenge, by providing researchers with metadata models that are easy for them to adopt, in order to reduce entry barriers to metadata creation.

Part II

Researchers engagement in metadata studies

5 A SURVEY ON RESEARCHERS ENGAGEMENT IN METADATA-DRIVEN RDM STUDIES

It is assumed that researchers in the long-tail are the main stakeholders in realistic RDM. Yet, the limited metadata quality and low levels of research data reuse are most likely linked to the lack of engagement of researchers in RDM activities. In this Chapter I outline a summary of RDM studies that use at least one technique to engage researchers in the development of tools, or to improve and assess their metadata practices. Studies carried out by disciplinary experts in which the authors themselves are taken as participants, or report on their own reality, are not considered in this survey.

To this end I searched the Scopus database[1] and obtained 219 RDM entries that feature the concept "metadata" in the title or as a keyword. Then, a double review process was conducted to determine the studies that have engaged researchers. For a broader coverage of publications I also assessed the list of 301 publications provided by the Perrier et al. scoping review [84]. The final corpus of analysis consisted of 14 studies that were coded according to their scientific domain, number of participants, study motivation, metadata context, methodological approach, metadata practices of participants, their main findings and recommendations.

Overall, metadata creation is not a generalized practice among researchers, and when it does it is mostly for personal purposes or influenced by the software in use. The interview is the most commonly used technique to engage researchers, followed by questionnaires. The adoption of flexible and comprehensive data description tools and the collaboration between researchers and data curators are highlighted as general recommendations to foster the production of metadata by researchers.

This survey makes it possible to assess the range of techniques used to involve researchers in metadata-driven studies. The dataset that supported this Chapter is available[2].

5.1 SAMPLING PROCEDURES

The sample universe was established by querying the Scopus database on entries pertaining metadata-related studies in the RDM context. As such the search expression was refined and finally set as a conjunction of conditions, as follows:

[1] scopus.com

[2] Castro, J. A. (2019). Research Data Management metadata-driven studies survey corpus and coding. INESC TEC research data repository. https://doi.org/10.25747/rv0g-1n12

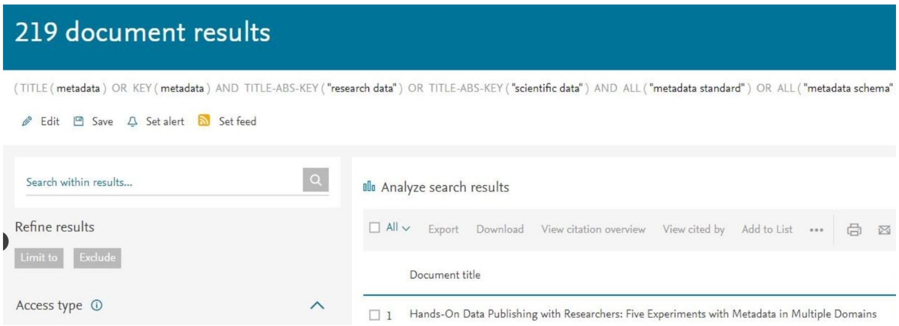
(1) metadata as the "title"or "keyword", to make sure that metadata is the focus of the study **AND**;

(2) research data or scientific data in the "title", "abstract" OR "keyword" to provide the context of the studies **AND**;

(3) metadata standard/schema(e); ontology or vocabulary in all fields, to specify the nature of the metadata.

Alternative search expressions, like having "research data management" as a concept, have returned a small set of entries giving the specificity of some expressions tested and the fact that the vocabulary in the RDM area is still being established. On the other hand, with a more flexible search expression I obtained a huge number of results that would become impractical to explore. Some exploratory queries returned over 1000 publications and quick checks revealed that many of the records were out of the RDM scope. In the end I was satisfied with the search expression used and its results. I noticed that all of the metadata publications that fit the criteria and that I was previously aware of appeared in the results.

Using this strategy a total of 219 entries were retrieved on 11 February, 2019 (Figure 8)[3] and I exported the *BibTeX* file containing all the bibliographic information and the *abstract*.



219 document results

(TITLE (metadata) OR KEY (metadata) AND TITLE-ABS-KEY ("research data") OR TITLE-ABS-KEY ("scientific data") AND ALL ("metadata standard") OR ALL ("metadata schema"

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

Figure 8: Number of entries retrieved from the Scopus database

After a brief review of the *BibTeX* dataset I excluded 18 entries that were either out of the RDM context, duplicates or conference proceedings summaries.

The resulting set of 201 entries was then evaluated in a double review process to determine which studies have used at least one technique to engage researchers. For this process I was assisted by a Ph.D. student involved in RDM, as a reviewer. Each of us read the abstract of all the entries. In each entry both reviewers annotated, in separate spreadsheets, if the abstract suggested the participation of researchers - if so (*Yes*), if not (*No*), and in doubt (*Maybe*). The annotations were then compared to check agreement. If one of the reviewers annotated with an *Yes* and the other with a *No* or *Maybe*, then the publication in question was read in full. However, if an entry was annotated with a *No* and a *Maybe* it was excluded. This decision was tested by reading some publications in this situation and confirming the absence in the studies. Moreover, 4 papers identified as positive in the review process where excluded, given that, although pertaining to a domain case study,

[3] For readability purposes the figure does not capture the search expression entirely

they were based in the authors' intrinsic knowledge and the full publication content did not confirm the use of a given technique to obtain data from participants.

Figure 9 gives and overview of the distribution of the 201 RDM related metadata studies over the years.

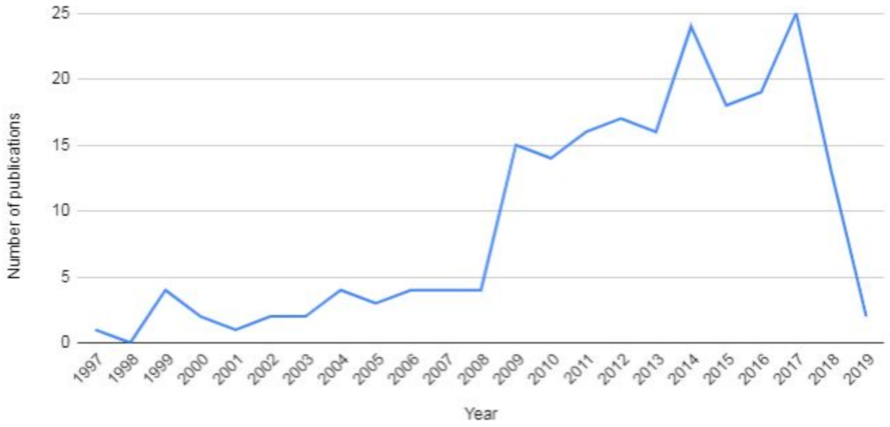


Figure 9: Distribution of RDM metadata related studies over the years

The oldest publication in my data is from 1997, and until 2008 the number of publications identified is limited. In 2009 there is a leap and the number of publications is 15. From there on the cadence of publication of metadata studies in RDM is higher. The peak was achieved in 2017 with 25 publications. As the search was performed in early 2019 the number of results for this year is small as expected. It is plausible that the emergence of RDM in recent years has led to an increase in the number of studies, to more works being accepted due to an increased sensibility of reviewers and also to more scientific events dedicated to this topic.

The final sample consists of 19 publications. Yet, I have been directly involved in 8 of these publications, and therefore these were excluded from the corpus of analysis. This decision has to do with the fact that the work described in these publications are part of the proposed data curator's workflow and its various use cases are described in Chapter 6, as well as other studies carried out by the TAIL team that follow the same approach.

For a broader scope I consulted the dataset that list the 301 publications included in the Perrier et al. scoping review [84], available at Plos One[4]. The purpose of this scoping review was to describe the volume, topics, and methodological nature of existing literature on RDM in academic institutions. The authors searched 40 literature databases encompassing several disciplines. The literature search resulted in 15,228 entries. In their study, after reviewing titles and abstracts, 654 potentially relevant full-text articles were retrieved, from which 301 publications were selected to include in the scoping review. The extent of the Perrier et al. review is very broad and therefore, in combination with my data, gives a high degree of confidence regarding the comprehensiveness of the literature mapped for this survey. Scanning the Perrier et al. bibliography for publications satisfying my criteria, I found three new publications that were added to the final corpus of this survey.

[4] https://doi.org/10.1371/journal.pone.0178261.s003

5.2 SURVEY CORPUS

A total of 14 publications made up this survey corpus and are listed in Table 1. I assigned a label to identify each publication as a way to refer to them when necessary. The publications labeled LabTrove, Archaeological Digging and Publishing Pushing were added from the Perrier et al. work.

Table 1: Survey corpus

Label	Publication
SPECTRa	SPECTRa: The Deposition and Validation of Primary Chemistry Data in Digital Repositories [37]
Institutional Influences	How institutional factors influence the creation of scientific metadata [73]
Personal Organization	Considering personal organization: Metadata practices of scientists [121]
Institutional Issues	Research Data and Metadata Curation as Institutional Issues [74]
BioWes	BioWes - from design of experiment, through protocol to repository, control, standardization and back training [29]
Organizing Behaviors	Descriptive Metadata for Scientific Data Repositories: A comparison of Information Scientists and Scientist Organizing Behaviors [122]
Linked Data	Linked Data for the Natural Sciences: Two Use Cases in Chemistry and Biology [126]
Site-based Geobiology	Site-based data curation on hot spring geobiology [82]
Swedish Case	Research data services: An exploration of requirements at two Swedish universities [62]
Metadata Workflows	Metadata workflows across research domain: Challenges and opportunities for support in the DFC cyberinfrasctructure [80]
Wiser	The WISER metadatabase: The key to more 100 ecological datasets from European rivers, lakes and coastal waters [106]
LabTrove	Creating Context for the Experiment Record User-Defined Metadata: Investigations into Metadata Usage in the Labtrove ELN [129]
Archaeological Digging	The challenges of digging data: A study of context in archaeological data reuse [44]
Publishing Pushing	Publishing and Pushing: Mixing Models for Communicating Research Data in Archaeology [58]

The second phase encompassed content analysis based on the full text of the documents in the corpus using the Catma 6.0 online tool [5] for the coding process. Given the exploratory nature of this study I did not use any pre-established coding scheme for text markup. The coding categories in Table 2, were defined on the fly after reviewing a sample of the corpus. Yet, not all the publications satisfy all the code categories in the same way. Some papers are more focused on describing a particular methodological approach, while others are oriented to present results and only vaguely mention the technique applied.

Table 2: Survey code categories

Code categories	Definition
RDM context	The RDM event that trigger the study
Motivation	How institutional factors influence the creation of scientific metadata
Participants Domain	The research domains studied.
Metadata context	The perspective from which the metadata is explored, e.g. vocabulary development
Methodological Approach	Technique(s) that mediated data collection for the study
Metadata Practice	Activities, or lack of, performed by the researchers to describe their data
Finding	Major conclusion(s) from the study
Recommendation	A suggestion on how to improve RDM results

[5] https://catma.de/

5.3 CORPUS DESCRIPTION

Considering the context that motivated the publications, the corpus can be divided in two clusters. These two clusters focus on: **1)** the development or evaluation of RDM platforms or tools; **2)** exploratory studies or participants metadata practices assessment.

The following publications are dedicated to the development or evaluation of RDM tools. The motivations are diverse and include the proposal of an institutional data repository with domain-specific metadata, the evaluation of a metadata infrastructure, the proposal of an Electronic Laboratory Notebook (ELN), and the improvement of data quality in a publication workflow. But there are also publications related to domain-specific vocabularies; namely a collaborative approach for the production of domain Linked Data and a framework consisting in minimal information to record metadata.

SPECTRa: Published in 2008, the main goal of this work was the development of an institutional data repository to address data loss in the chemistry domain. Thus, the practices of chemists in archiving and disseminating primary chemical data were explored. The authors reported the use of standardized chemical metadata to enable long-term data reuse.

Wiser: Within a EU-funded project a metadatabase was designed to summarize the available data from research in Europe, therefore providing researchers with a way to find optimal data for their analysis. The authors reported in 2013 on the evolution of the metadata base, while evaluating if generally used metadata standards were suitable for some of its specific purposes. In this platform metadata ideally comprise the necessary information to enable data reuse.

BioWes: This work, published in 2016, describes an RDM platform that was proposed to mitigate the lack of data and metadata management infrastructures. This platform was meant to support researchers from protocol design to data sharing. The system was tested in the framework of an international project.

LabTrove: In this 2014 publication the authors presented a Web browser-based ELN that enable users to add their own defined metadata to notebook entries to describe entries and experiments. The argument in favour of the adoption of such tool is that capturing experiment metadata early with an ELN facilitates data curation.

Linked Data: The aim of this work, published in 2013, was to present a methodology focused on the collaboration between data curators and researchers to translate scientific domain knowledge into Linked Data. This study was instantiated with two use case scenarios in the chemistry and biology domains.

Site-Based Geobiology: This 2017 publication reported an use case from a data curation project where the goal was the development of an approach for retaining the value of digital data collected, from scientifically significant sites, for reuse across disciplines. The main challenge was the existence of few criteria to guide the production and management of open datasets to as-sure their fitness to reuse. Since systematic documentation is a requirement for geobiology data reuse, a framework for recording the minimal necessary metadata was proposed.

Publishing Pushing: This case study, carried out between 2012 and 2013, involved archaeological researchers that published data in an online platform and reported on how different editorial and collaborative review processes improved data documentation and quality. Data publishing practices were discussed to better understand different data management needs.

The following studies are more exploratory in nature. In general, the aim was to know the practices of researchers to help decision making in the improvement or development of RDM services. The following studies generate knowledge that can be translated to several realities, no matter how specific the domains in some of them may be.

Institutional Influences: This publication describes a distributed ethnographic study of laboratory and field work research groups, carried out in 2011, with the aim to demonstrate how "frictions" arise in creating and managing metadata. These metadata frictions are challenges and problems that arise during the creation, handling, management and preservation of metadata products [40].

Institutional Issues: Based on a prior conceptualization of institutions, the author proposed a theoretical framework outlining five "institutional carriers" for data practices, corresponding to norms and symbols, intermediaries, routines, standards, and material objects. This 2015 publication describes the application of the framework in three case studies to assess how institutional support extended through a single organization or specific disciplines.

Personal Organization: This paper assumed that information professionals need to understand the personal metadata and research data organization practices of researchers to address the challenge of integrating research data and datasets into library collections. Therefore, the author gathered information about organization practices and perceptions of researchers, in 2008.

Organizing Behaviors: By examining and comparing how information professionals organize their data for personal use and deposit the authors' aim was to obtain insight on how to improve repository systems designed to accommodate the special needs of scientific data sets. This work was published in 2008.

Swedish Case: This exploratory study of researchers' needs regarding RDM was conducted at two Swedish universities between 2015 and 2016. The goal was to inform the development of adequate RDM services, since these services, although emergent, were not yet generalized. The author identified characteristics, requirements and needs related to RDM issues.

Metadata Workflows: The goal here was to study the role of people, and of automated processes, in the creation of metadata during the data life cycle. This work was outlined in a poster presented in 2014, and its underlying motivation was to know what could be achieved to improve the quality of metadata workflows in general.

Archaeological Digging: Between 2011 and 2012 the authors examined the data reuse practices of archaeologists to consider how metadata standards might be extended to preserve the meaning of cultural heritage materials.

These metadata-driven studies took place in a diversity of scientific disciplines and also had a distinct number of participants, as shown in Table 3.

Table 3: Assessment of domains and number of participants, by publication.

Publication	Domain	Participants
SPECTRa	chemistry	Not clear
Institutional Influences	geosciences; biodiversity; ecology physics; network sensing	over 100 RDM stakeholders
Personal Organization	evolutionary biology	7
Institutional Issues	geosciences; biodiversity; ecology	33 (some staff members)
BioWes	biomedics	16 institutes
Organizing Behaviors	biology	16
Linked Data	chemistry; biology	2
Site-based Geobiology	geobiology, geochemistry; microbiology	9
Swedish Case	archaeology; biology; social sciences	12
Metadata Workflows	oceanography; social sciences; computer science	14
Wiser	ecology	multi-agency collaboration
LabTrove	chemistry; biology; physics	202 authors entries; ? interviews
Archaeological Digging	archaeology	22
Publishing Pushing	archaeology	12

The natural sciences are well represented with several studies in domains like chemistry and biology, as well as studies related to environmental and earth sciences. On the other hand, the social sciences, humanities and engineering domains seem under represented. For instance, only in the *Swedish Case* and *Metadata Workflows* publications is mentioned the participation of social scientists. However, in these studies the sampling was open to different domains. Archaeology is the only domain of the humanities featured in the corpus, as it is the focus of the *Archaeology Digging* and *Publishing Pushing* studies.

As to the number of participants it ranges from 2 in the *Linked Data* case studies to 104 notebooks analysed with entries from 202 researchers in the *LabTrove* survey, although the number of interviewed researchers in this study is not clearly stated. Likewise, the *Institutional Influences* ethnographic study had the participation of over 100 stakeholders, but how many of them were researchers is not mentioned. The publications in which the number of participants is not declared were conducted in the context of large projects or research infrastructures, in which a great number of participants is likely. This is the case of the *Wiser* and of the *BioWes* studies. The former corresponds to a multi-agency collaboration within an EU project. The *BioWes* survey included 17 project partners, from which 16 were research institutes and one was a private company. Moreover, these studies have used questionnaires as the preferred technique, which does not involve face-to-face contact, thus enabling them to reach more participants.

5.4 TECHNIQUES TO ENGAGE RESEARCHERS

Interviews and questionnaires were the most popular techniques to engage researchers in the surveyed RDM metadata-driven studies. As shown in Table 4 the interview was applied in 7 of the 14 studies and questionnaires in 5 of them. Moreover metadata creation of some sort was explored in 4 studies. Content analysis has also been applied on 4 publications. This is a technique that does not involve researchers directly, yet it should be taken into account because it allows data on participants to be obtained. Other techniques applied to involve researchers were meetings, workshops and usability testing.

Table 4: Techniques to engage researchers in the surveyed studies.

Publication	Interview	Questionnaire	Content analysis	Other
SPECTRa	✓	✓		
Institutional Influences	✓		✓	observation
Personal Organization	✓			
Institutional Issues			✓	observation; meetings
BioWes		✓		observation
Organizing Behaviors		✓		metadata creation
Linked Data	✓		✓	meetings
Site-based Geobiology				workshop
Swedish Case	✓			
Metadata Workflows		✓		
Wiser		✓		metadata creation
LabTrove	✓		✓	usability testing
Archaeological Digging	✓			
Publishing Pushing			✓	workshop

Interviews in the *Institutional Influences* ethnographic study took place in three research sites and included 100 researchers including staff members, software developers, and data managers. The aim of these interviews was to assess data sharing, collaboration, research processes and their outputs. Moreover, several of the participants´ publications, as well as individual and project web sites, were analysed as supplementary data. The *Institutional Issues* study was framed in the context of the latter, therefore it may report the same interactions. In this case the authors mentioned 14 semi-structured interviews, averaging 43 minutes in length, in one research site. The participants were recruited via snowball sampling, where interviewed people named other people involved in their projects to participate. The observations reported in these two publications amount to 200 hours and 16 trips to lab or field settings in one of the research sites. In another research site the author said to have held more than 30 meetings with lab heads and researchers. For the third research site the author focused on data practices reported in several studies described in the literature.

Snowball sampling was also employed in the *Archaeological Digging* study, in this case combined with convenience sampling. The authors started with people associated with their collaborators and moved on to recruit more participants in scientific events. In the end 22 semi-structured, hour-long interviews, focusing on data reuse experiences, were conducted.

To prepare the *Organizing Practices* interviews, the participants had to choose a publication for which they still owned the related datasets, while the interviewer read the publication for a better understanding of the dataset underlying the conversation. Since the interviews took place in the participant's personal office or lab it was easy to consult the dataset during the interview. The 7 participants of this study were recruited by convenience sampling upon recommendation of repository team members. The interviews duration ranged from 15 minutes to one hour. The *Swedish Case* authors based their interviews in the Data Curation Profile Toolkit [19] with additional metadata-oriented questions to collect data, specifically to know what were the practices of using subject metadata to describe their data. The interviews with the 12 participants lasted between 46 and 119 minutes.

The *Linked Data* author applied the interview to identify the research interests of the participants and as a way to introduce them to Linked Data. Ideas and expectations were exchanged in order to build a common understanding of the goal and scope of the *Linked Data* project.

Questionnaires were applied in the *SPECTRa*, *BioWes*, *Organizing Behaviours*, *Metadata Workflows* and *Wiser* studies, but in some cases the approach is not detailed. In the *Wiser* study an online questionnaire was developed so that its entries would populate the metadata base. Project partners were required to complete one questionnaire for each dataset contributed. According to the authors, the user interface of the questionnaire was built using standard web technologies, and to facilitate data entry most of the fields were designed as check boxes, radio buttons or selection lists. Moreover, comments fields were available for additional information and a handbook with supplementary information was compiled to support the participants. A complementary questionnaire with 28 questions was also devised to evaluate the usage of computers and of the Internet. The questionnaires in the *Organizing Behaviours* study were used to gather information about participant demographics and processes. After the deposit and description of data in a repository the author applied a follow-up questionnaire to gather insight of information organization issues. In order to do so, 22 codes were created to highlight these issues in the narratives and short answer responses from the follow-up questionnaire.

To gauge the patterns of metadata usage in an ELN the authors of the *LabTrove* study have surveyed more than 100 notebooks from diverse disciplines. The development team undertook a variety of activities to infer user behavior and attitudes towards metadata. Usability tests with new users and trials with students were combined with interviews with people interested in adopting the ELN. According to the authors, these activities provided the opportunity to examine user expectations, their understanding about metadata and how the ELN design might affect their metadata use.

In the *Site-based Geobiology* and in the *Publishing Pushing* studies the authors have organized workshops. The latter has brought together researchers to publish and integrate data from 12 archaeological sites, to explore the challenges of data reuse. Participants then analyzed subsets of the integrated data and presented their results to the group. Moreover each participant addressed a specific topic using a subset and later presented their feedback in a workshop, so that data contributors could improve the quality of their publications. As for the *Site-based Geobiology* study, a panel of experts was enlisted to solve problems through a process of consensus development. The process started with a stakeholder workshop that enabled the authors to gain knowledge of their practices in the field, expectations for data quality and reusability, opinions regarding data sharing, and initial criteria for a minimal information framework. On top of that, a two-day workshop, designed to interrogate data value and reuse factors through a set of roundtables, exercises, and focus groups, laid the foundation for the engagement of researchers with the resource management personnel from the site. Content analysis of research artifacts and participatory engagement after the workshop led to development of the framework for recording minimal metadata. This study also included a trial run in the field, using a custom data entry template rooted in the framework. This made it possible to observe students entering data into the template, and their feedback informed the final revisions of the metadata framework.

5.5 METADATA PRACTICES OF STUDY PARTICIPANTS

It is not surprising that metadata practices are not uniform and systematised among the participants in the various studies. A generalized conclusion is not possible, in most cases, since the studies reported various approaches that participants employed to create metadata. Many studies described one-off practices, where one participant, or group, may use one technique while the others may use another. For instance in the *Archaeological Digging* study, the availability of context for archaeology artifacts in older and contemporary research projects was considered uneven. The metadata procedures in older research projects were described as ranging from meticulous to sloppy, while in contemporary work the challenges were related to the transition from paper-based to digital recording procedures.

In the *Swedish Case* the number of approaches to describe data was identified as diversified, encompassing metadata supported by metadata schemes, codebooks, a paper filing system using plastic folders ordered geographically and by topic, and tables in spreadsheet files. In many research fields, the lack of a common integrated data infrastructure often results in non-standardized, local data management practices. Participants from the *Metadata Workflows* created metadata at different stages of the data collection process, but it is more likely that metadata are created manually after data collection. Two participants reported that computer-generated metadata is created before data collection and 9 stated that this occurs during or after data collection. Standards like DC and EML were mentioned by the participants as the tool to support metadata creation. The *Institutional Influences* study verified that standards are often rejected due to their complexity, or lack of technical support. In some cases researchers were unaware of applicable standards. Therefore, the authors concluded that most practices were *ad-hoc* with only a few researchers using metadata standards. However, one of the research sites is responsible for the development of the EML standard, and this standard was consistently applied in one domain. The *BioWes* participants were found to occasionally use standard terminology, or terminology specific to the group, in the description of experimental conditions. Nevertheless, missing metadata and RDM tools decreased the potential of experimental reproducibility.

The authors of the *Site-based Geobiology* study observed adequate processes for cataloging and tracking of physical specimens and collected artifacts, but not equivalent procedures for recording, collecting and preserving data. Moreover, researchers showed great awareness for each other's work but there was not much collaborative and coordinated work with collective data resources. In the *LabTrove* study the authors pointed out an interesting observation about the metadata records in public and private notebooks. The reluctance to make data public is a known inhibitor to metadata use, however notebooks with privacy settings had some of the highest average figures for metadata use. Overall, some groups of researchers were found comfortable with metadata and were able to produce effective metadata structures, yet most researchers only recorded the minimal metadata required by the notebook.

Personal guidelines were part of the metadata processes of the participants in the *Organizing Behaviors* study, despite the influence of certain standardized policies in information organization techniques. These personal guidelines consisted of personal preferences or long-term habits in metadata creation. Nevertheless, the same author acknowledged in the *Personal Organization* study that practices would change depending on whether the data was destined for personal purposes or intended for a larger audience.

5.6 final remarks

The surveyed publications shed light on the reality of researchers in different RDM scenarios related with metadata. Their findings enable to draw general recommendations to improve tools and services in which metadata are, in one way or another, pivotal. A myriad of metadata challenges have been spotted in these studies with consequences to metadata creation.

Reluctance to provide detailed metadata to enable data reuse was associated with the lack of financial resources and to the fact that this is a time-consuming activity in the *Wiser* publication. Additionally, it was also mentioned that some participants believed that even making their metadata available means giving up on their intellectual property rights, and that data might be used for incorrect purposes. Others have verified an interest of researchers in the potential of LOD, but few were willing to invest their data, time and knowledge to cooperate in such effort (*Linked Data*). Lack of awareness also emerged as a limitation to metadata creation, particularly the lack of knowledge of the availability of institutional repositories (*SPECTRa*), as well as unawareness regarding the benefits of creating metadata (*LabTrove*).

Furthermore, the *LabTrove* authors also found that the absence of a defined metadata schema is an inhibitor for metadata creation. In what they called a "black-canvas effect", where users may be willing to add metadata but do not know where to start. Therefore, the authors argue in favour of need to provide researchers with more assistance to help them create appropriate metadata for their experiments, considering that the best way to deal with the "black-canvas effect" is to start with generic metadata models that can be extended to meet the requirements of the research groups. Likewise, the *Publishing Pushing* authors suggested that curators developing ontologies and controlled vocabularies have to be responsive to community needs, arguing that without a flexible approach vocabularies would be of minimal interdisciplinary applicability for data integration and linked data applications. Other argue that although the development of domain vocabularies is not the primary objective of data curators, they still need to have a good understanding of methodologies for the creation of ontologies (*Linked Data*).

The collaboration between the data curators and researchers was also a recommendation that stood out. Training researchers to correctly describe their data has been indicated as a valuable service (*Swedish Case*). There is a need for appropriate RDM and metadata training to overcome the lack of knowledge about what metadata is and how to use it (*LabTrove*). Researchers and data curator's have distinct approaches to metadata creation (*Organizing Behaviours*). Without formal training in metadata creation researchers devised their own organizational schemes and metadata creation policies to suit their daily challenges, and their metadata is more focused on the details when compared to the metadata created by the data curators. According to the *Personal Organization* author, repository staff and other collaborative services could benefit from getting insight of personal metadata and organization practices of their potential contributors.

This collaboration is also important when supporting researchers in data deposit. The *Swedish Case* authors see benefits in combining the possibility for researchers to submit data themselves, with a service where curator's can deposit data in their behalf, hence promoting data sharing for researchers who want control during the process and for those who do not have the time to do this themselves. Moreover, through training, researchers would be incrementally more self-sufficient in data sharing. For the *SPECTRa* authors, the responsibility to capture data is primarily of the departments, but with liaison with the libraries or other services. In the *Archaeological Digging* study the authors concluded that the transparency in data collection and curation procedures, along with the amount of metadata that data repositories hold, contributes to the overall reputation of the repository, which helps to increase the perceived quality of a dataset by its users.

To sum up, the well grasped notion, by the *Institutional Influence* authors, that the "*one size fits all*" approach to data and metadata management will not work due to the variability across disciplines, projects and data types, is a remark that applies after reading these publications. Their categorization of metadata frictions showed that institutional support for RDM is not consistent. In the "long tail" they have found great support for a small number of projects and many project with scarce support. At the level of individual disciplines the support is likelier even smaller, but still not uniform. The authors recommended that iSchools can contribute with expertise and research to multiple kinds of research institutions, and help to make data and metadata management efforts responsive to differing needs across institutions.

The increase in the number of RDM publications related to metadata over the years, taking into account the retrieved entries from Scopus, seems in line with the different policies that have come into force in recent years. However, of the 201 valid publications returned, only 19 were assessed has having directly involved researchers with at least one technique. My opinion is that researchers, as data creators, and main RDM stakeholders, should have a more effective participation in related studies, even considering other potential studies who were not included in this survey. Nevertheless, the surveyed studies yielded valuable insight to underpin my work contributions. In the next Chapter I outline my proposal to engage researchers in a data curator's workflow mainly dealing with the development of domain-specific metadata models.

Part III

Engaging researchers in the data curator's workflow

6 INVOLVING DATA CREATORS IN AN ONTOLOGY-BASED DESIGN PROCESS FOR METADATA MODELS

Adequate RDM strongly depends on accurate metadata records that capture the production context of the datasets, thus enabling data interpretation and reuse. Research domains are diverse in nature and comprise very specific concepts, making it necessary for researchers and data curators to work together in order to describe datasets. This is even more prevalent in the long-tail of data. This problem is aggravated in the current context of massive data creation, particularly in research groups with access to limited resources [16].

This Chapter consists in the proposal of the data curator's workflow for the development of domain-specific metadata models. The focus is to foster the collaboration between RMD stakeholders, namely between data curators and the researchers.

I start by introducing the role of data curators and researchers in data description, and then I detail the data curator workflow with the steps in the development of the domain-specific metadata models. Since the metadata models are formalized as lightweight ontologies, I explain the approach behind the development of the lightweight ontologies. The data curator workflow is instantiated with 4 case studies, carried out by different TAIL members who assumed the role of the data curator. These use cases are in the following domains: Vehicle Simulation; Hydrogen Production; Biological Oceanography and Social Sciences.

This Chapter is based on the following book chapter [20]:

- Castro, J. A., Amorim, R. C., Gattelli, R., Karimova, Y., da Silva, J. R. and Ribeiro Cristina (2017). Involving Data Creators in an Ontology-Based Design Process for Metadata Models. *Developing Metadata Application Profiles*. IGI Global.

The lightweight ontologies development approach was implemented during the development of metadata models for the Fracture Mechanics and Pollutant Analysis domains, detailed in the publication [23]:

- Castro, J. A., da Silva, J. R. and Ribeiro, Cristina (2014). Creating lightweight ontologies for dataset description. Practical applications in a cross-domain research data management workflow. *In IEEE/ACM Joint Conference on Digital Libraries (JCDL).*

As the number of use cases grow I felt some limitations with the communication with researchers, so I seek complementary approaches to this workflow. These approaches are addressed in the following chapters. Chapter 7 will detail the process of applying content analysis to improve the communication with researchers and in obtaining candidate concepts for descriptors. Chapter 8 refers to a use case where an existing standard was adopted to involve several researchers from the same scientific institution in data description.

6.1 THE ROLE OF RESEARCHERS AND DATA CURATORS IN DATA DESCRIPTION

Academic institutions are ideal backgrounds for providing RDM services, and if some are already engaged in research data activities, others are considering doing so. In fact, data management services were pointed out has one of the top trends for academic libraries [113]. Ideally, institutions should provide the infrastructures and services to support data management, sharing and reuse [53].

Yet, institutions often lack the resources and struggle to support RDM requirements [96]. A possible solution relies on having data curators, or other information specialists, as stakeholders on these services, as they can be part of grant proposal teams as data curation consultants for example. In big data projects it is not unusual to have data scientists with domain expertise to perform data-related activities. Nevertheless, the same does not apply in the long-tail of science.

Data curators are aware of metadata best-practices and are becoming very active in this environment, as they can assist researchers to foster data dissemination by improving metadata quality [50]. Data curators and information scientists in general, can make good use of their skills, but in the long run their contribution can be less effective if researchers are not motivated to collaborate in the overall RDM process. Data curators have limited knowledge concerning domain-specific disciplines or research endeavours in general, being counter-productive for them to address data documentation activities in a wide variety of fields. Thus, the heterogeneity of scientific disciplines can prove to be overwhelming for data curators. Exclusively depending on data curators for data management can delay the whole process, considering that most institutions cannot delegate a data curator for each department or research group and, if this is done, it could yield unsustainable costs for small or medium-scale institutions. Therefore, despite their general metadata skills, by themselves, data curators will not be capable to provide timely metadata to face the fast pace at which research data is created, and eventually, this situation will result in a bottleneck in the research data workflow [130].

Given these circumstances, researchers should be considered key stakeholders in data description and in the development of data management tools. Taking into account their expertise in domain terminology and regular involvement in research environments, researchers, as data creators, are valuable candidates to produce accurate metadata records [34]. In this sense, collaboration between researchers and data curators is crucial, and both parties should co-exist in the development of the vocabularies to support metadata activities—researchers by providing insight on the domain terminology, and data curators, as information management experts, by working together to make datasets reach a larger audience.

Taking into account that every research domain, or experiment configuration, is likely to induce new data description requirements, a close collaboration between a panel of researchers working in different research domains at the University of Porto was assembled.

During my collaboration with researchers at the University of Porto, I had the opportunity to develop metadata models for several domains. In experimental domains we relied solely on an interview form, complemented with descriptors that researchers were able to suggest based on their perception. These were often influenced by the difficulties of reaching an agreement on metadata conceptualization.

This panel of researchers was broad enough to gather a collection of datasets that correspond to the different types of research data: experimental, observational and computational data. The research domains represented in the panel was diverse, including members from pollutant analysis, fracture mechanics and hydrogen production research groups, which generate experimental data. Other groups were working mostly with observational data and were related with an astronomy laboratory, biodiversity campaigns and social and behavioral studies. Other members of this panel worked in computational research environments, namely an operational research team that studied optimization problems, a research group evaluating the performance of electrical buses, and a team from the computational fluid dynamics area. Altogether, this research panel - and their datasets - provide a rich testing scenario for the definition of the metadata models that are convenient to apply in very particular data production environments, while taking into consideration a broader application and shared needs with other research contexts.

The main objective of this collaboration was to develop the metadata models that best suit the description needs of researchers, by selecting a set of descriptors that meet the daily terminology they are applying when working or communicating with their colleagues. Capturing familiar concepts that researchers can actually understand, and use in more casual descriptions, will likely mitigate existing barriers to data description processes, as it can reduce the complexity of using scientific metadata standards tailored for describing data at the end of the research cycle by trained professionals [88]. Hence, it is of utmost importance to promote the engagement between data curators and researchers in the definition of the metadata tools that will fit the latter expectations [69].

6.2 STEPS IN THE DEVELOPMENT OF THE DOMAIN-SPECIFIC METADATA MODELS

A first moment in the definition of the domain-specific metadata models is a meeting with both stakeholders, the data curator and the researchers. This initial meeting consists in an in-depth interview conducted by the data curator. The interview is a good methodological approach in this case since it provides the curator rich insight on domain knowledge. This knowledge includes information about the procedures and instruments that are used in research activities and data collection methodologies, in addition to how the research teams are handling, storing and sharing their data with their collaborators.

The interview is also the moment where researchers, in most of the cases, become aware of RDM problems and start to formulate them for the first time (struggling to find an illustrative in their archives; not being able to interpret a dataset originated by a colleague). During the interview, participant researchers also identify, by themselves, some RDM opportunities, such as the ability to partially disclose their data or consenting the access to the associated metadata record only, thus allowing other researchers, provided they are interested, to request access to the corresponding dataset.

I applied an adapted version of the Data Curation Profile Toolkit script, translated into Portuguese, in order to "capture requirements for specific data generated by researchers as articulated by the researchers themselves", helping in the data processes decision making, while being flexible to be applied to any scientific sub-domain. A good practice before running this interview is to allow researchers to read and consider the answer to each question. For instance, by sending the interview form by e-mail beforehand may result in more detailed responses. This script will be further described in Chapter 9, to provide the multi-domain data description sessions context.

With this interaction data curators will obtain domain knowledge from the domain expert vision, and the domain expert can also gain a perspective on RDM from someone with prior experience on this matter.

Furthermore, and to prevent researchers unavailability to participate full-time in the definition of the metadata models for their domain, I explored applying content analysis to the documents produced by the researchers, depending on their availability. In this work content analysis was performed manually in order to extract the main domain concepts to include in the metadata models as descriptors. The role of content analysis in data curation is evaluated in three use cases described in the next chapter.

Both the interview and the content analysis are useful tools to elaborate conceptual maps for the domain that are, in turn, at the core of the developed metadata models. To design a conceptual map is therefore the third step in the process to structure and formalize the knowledge of the data workflow for a selected domain. After the definition of the conceptual map, and having established the relation between classes and their properties, the key concepts are sought in metadata standards, particularly scientific ones, giving preference to those already formalized as ontologies.

After a selection of domain descriptors, another session is scheduled to propose those concepts to the researchers, and they are then also asked about the contextual information necessary to provide enough scientific evidence for others to verify, replicate, and reproduce the experiments from which the datasets were gathered. Finally the researchers are invited to validate the metadata model by evaluating the recommended concepts, and they can suggest new ones, remove or even rephrase the concepts. After the selection of domain-specific concepts, these concepts are soughted in existing ontologies and in controlled vocabularies (for example the IEEE thesaurus[1]). All the remaining vocabularies that could not be reused from existing vocabularies are introduced in a domain-specific namespace.

From my experience, the interaction with the researchers takes a total of three sessions. The metadata model is then formalized as a lightweight ontology.

[1] https://www.ieee.org/publications/services/thesaurus.html

6.3 LIGHTWEIGHT ONTOLOGIES

Working around the complexity of standardized metadata schemas, some authors have started to select sets of metadata descriptors suited for their particular application. This "mixing and matching" approach has yielded the notion of Application Profile [52]. However, even application profiles can be hard to understand and adopt; moreover, they are self-contained, meaning that they do not evolve incrementally and through reuse like ontologies. For datasets in a fast-paced, multi-domain research environment, a more incremental approach is desirable.

Ontologies have been presented as a possible solution for research data description. They satisfy all desirable metadata requirements [38] and are capable of representing the specific semantics of each research domain, while being able to evolve asynchronously. Yet, the attempts to model every aspect of each domain make it hard to use ontologies in an actual cross-domain RDM environment. OBOE [68] is an example of a domain model whose concepts are very specialized. Like many domain-specific ontologies, its modelling granularity is too fine for it to support a data management system. EXPO [108], is another case of correctness from a modelling perspective, but with a granularity making it unwieldy for usage in an operational data management workflow. Others like ESG [87] and CERIF [57] model cross-domain research concepts representing activities, entities or artifacts relevant for the research workflow that can be use for dataset description—for example *Experiment*, *Project* or *Participant*.

In the literature, ontologies with a large number of formal axioms and constraints have been defined as "heavyweight ontologies", while those with a simpler approach are called "lightweight ontologies" [63, 31]. DC[2], for instance, is currently a widely used lightweight ontology, on par with FOAF (Friend of a Friend). Their simplicity and weak constraints make them easily processable by machines, and both have been directly incorporated both directly in the Dendro platform as sources of descriptors. By defining a limited number of classes, avoiding the definition of many object properties, and living out constraints and axioms, these ontologies become viable alternatives to support the data model of a data management system, while being more easily manageable by curators.

Part **1** of Figure 10 shows the complexity incurred in representing a *temperature* metadata value using two heavyweight ontologies (EXPO and OBOE). Such complexity is undesirable in an operational system despite its semantic richness, so it is suitable to use a simplified representation via a lightweight ontology (part **2**). However, both approaches can co-exist: metadata can be represented using lightweight ontologies in one system and then evolve, via ontology relations, to more heavyweight representations if the need should arise.

Part of the proposed data curator workflow builds on past experience obtained from the implementation of a solution for collaborative RDM using Semantic MediaWiki [99]. This solution has been improved by employing ontologies and a triple store, dispensing with a relational database—an approach also followed by a previous architecture designed for extensible, domain-agnostic data management [65]

The domain-specific lightweight ontologies do not intend to comprehensively portray a scientific domain, but are focused in the data description

[2] The OWL version available at http://bloody-byte.net/rdf/dc_owl/, was implemented in Dendro

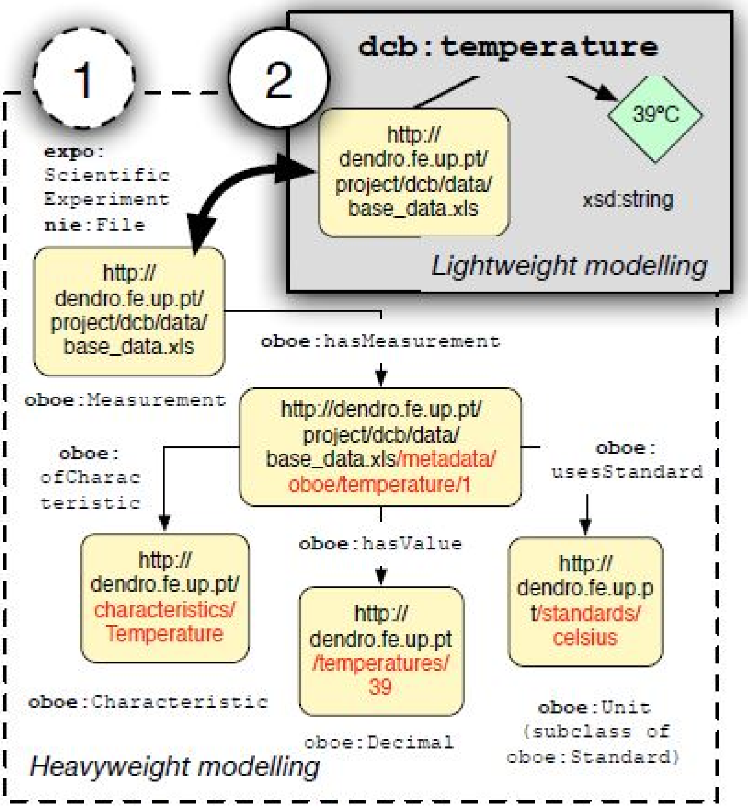

1
2
dcb:temperature
http://dendro.fe.up.pt/project/dcb/data/base_data.xls
39°C
xsd:string
Lightweight modelling
expo:ScientificExperiment
nie:File
http://dendro.fe.up.pt/project/dcb/data/base_data.xls
oboe:hasMeasurement
oboe:Measurement
http://dendro.fe.up.pt/project/dcb/data/base_data.xls/metadata/oboe/temperature/1
oboe:ofCharacteristic
oboe:usesStandard
oboe:hasValue
http://dendro.fe.up.pt/characteristics/Temperature
http://dendro.fe.up.pt/temperatures/39
http://dendro.fe.up.pt/standards/celsius
oboe:Characteristic
oboe:Decimal
oboe:Unit (subclass of oboe:Standard)
Heavyweight modelling

Figure 10: Recording dataset metadata using heavyweight and lightweight ontologies

needs of small researchers groups. It is also important to notice that these metadata models do not convey the notion of application profile, but rather a set of concepts that were identified together with domain researchers, and can be combined with other descriptors to obtain richer metadata records. All the concepts were given annotations specifying their *rdf:labels* and their *rdf:comments*, since a natural language description of the concepts is adequate to facilitate their interpretation by humans and, mainly, because Dendro use the annotation properties in the ontologies to build its resource description interface.

6.3.1 An extensible lightweight ontology

In order to model research concepts to match the *File-Folder* representation of datasets, a lightweight *Research* ontology was developed. The *Research* lightweight ontology consist in few classes that represent the structure of research types and comprehensive domain-agnostic properties such as the instrumentation, software, or method applied to capture the data. This ontology serves as "extension point" from which other domain-specific ontologies can be derived in order to represent the descriptors required for each domain. It contains concepts such as *Experiment* or *Paper*, that can be reasonably mapped as *Files* or *Folders*. The assumption here is that the directory structure closely follows the different activities of a research project—for example, the Paper concept to represent all the assets and activities in the process of creating a paper, and not the paper as *concrete entity* (unlike EXPO for example). When creating a lightweight ontology for experiments in a specific domain, the curator can also subclass *Experiment* to create specific types of experiments with their own properties, depending on the domain.

The concepts covered in the *Research* ontology range from the level of the research experiment to the level of the data file. This means that the semantics of file contents is not represented, nor the organizational and administrative concepts at the *research project* level (these can be found, for instance, in CERIF). The two ontologies, however, complement each other: CERIF models highest-level organizational concepts (project-level), *Research* is targeted at the individual experiments.

By the time a new lightweight domain-specific ontology is developed one can subclass *Experiment* with a specific type of experiment, such as Hydrogen Production, from which the data properties identified to describe datasets in this domain can be instantiated. The developed lightweight ontologies are then loaded into Dendro.

Figure 11 depicts the modulation of domain-specific concepts in the *Research* ontology, in this case concepts from the Vehicle Simulation domain.

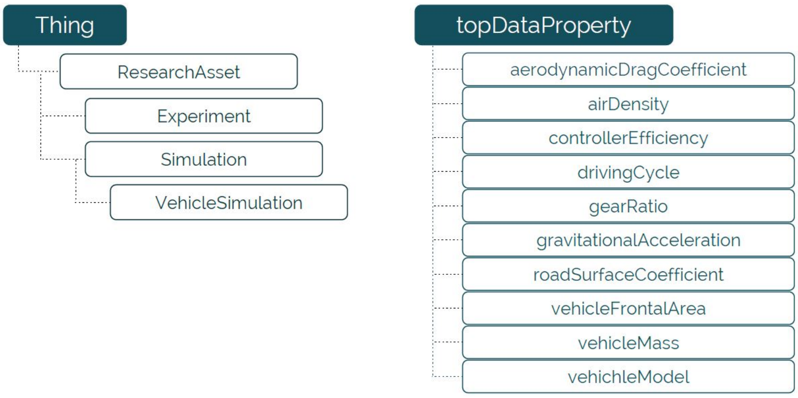


Figure 11: Integration of domain-specific descriptors in the *Research* ontology

The *Research* ontology is therefore focused on representing metadata for parts of a research project. This means that, for instance, a *temperature measurement* stored in a file will not be represented in the ontology. But the ontology may include a temperature property so researchers can represent the temperature at which an experiment was conducted. An instance of the metadata temperature property can be associated to the *File* or *Folder* of the experiment.

6.3.2 Instantiating the process

In order to demonstrate the applicability of this process, the process was first instantiated in two cases—one from Fracture Mechanics, and another concerning Pollutant Analysis.

- The Fracture Mechanics, double cantilever beam experiments (DCBExperiment), consist in testing a given material to study its resistance. A specimen is subjected to pressure so that researchers can evaluate the propagation of cracks in it, recording force values applied and the corresponding specimen displacement. These values are then processed with appropriate methods and converted into energy values.

- Pollutant Analysis is a type of experiment carried out by a analytical chemistry research group. This research group performs routine analysis regarding the concentration of certain pollutants in water and sediments collected *at a given time and place*, in what they call runs. These samples are taken and analysed using specific equipment and methods.

One of the main points assessed was the small amount of detailed information associated to each dataset given that these are easily interpreted by researchers from the same domain. Processed data on the other hand required more expertise regarding the production methods and context in which the experiment was carried out, so that the dataset can be interpreted and thereby cited. Produced data is saved in spreadsheets, where it is statistically processed and the final results are written in a word processor, which is a common workflow in research efforts.

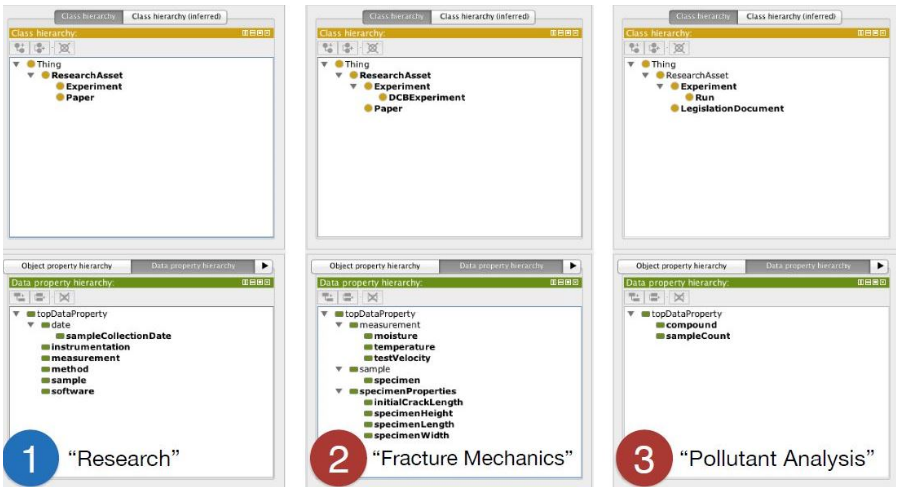


Figure 12: Instantiation of the lightweigh ontologies in the Fracture Mechanics and Pollutant Analysis domains

Figure 12 shows the lightweight ontologies instantiated in the two case studies. The generic *Research* ontology is shown in **1**, the Fracture Mechanics ontology is shown in **2** and the Pollutant Analysis ontology is shown in **3**. The *DCBExperiment* is derived from *Experiment* to provide faceted representation (i.e. distinguishing the DCB datasets from the remainder). DCB experiments metadata must include the ambient *Temperature* and *Moisture* percentage at the location of the experiment, and the velocity at which the specimen was pressed (*Test Velocity)*. It is also important to record the specimen that was tested and its properties (*Specimen Lenght*, *Specimen Width*, *Specimen Height* and its *Initial Crack Length*). These are subproperties of *Specimen Property* that can also be instantiated, allowing researchers to record other metadata.

Researchers from the Pollutant Analysis domain need to know the number of samples used (*Sample Count*), as well as the temporal and spatial information of the collected samples. To do so concepts from the DC ontology—namely *Spatial Coverage*—to identify the place where the samples were taken, and *Sample Collection Date* as a subproperty of DC *Date* can be used. Since the latter property is cross-domain, it was included in the *Research* ontology (**1**). Experiments are named *runs* by their creators, so *Run* was added as a subclass of *Experiment* and, as researchers compare *Compound* values with legal limits, *Legislation Document*, a *Research Asset* subclass was created to represent relevant legislation (3).

The advantage of this modelling process is the consistency and interoperability between lightweight ontologies, that allows Dendro to directly ingest and process them as sources of descriptors that researchers can use in the annotation of their assets.

6.4 USE CASES IN THE DATA CURATOR'S WORKFLOW

Recognizing that every research domain, or experiment configuration, has different data description requirements, a collaboration with a panel of researchers from several domains at the University of Porto was established. The goal was to provide researchers with the descriptors that enable them to obtain comprehensive and accurate metadata records. These domain-specific descriptors are expected to be simple and within the terminology used by the researchers, so they can be easily adopted in the production of well-documented versions of their datasets.

It is worth mentioning that the descriptors are only defined once in the domain-specific ontologies. This means that if two different domains require the same descriptor in their metadata models, this descriptor is only included in one of the ontologies, due to Dendro ability to draw descriptors from many ontologies at the same time and combine them in a single metadata record. This saves time, but the best benefit is fostering interoperability through concept reuse.

6.4.1 Vehicle simulation

At the time of the interview the Vehicle Simulation research group was conducting experiments to assess specific parameters related to the performance of electrical buses in an urban environment. This performance evaluation is highly dependent on datasets containing the bus routes, such as the geographic coordinates, latitude and longitude where the bus will go through. These data were provided by a bus company and each route had an associated file with the line schedules, allocated driver and distances covered. Other files contained technical vehicle properties provided by the manufacturer. To complement these data, researchers also needed specific environmental information, such as the air coefficient or the surface roughness, which can easily be retrieved from the web, according to the interviewed researcher.

When this information is gathered researchers are in condition of running a simulation as close as possible to reality. After each simulation new datasets are created, and those are liable to different interpretations and can be analysed, or reused, according to any specific research criteria, thus justifying the potential value that these datasets hold to launch new projects. For instance, traffic engineers can use the data to study congestion points; others can use them to optimize the bus routes. Access to these data, in the words of the researcher, can also reduce field work endeavours, since the alternative is to go to the street and manually count the traffic data.

At that time the research group working on electrical bus performance did not follow any particular data management guidelines. The datasets were mainly organized as spreadsheets, and when new external data arrived it was stored via Dropbox, and regular backups were made. For the purpose of searching for data, researchers basically trust their personal e-mail to keep track of all the entries. The research group did not describe their data, although the simulation variables could be part of a "ReadMe" file. The exploration of the data was harder than if the data were registered as metadata entries in a proper information system.

To calculate specific electric bus performance parameters the vehicle simulation researchers have developed a mathematical model, prone to change over time. This model includes several subsystems; one that computes the required energy for a vehicle to complete a driving cycle and another that uses the kinetic energy of the vehicle to calculate the possible amount of energy that can be recovered from the regenerative braking. Other subsystems are related to the batteries and supercapacitors and evaluate if these are capable of absorbing the energy from the braking. There are high-level entities that are essential to contextualize the electrical bus simulation setup, like the vehicle itself, and the driving cycle from which all the vehicle calculations are based [85]. Both the tractive force, that compels the vehicle forward, and the kinetic energy, have a great influence on the way the vehicle behaves, and the input values underlying the tractive force and kinetic force must be documented.

The vehicle simulation ontology uses properties related to the identified high-level entities. For instance a *Vehicle* property, corresponding to a vehicle category, like "electric bus" (or other type of vehicle depending on the study) and a *Vehicle Model* property that records a very specific property used in the simulation (eBus-12), where defined in the ontology. Likewise, since there are many available driving cycle standards to be used in vehicle simulations, the *Driving Cycle* descriptor was also defined. These are descriptors with the potential to create access points to the data, as they can yield information that distinguishes a dataset from others. All the other properties deal with a set of variables that constrain the entire simulation and are tied to the calculation of the tractive force and of the kinetic energy. Values concerning the *Aerodynamic Drag Coefficient*, and the *Road Surface Coefficient*, are contextual environmental variables that influence the performance of the vehicle under scope, and therefore must be annotated to help others interpret, or reproduce, the outputs from a vehicle simulation (Table 5).

Table 5: Vehicle Simulation ontology

Descriptor	Definition
Aerodynamic Drag Coefficient	A number used in calculating the aerodynamic drag of a vehicle
Air Density	The mass per unit of air in terms of weight per unit of volume
Controller Efficiency	The efficiency of the motor controller
Driving Cycle	A series of data points representing the speed of the vehicle versus time
Gear Ratio	The relationship between the number of turns made by a driving gear to complete a full turn of the driven gear
Gravitational Acceleration	The acceleration of an object cause by the force of gravitation
Road Surface Coefficient	Used in determining the influence of the road surface properties in rolling resistance
Tire Radius	The forward speed divided by the spin rate, for a free rolling wheel
Vehicle	The vehicle used in the simulation
Vehicle Frontal Area	The vehicle front end dimension
Vehicle Mass	The mass of the vehicle
Vehicle Model	A parameter used to designate the vehicle

However, since the mathematical model is expected to evolve, so does the vehicle simulation ontology if needed. At a given time researchers can be focused on evaluating the battery's performance and battery attributes can easily be added to the ontology.

6.4.2 Biological Oceanography

The Biological Oceanography domain in this case study includes researchers from three different groups at the Universidade Federal do Rio Grande, in south Brazil, namely: the Decapods Crustaceans Laboratory, a Laboratory of Ichthyology, and the Ecological Benthic Invertebrate Laboratory. A researcher from each group have collaborated with this work, thus the metadata model, and corresponding ontology for this domain, uses concepts that relates to each one of them.

The biological material collected consists of the organisms studied by the researchers (fishes, crustaceans and benthos) and sediments (substrate deposited on the bottom of water bodies). Environmental parameters may be recorded during the collection events or be independent, collected according to previously stipulated intervals. They are called by the researchers as "abiotic data" and the main ones are: water temperature, salinity, transparency and depth of the water. The two field activities apply methods and use specific instruments and tools to gather all sorts of biological material and environmental variables.

At the time of the interview researchers were mainly using spreadsheets to manage their data, first in paper for field data, and then the spreadsheets for aggregated data. The electronic files were simple and contained, at the headings, the performed measurements, abbreviated temporal and spatial references, sometimes accompanied by captions easily interpreted by the laboratory staff. The data were often organized in the available computers, and eventually saved in the cloud, depending on each researcher. The interviewed researcher prioritised raw data for storage and preservation, since raw data can originate new studies, and processed data are already documented in publications.

Regarding data sharing, it was usually up to each researcher to decide whether data from their projects were disclosed or not, because there were no established guidelines or commitments for this purpose. As a result, these initiatives were not based on standardized procedures. Ultimately, research data was only shared if two or more institutions were involved in a single project. Nevertheless, the research team acknowledged that this collaborative scenario was gaining relevance and becoming more frequent as new projects began. This collaborative environment flourished mostly within the University, where data sharing occured through the exchange of digital files containing biotic and abiotic data between laboratories.

For both field activities sampling events, spatial and temporal data were recorded. These were key elements during the process of building a Biological Oceanography ontology. Spatial data elements refer to the name of the place of a given event, specific sampling points, coordinates, among other. The temporal information is represented via the date of the events, their periodicity and the season when they occur. After the field events the collected material is processed in the laboratory. The biological material captured consists in organisms and sediment, which are separated in a triage process. At this stage, researchers calculate the sediment elements, the individuals are separated by species and an inventory is made along with several measurements. In order to describe some of these processes the following properties were defined in the Biological Oceanography ontology: *Individual Count*; *Individual Per Species*; *Species Count*; *Observed Weight*. Other parameters also include the species *Scientific Name*, *Sex* and *Life Stage*.

To address methodology issues it is important for researchers to annotate a description of the sampling procedures used in the research project, and the final destination of a sample after the analyses are made also needs to be recorded.

The Biological Oceanography ontology combines elements from different metadata standards or ontologies, namely descriptors taken from the EML (eml. prefix), from the Darwin Core standard (dwc. prefix), and others from the Ocean Biogeographic Information System database repository metadata profile, which is an extension of Darwin Core (obis. prefix). All the remaining descriptors were included as suggested by the researchers from this case study, as illustrated in Table 6.

Table 6: Biological Oceanography ontology

Descriptor	Definition
eml. Begin Date	A single time stamp signifying the beginning of some time period, like a sampling event period
eml. CommonName	Specification of applicable common names, may be general descriptions of a group of organisms if appropriate
eml. End Date	A single time stamp signifying the end of some time period, like a sampling event period
eml. Geographic Description	A short text description of a dataset's geographic areal domain
dwc. Life Stage	The age class or life stage of the biological individual(s) at the time the sampling event
dwc. Individual Count	The number of individuals represented present at the time of the sampling event
Individual Per Species	The quantity of individuals caught per species in a sampling event
obis. Observed Weight	The total biomass found in a collection/record event
Species Count	The total number of species caught in a sampling event
Sample Destination	Describes the final destination of a sample after used in the research analysis
Sample Identification	An identifier created at collection time to identify the specimen collected
eml. Sampling Description	Allows for a text-based/human readable description of the sampling procedures used in the research project.
dwc. Sampling Effort	The amount of effort expended during an event.
Sampling Periodicity	This field expresses the time interval between sampling events.
dwc. Scientific Name	The full scientific name, with authorship and date information if known.
dwc. Sex	The sex of the biological individual(s) collected.
eml. Single Date Time	Is intended to describe a single date and time for an event.

6.4.3 Hydrogen Production

The research group at the CEFT (Transport Phenomena Research Center, Energy branch) is focused on studying phenomena related to large-scale hydrogen production via chemical hydrides. The main purpose of this group was to instantaneously produce hydrogen to feed diverse Proton Exchange Membrane (PEM) fuel cells that can be used in a variety of portable devices such as cell phones or MP3 players. The experiments in this area focus on five main objectives, as follows:

1. Reactor optimization (smaller and lighter, with an ideal geometry); 2. Performance improvements through systematic feeding of reagent solution; 3. Storage of hydrogen in a liquid-based state through diverse additions (polymers, ionic fluids and other solutions); 4. Reaction output recycling; 5. Development of a kinetic model for the whole reaction.

Experimental data from this group was mainly stored in spreadsheets along with information associated with the environment where data was produced - temperature, involved compounds, pressure and other relevant measurements. Several connected sensors are used to extract this information, which was gathered with a specific software - that then exports them to the spreadsheet. Their workflow was divided in three main stages: first they produced raw data, which was then subject to error and consistency checking. At the last stage, raw data was then processed and refined to extract results and obtain conclusions about the performance of the procedure and the quality of the outputs. From the preservation point of view though, researchers identified the outputs of the first stage as the most important data to be deposited and preserved.

Concerning the collaborative scenario, the research team often resorted to traditional communication tools such as e-mail to share documents and data among them. In other cases data was copied from the researcher's personal computer to external hard drives, individually managing each access request.

When presented with the advantages of having research data published and accessible to either the workgroup or the scientific community, the interviewed researcher stated that it would be convenient to be able to test the reproducibility of their data and to retrieve the data associated with a specific publication, together with the associated metadata. To make their data findable and in conditions to be reproduced, hydrogen production researchers need to ensure that their descriptions include the predefined settings of the experiments. These settings involve the *Additive*, *Catalyst*, *Reagent*, *Reactor Type*, and the type of *Hydrolysis* used to perform the analysis. This kind of information serves as a pointer to facilitate the retrieval of datasets in this context, as a researcher may be interested only in a dataset containing the results of a powder reused Nickel-Ruthenium based catalyst experiment [46]. Additionally, since the amount of water used in the experiment (*Hydration Factor*) and the number of times a catalyst was reused (*Number of Reutilization*) influence the hydrogen generation results, the corresponding values also need to be recorded. In the final stage of the experiment the researchers evaluate the results in conformance with the *Gravimetric Capacity* and the *Hydrogen Generation Rate*, which determine if the experiment performance was positive or negative. All the datasets have valid results but only a few have satisfactory battery performance, and these are candidates to be used in the development of the fuel cell type PEM. If these values are registered the researchers can later easily identify the datasets that contain positive or negative values, and all results can be important for further analysis. Table 7 shows the descriptors that were selected to describe hydrogen production datasets.

Table 7: Hydrogen Production ontology

Descriptor	Definition
Additive	Type of additive used in the experiment
Catalyst	Type of catalyst used in the experiment
Gravimetric Capacity	Gravimetric hydrogen storage capacity
Hydration Factor	Amount of water used in the experiment = 2+x
Hydrogen Generation Rate	Amount of hydrogen per minute
Hydrolysis	Type of hydrolysis reaction
Number of Reutilization	Number of catalyst reutilizations
Reactor Type	Type of reactor used in the experiment
Reagent	Type of reagent used in the experiment

6.4.4 Social Sciences

Researchers from the social sciences domain deal with data from diverse sources related to interventions at different levels. The cases considered here concern the direct interaction with a group of social scientists. The produced data was often extracted from interviews - structured or unstructured transcriptions from such interviews, personal observations or reports, photos, videos or other multimedia-based support.

Data were considered to be very sensitive, as it concerned personal private information that needed to be hidden before the disclosure of data. Researchers already had procedures to anonymize data such as encouraging interviewees to use fictitious names, maintaining the overall validity of their data. When first approached, researchers from this domain showed awareness of recent, emergent data management practices. Projects from this domain were occasionally referred for data preservation, and there were already some guidelines for this purpose.

However, the guidelines were not available at the beginning of projects, so some did not have actually put them into practice. The result was that their daily outputs were mainly deposited and managed individually by each author in their own storage solutions, lacking the other contextual description that would otherwise accompany such resources. Due to their diversity in terms of data sources, these researchers resorted to different tools to allow them to individually and occasionally share specific items. This often implied extra time consuming tasks as some of these artefacts were hard to find. Among these tools, a cloud platform and personal email were the main platforms in place to achieve such a collaborative environment.

From the interview with one of the lead researchers, having access to the data associated to a publication would be of a great interest if this data could be fully interpreted. This requires recording parameters such as the date of production, the characteristics of the interviewed population (if applicable), and the point of collection. Nevertheless, from a reuse point of view, this researcher did not saw a substantial benefit in having data available to the scientific community and would rather have it managed within a smaller community. This was due to the fact that each newly created project already involves gathering new data, as old data tends to lose meaning and importance in this area.

After the domain analysis and the interview with a lead researcher, it was possible to identify some of the data description requirements, specifically tailored to help researchers from this domain to better understand data produced by other collaborators in the same area. Some initiatives already have impact in this field, such as DDI. From this stage, we selected a set of DDI (ddi. prefix) elements that were mostly needed for the purpose of data description in this specific case, as data from this domain greatly benefits from clear identification of the interviewed population, spatial coverage, involved methodologies and time span for the interviews.

As the number of contacted researchers in the social sciences grew there was the need to incrementally extend the ontology with more descriptors adding more descriptors to this DDI subset. The selected descriptors are listed in Table 8.

Table 8: Social Sciences ontology

Descriptor	Definition
ddi.Data Collection Date	Provides a data range of dates for the described data collection event
ddi.Data Collection Methodology	Describes the methodology used for collecting data
ddi.Data Collection Software	Describes the software used for collecting data
ddi.Data Source	Describes the source of provenance of the data
ddi.External Aid	Any support given to the interviewee such as text cards, images or audiovisual aid
ddi.Kind of Data	Briefly describes the kind of data documented in the logical product(s) of a group unit such as qualitative, quantitative or mixed
ddi.Methodology	Metadata regarding the methodologies used concerning data collection, determining the timing and repetition patterns for data collection, and sampling procedures
ddi.Sample Size	Size of the sample from which data was requested
ddi.Sampling Procedure	Describes the type of sample, sampling design and provide details on drawing the sample
ddi.Universe	Describe a universe which may also be known as a population
ddi.Other	ddi.Analysis Unit; ddi.Based On; ddi.Deviation From Sample Design; ddi.Independent Dimension; ddi.Summary Statistic Type; ddi.Variable

6.5 LESSONS LEARNED

The proposed data curator's workflow is based on the premiss that researchers are accountable for the description of the data they produce, as long as expressive descriptors are provided to them. Thus, researchers are not only stakeholders in data description, but also in the definition of the concepts to include in the domain-specific metadata models. The design of the lightweight ontologies itself is a rather straightforward process, since these are not exhaustive and complex from a modelling perspective, while their implementation as a part of the data model of Dendro is expedited.

The main challenges of this work were related to the interaction between the data curator and the researchers, mostly due to the lack of published material in this context, which certainly would add value to the dynamics of the first interactions with the researchers.

The data curator's also dealt with the researchers' neglect regarding RDM, specifically when data description is not, yet, a common practice for the majority of the research groups. In the conversations with the researchers, it was observed a general belief that their current attitude towards data management was already good enough, despite the difficulties in sharing research data between group members, or finding a dataset of interest.

Furthermore, this collaboration started upon request from the data curator side, which could have limited researchers' willingness to actively participate. Therefore, it is important to clearly show researchers the benefits of having their data described in order to motivate them. Otherwise, the many deterrents will prevent them from performing data description. However, the research groups in these case studies were motivated to be engaged in this work.

Another aspect to consider is the researchers' availability to participate in the meetings. Although the same approach was adopted in all of the case studies, the amount of time to complete the process was very disparate between them. For instance, the meetings with the researchers from the Hydrogen Production group were conducted over a period of time of no more than two weeks and the meetings occurred in their work place. On the other hand, the meetings with the Vehicle Simulation researcher were scattered over a period of three months and, mostly, in our work place. So, this kind of engagement is time expensive and needs to be carefully planned.

The developed ontologies were loaded to Dendro as a part of a data description experiment to assess the use of the ontologies and of the Dendro recommendation system [98]. The results showed that researchers make good use of the descriptors that have been selected for their domains, while also using descriptors from other domain-specific ontologies. Moreover, by exploring the vocabularies that they helped to build in the first place, in a concrete data platform use case scenario, researchers have developed their awareness towards RDM, and introduce their own recommendations to improve the proposed workflow.

It was also verified, in a very informal fashion, that one researcher was not interested in the concepts defined in collaboration with the Simulation Vehicle researcher. The difference was that the first was working in Traffic Simulation, and the datasets they create may use a different set of concepts, making it hard to define a metadata model comprehensive enough to describe research data in similar domains. This researcher stated that the Dendro platform was not useful for him since it had no descriptors of interest. This feedback alone strengths the idea that researchers will not adhere to data description unless familiar descriptors are provided to them.

Further work is necessary to estimate which of the defined concepts are really relevant to the researchers and which are not, and to consider extended versions of the ontologies. In this scenario there is space for researchers and data curators to partner and take advantage of their combined skills to develop domain vocabularies. The ontologies, derived from the metadata model design process, are not intended to fully represent the domains from which they derived; instead they capture the particular data description needs of the panel of researchers in the use cases. If the researchers need to provide additional contextual information, the corresponding properties can easily be added to the ontologies and incorporated by the Dendro platform.

7 ROLE OF CONTENT ANALYSIS IN IMPROVING THE CURATION OF EXPERIMENTAL DATA

Researchers are increasingly seeking tools and specialized support to perform RDM activities, with metadata atop, and the collaboration with data curators can be fruitful. Yet, establishing a timely collaboration between researchers and data curators, grounded in sound communication, is often demanding. In this chapter I propose manual content analysis as an approach to streamline the data curator workflow. I argue that with content analysis curators can obtain domain-specific concepts used to describe experimental configurations in scientific publications, to make it easier for researchers to understand the notion of metadata and for the development of metadata tools.

However simple it may seem, as I contacted with more researchers, I realized that even basic concepts as metadata were hard to convey and that building a communication channel with the researchers is a key factor. Some of the sessions with researchers, lasting at least one hour, have required a considerable amount of mental effort, without any satisfactory results. These unproductive interactions may be a deterrent for a researcher with no previous engagement in RDM or that has to be convinced of its benefits.

Thus, I increasingly started to explore content analysis [110], in an ad-hoc fashion, with the goal of improving the workflow by 1) preparing data curators to talk with the researchers using domain terminology to ease communication; 2) making the data curator proactive in the definition of domain-specific metadata models. Content analysis has great potential to improve this data curator´s worklow, and metadata production in particular, but the process needs to be formalised in order to become systematic.

Therefore, I proposed this task to be performed by a student of the master in Information Science, who took on the role of data curator. The data that support this chapter is publicly available[1].

Three use cases from experimental domains are presented in this chapter, one related to Sustainable Chemistry, one to Photovoltaic Generation and another to Nanoparticle Synthesis. These domains are useful for this proof of concept since experimental data is often reproducible if the procedures and relevant variables are well documented [128], while the diversity of research configurations calls for tailor-made metadata models rather than cover-all standards.

The curator started by performing content analysis in research publications, proceeded to create a metadata template based on the extracted con-

[1] Castro, J.A. and Landeira, C. (2019). Content Analysis of Publications in Experimental domains. INESC TEC research data repository. https://doi.org/10.25747/kh8b-xx50

cepts, and then interacted with researchers. The approach was validated by the researchers with a high rate of accepted concepts. Researchers also provided feedback on how to improve the proposed descriptors. Content analysis has the potential to be a practical, proactive task, that can be extended to multiple experimental domains and help to bridge the communication gap between data curators and researchers.

This chapter is based on the following publication:

- Castro, J. A., Landeira, C., da Silva, J. R. and Ribeiro, C. (2020). Role of Content Analysis in Improving the Curation of Experimental data. *International Journal of Digital Curation* [21]

7.1 CONTENT ANALYSIS RELATED WORK

Information about the *methods* to generate data is important contextual metadata, and it has been studied how different scientific metadata standards support the description of methods, but there is potential for more comprehensive elements. In this context, research papers were identified as a rich source of information [27]—however, to the best of my knowledge, content analysis is not a methodological technique usually associated to RDM, nor is it a common approach when developing tools for metadata production.

Nevertheless, an exploratory study based on the literature for soil science showed that journal publications hold relevant information for metadata production [28]. While this study focused on the actual data description task rather than the selection of descriptors, it shows the need to systematize and the possibilities to extend the approach to other disciplines. Reading the papers that report on the experiments where the data were collected was also suggested as a task the data curator must perform, to quickly get a grasp of the research domain of the data being described, even if becoming a domain expert is not the goal [126].

Since curators in institutional data services are expected to describe many datasets from a myriad of domains and in a very limited time, one must look towards automating the process as much as possible. Automated methods such as Named Entity Recognition, present in packages such as CoreNLP [47], or Keyword Extraction, implemented in the YAKE! framework [18] can be used to highlight the most important concepts referred in the research texts that usually are related to datasets. At the same time, they can help highlight relevant parts of a document so that the curator can more easily spot possible metadata to include in the dataset record.

With this in mind, content analysis in an automatic-fashion was evaluated in the TAIL context with the goal to discover those concepts that could be mapped to domain-specific descriptors, in this case properties from different ontologies, which were drawn from DBpedia or the LOV catalog [119] after a keyword extraction step using both CoreNLP and YAKE!. The results showed the complexity of the task, as even after keyword extraction there is a large set of possible ontology properties to choose from, and highlighted the indispensable role of the curator in the process for systematically validating and complementing the results of any automatic tool [76].

Information extraction from documents has been applied to RDM before: in the chemistry domain, for example, it has been used for the development of ontologies and predictable models from data. The result was considered useful to deal with significant amounts of data and structured documents,

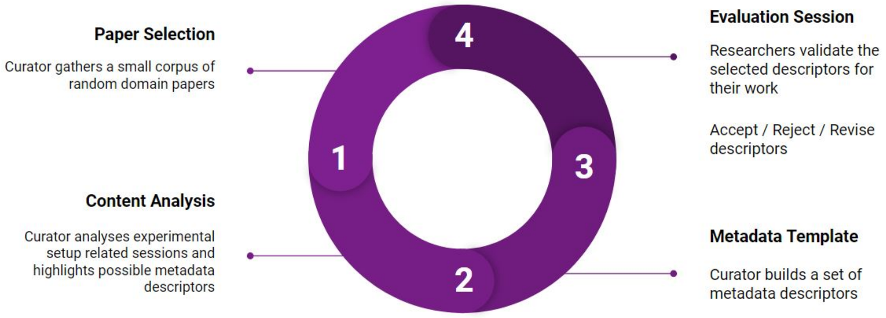

Paper Selection
Curator gathers a small corpus of random domain papers
Content Analysis
Curator analyses experimental setup related sessions and highlights possible metadata descriptors
Metadata Template
Curator builds a set of metadata descriptors
Evaluation Session
Researchers validate the selected descriptors for their work
Accept / Reject / Revise descriptors
1
2
3
4

Figure 13: From content analysis to the metadata template evaluation

but not effective when applied to less structured descriptions of chemical procedures [117]. Accordingly, chemistry librarians have argued that humans are able to efficiently summarize and to present information as opposed to the limitations that a fully automated approach might entail, such as many false positives in the selected concepts and overlooked details [75].

7.2 CONTENT ANALYSIS APPROACH

My approach to content analysis in the data curator's workflow (depicted in Figure 13) comprehends the identification of relevant segments of text in scientific publications, in a particular section reporting the experimental set-up, like the *methodological approach* or the *experimental configuration*. The experimental set-up section is particularly interesting as it systematically describes the parameters of a given experiment, that is, what were the procedures and provides the context for the production of data, which is a requirement for scientific metadata [88]. On the other hand, sections covering the results, although important to know more about the domain, are the output of an experiment with a greater focus on the data itself than on the context of production. Moreover, if the proposed approach entail the integral reading of the papers it would be a counterproductive task. However, a brief reading of the introductory section of each publication, or any other additional section, will give the curator a broader scope of the domain that may be useful to the overall task and to the conversations with the domain experts that follow.

The selected text segments are those where the researchers assigned a specific value to a property or made an environmental characterization. For instance, if the researcher writes *"The ozonation and the experiments with ozone-based AOPs were conducted in a bubble-column semi-batch reactor"* it can be infered that the *ozonation reactor* is a candidate metadata element. To ensure that the process is as realistic as possible from the data curator standpoint, without being dependent on their degree of specialization, the task was developed in the context of a master degree dissertation in Information Science, therefore performed by someone with limited RDM expertise, and a time frame of no more than two weeks, shared with other tasks.

Only a small corpus of publications for each domain was considered given that content analysis in a large sample might be more appropriate for an automatic approach or if the goal is to retrieve more values for the development of controlled vocabularies. Also, it is assumed that if some piece of information is relevant in a particular domain that kind of information will be present even in a small number of papers. On the other hand the assessment of a large number of papers would only contribute to increase the number of possible descriptors.

After processing the text and setting up a list of metadata candidates, the data curator prepared an informal metadata template in a shared document for the researchers to fill in.

This template is a simple two-column form, with the proposed descriptors in one column and an empty one in the other for the researcher to add the corresponding value. The data curator then asked researchers to insert values for the descriptors they considered appropriate or else to comment on how to make the descriptor more appropriate, in case they believed that the concept could be improved. If the researchers did not insert a value it was requested a further comment on the reason for this. The experimental domains to use in this experiment were determined by existing contacts with researchers producing data and on their agreement to participate in an evaluation session. Hence, the approach was applied in three experimental domains. The publication corpus for each domain was collected based on keywords related to the experimental configurations that the researchers in the three cases perform regularly. One researcher from each domain participate in the evaluation.

7.2.1 Sustainable chemistry, degradation of pollutant particles

Human activity waste accumulates in the environment and contaminates water, soil and atmosphere, triggering all sorts of hazards. Global warming, water shortage, health risks and malformation are some of the issues amplified by pollution. It is therefore important to eliminate pollutants or make them less offensive to the environment, and solutions to achieve these goals are being studied.

To reconstruct the context of an experiment related to the degradation of pollutant particles, it is necessary to capture the properties of multiple samples being studied, to record the instruments applied in the experimental workflow, their characteristics and calibrations, according to their influence in the final results. Metadata for the methods and techniques applied give a broad view of the experiments. On top of that, details of the duration of certain experimental events, measurements, and environmental controlled conditions, contribute to metadata quality and to the trustworthiness of data.

7.2.2 Photovoltaic Generation, thin film experiments

Solar energy is growing as a renewable and clean energy type. Photovoltaic energy generation is a method to convert solar light into electric energy. This transformation happens through semiconductors that can release energy when stimulated by light. Due to the demand for clean energy solutions, ensuring the provision of sustainable electrical power, many studies, as the study of thin films and the study of optical properties of copper, gallium, selenium and others, have been developed.

This kind of experiment involves different methods and techniques that influence the experimental configuration. For instance the effect of the so-called annealing temperature is only required if the researcher is adopting a technique that submits the sample to this factor. For the contextualization of photovoltaic thin film data one needs to know the technique of elaboration of the absorbing layer, the precursors used for the elaboration of the thin films, the optimized experimental conditions, the deposit substrate and its properties, the electrical and optical properties of the thin films produced. The researcher conducting thin film experiments was not available in person, so the evaluation task was done remotely. The data curator shared the link to the metadata template file for this domain, and used an online chat platform to provide instructions to the researcher and get feedback.

7.2.3 Nanoparticle synthesis

Nanoparticle synthesis refers to the methods for creating nanoparticles, and the development of this type of experiment are relevant since there is an wide range of applications in a diversity of areas. Research opportunities in this domain are rich given the capacity to revolutionize the characteristics and functionalities of materials on a nano scale. In construction it can have the objective to contribute to improve building conditions with materials that last longer, or with additional functionality. Nanoparticles can be applied, for instance, to enable the materials to autonomously remain clean. Moreover, this line of research has an impact in health applications, in diagnosis, transplants and tissue engineering, as nanostructures allow cell interactions previously prevented due to their size.

For this case a total of 74 potential descriptors were identified by the curator. However, after the two previous evaluations, the data curator choose to represent in the template only a smaller set of descriptors to see if the researcher, with a less exhausting task ahead, since the number of descriptors to evaluate was significantly smaller, would show a different behavior in the task. For example, more time considering options, not being influenced by a large number of concepts to have more room to suggest new ones and to verify if these were in line with those that were omitted. Therefore, the number of descriptors in the metadata template for this evaluation was only 23.

7.3 CONTENT ANALYSIS EVALUATION

For each case it was prepared a form with the proposed descriptors, where some are common to the three templates. Researchers were provided with instructions to fill in or to make a comment when the insertion of a value was not considered.

7.3.1 Sustainable chemistry evaluation

In the sustainable chemistry metadata template a total of 60 metadata fields were included, from which 53 were understood by the researcher. From these, 38 were directly filled in or approved, while the researcher has suggested improvements on the remaining. On the other hand, 8 descriptors were rejected, either perceived as redundant, such as the *Gas superficial velocity*, understood but not used or did not make sense for the researcher (see Table 9). The evaluation was concluded in a single session taking about one hour. At the end there was an informal conversation with the researcher to obtain additional feedback.

The researcher stated that there is a need to record the experiments in laboratory notebooks, that work as a minute of experimental configurations, specially when unexpected events occur. When asked if the metadata template was comprehensive enough to describe data on the degradation of pollutant particles, and capture the minute information, the answer was positive, but the researcher also noted that more elements are required. Nevertheless, the researcher pointed out that some fields should be more specific, e.g. *Solution* might be replaced by *Solution concentration*, from which the solution can be identified. Another suggestion was *Analysis method*, rather than *Polyphenol analysis method*, that was considered over-specialized.

The metadata template for this case included both a generic *Instrument* field, and additional ten descriptors for the identification of specific instruments, e.g. *Light radiation instrument*. In this case the researcher rejected *Instrument*, given the difficulty posed by the diversity and number of instruments used in a single experiment. Fill in the generic *Instrument* descriptor with information about all the specific types of instruments that were used in an experiment and their purpose, was perceived as a burdensome task by the researcher. Therefore, the researcher prefer to have available descriptors for each specific instrument and only introduce a value, for instance, its model. From the list of specific instrument descriptors proposed only the *Spectral measurements instrument* was rejected.

To make the data curator workflow more efficient the researcher suggested that the process might include metadata generated by the many instruments that comprise the experimental set up. However, it was recognized that it would be a challenge for the data curator to gain access to all instruments, as they are spread across different laboratories and require authorization by the lab coordinators. This is something hard even for the researcher, as noted.

7.3.2 Photovoltaic generation evaluation

Since the evaluation on the photovoltaic generation case was done remotely, the metadata template was completed in more than one interaction, over a few weeks, according to the agenda of the researcher. Although this process was constrained by the lack of personal feedback, it also offered the possibility to obtain results with less assistance from the data curator.

In this case the data curator listed a total of 56 fields from which 45 were understood, with the researcher proposing improvements for two of them. The 11 remaining were rejected with the justification that the metadata element was useless, such as the *Electrical resistance* and the *Spraying time* descriptors, as listed in Table 10. The researcher confirmed the proximity between the proposed concepts and the information they usually record in a notebook or text file, but alluded to the lack of a metadata element for the precursors concentration, which was not included in the metadata template. Thus, more metadata fields are required to capture all the relevant information in this case.

Similarly to the sustainable chemistry case, the photovoltaic energy researcher made considerations about the granularity of concepts. For the *Optical transmittance* descriptor the researcher suggested the adoption of the broader term *Transmittance*, while for some other cases more specific descriptors were suggested, such as subdividing *Dielectric constant* into *Real part of dielectric constant* and *Imaginary part of dielectric constant*. The description of instruments is also relevant, however only the *Instrument* field was available in the template, given the lack of expertise of the data curator to assign different types of instruments in this context. Nevertheless, this field was completed without further comment by the researcher.

7.3.3 Nanoparticle synthesis evaluation

As for the nanoparticle synthesis evaluation, from the 23 descriptors in the template a total of 20 descriptors were accepted, 16 directly filled in. The researcher recommended improvements on three others, while one was considered not very precise. The remaining three were rejected, considered meaningless or not needed for the researcher. The metadata template also included three descriptors for the representation of very specific types of instruments, the *Optical properties analysis instrument*, the *Sample synthesis instrument* and the *Radiation emission instrument*. The researcher added a description in two of them without further comments, although it has shown preference for the specification of the instrument when asked a generic *Instrument* descriptor would be suitable. A field for the description of the instrument producer would also be useful in some experimental contexts, according to the researcher. For the descriptor *Sample Coat* the researcher stated that this is not the most suitable term, but recognized that many colleagues can use it regularly, not knowing what the alternative term might be. The proposed *Reducing agent* descriptor was one of the concepts that was understood as useful but that was not used in the experiments performed by the researcher in this evaluation.

For the descriptor *"Laser pulse width"* the introduced value was *"(248 nm (wave-length); 500 mJ (pulse energy); 10Hz (pulse frequency); 20 nS (pulse duration)"*. This suggests that the researcher would prefer a structured description instead of having to fit structured information in an unstructured descriptor. It is also important to highlight that the list of hidden descriptors from the metadata template already included most of the necessary fields for the needed representation, namely *Pulse wave-length*, *Pulse frequency* and *Pulse duration*, while the *Laser energy per pulse* was a potential descriptor identified by the curator. However, if the *Radiation emission instrument* is available there is no need to record the wave-length, according to the researcher. When asked if there were missing descriptors, the researcher said that those presented were enough and that the metadata would be useful for other researchers as well.

After the evaluation of the 23 proposed descriptors, taking into account the duration of the session (about 30 minutes) and the availability demonstrated by the researcher, the data curator suggested a quick observation of the remaining descriptors. From the list of 51 remaining descriptors, one was seen as ambiguous. At a certain point of the evaluation the researcher concluded about the importance of the descriptors that *"their use will depend on the experiment"*. It depends of the method, technique, sample and instruments chosen, for instance the *Synthesizing vessel dimension* is only necessary if a synthesizing vessel is used.

When asked if the session was useful and if it was easy to participate in the task, the answer was positive, yet the researcher consider that if the descriptors in the template were organized and not *"all mixed"* it would ease the description, acknowledging that a correct organization of concepts would be a difficult task for someone who is not an expert in the scientific domain. The nanoparticle researcher also mentioned that there is a need to annotate the experimental context and that *"can not work in chaos"*, and prior to this experiment already discipline herself to annotate all the contextual information in her experiments. These annotations are made using slides, so a presentation is always ready when necessary. Other methods, like keeping a notebook, were explored but the researcher could not organize the information so efficiently.

Table 12 depicts the overall results. A total of 139 descriptors were presented, from which 117 were accepted by the researchers, some with suggested revisions. The remaining 22 were rejected, most viewed as unnecessary by the photovoltaic generation researcher, while some were repeated concepts with overlapping semantics or were not understood. The overall acceptance ratio was 84 per cent.

Whenever a proposed descriptor is rejected, this does not mean that it is not suitable for the specific domain. It may well happen that another researcher in the same domain applies some techniques with properties that are unfamiliar to the researchers in our case. The same is true for potential descriptors not identified by the data curator, as different subareas may demand additional ones.

7.4 CONCLUSIONS

The assumption in this work was that content analysis positively impacts the data curator workflow by improving the communication with the researchers and making the data curator proactive in the definition of domain-specific metadata models. With the metadata template evaluation I obtained tangible indicators that support this hypothesis, and our past experience carrying out RDM tasks with researchers makes it possible to reflect on more elusive indicators.

The content-analysis step is likely to decrease the communication gap between researchers and data curators. This is due to the increased awareness and interest on the researcher side, and also to the domain expertise gained by the data curator. During my several interactions with researchers I systematically requested feedback on how to make the interaction better, and many considered the adoption of domain terminology as something that helps them to quickly understand RDM benefits and practices. Performing manual content analysis provides domain knowledge to the data curator even before the first interaction and throughout the process, facilitating communication with the researcher. Hence, in the interaction with the researchers, the data curators can adopt domain concepts to illustrate RDM scenarios, as opposed to more generic metaphors, e.g. based on DC metadata. Disciplinary examples is something that researchers tend to ask for.

By showing researchers a template with familiar concepts the data curator denotes interest in their domains, establishing a productive, empathetic relationship that makes talking about metadata less demanding. Moreover, starting an interaction with good communication also leads to faster input from the researchers and raises their awareness in a effective way. For instance, it has motivated the Sustainable Chemistry researcher to make a suggestion as soon as the metadata template evaluation was completed.

Additionally, the high acceptance rate of the descriptors in the metadata template evaluation provides evidence of the advantages of performing content analysis before the first meeting with the researchers. In this line, a data curator can assume with some confidence that many potential descriptors resulting from content analysis will be included in a domain-specific metadata model. Moreover, the identification of descriptors was recognized as a realistic activity from the data curator point of view, since it was performed in a reasonable time frame and did not required in-depth domain expertise.

In this evaluation the data curator adopted an exhaustive content analysis approach and still the task was seen as practical. I think that if the data curator has adopted a principle of minimal effort, only capturing high-level descriptors, that might still be enough to start the conversation with the researchers and let them contribute with finer metadata requirements. Likewise, the greater the number of descriptors specified by the curator the narrower the possibility will be of the researchers contributing descriptors of their own, hence making it harder for the curator to determine which are the most relevant descriptors for each domain.

Regardless of its merits there is still room to improve this approach. An example is the introduction of tools for entity extraction from the texts provided by researchers at the start of the process. While these automated approaches show great potential in helping the curator navigate larger collections of texts, the results of our past work with keyword extraction approaches for metadata production show that they cannot be seen as a replacement for the expertise and engagement that the curator brings to the process, but rather as a complement.

I believe content analysis is a complementary task in the development of metadata tools, such as domain-specific ontologies. However, even a very specific domain or a particular type of experiment can encompass several techniques, each with its own metadata requirements. The expectation is that as the number of descriptors grows researchers can then combine suitable descriptors for each dataset, depending on the experimental setup that originate them.

Automatic ontology-learning approaches were also considered under the TAIL project [76]. The results were promising, especially for recall, but precision needed to be improved. From a data curator perspective, manual and automatic approaches can go hand in hand. Although automatic content analysis can expedite the process and deal with larger corpora, making sense of the large number of automatically extracted concepts still requires decision making, considering the subjectiveness involved in giving context to the extracted concepts, in order to infer the descriptors.

To conclude I do not anticipate manual content analysis as an activity to be performed regularly by a data curator. In fact, many researchers already have well-detailed experimental protocols and scientific metadata standards are available, some in experimental domains [128]. Even so, this approach can be adopted as long as there is a need to define metadata requirements from the beginning or to specialize extant tools.

Table 9: Results from the sustainable chemistry template evaluation

Curator input	**Researcher comment / recommendation**
Chemical compound	It causes doubt; everything is a chemical compound, so it has no specificity. *Sample*
Mass transfer coefficient	Do not know what to describe. *Degraded compound amount*
Oxidation agent potential	The value would be the same for Oxidation potential. *Oxidation potential*
Interfacial area	Not all samples have an interfacial area. *Interfacial distance*
Chemical demand (oxygen)	It does not make sense this way. The two are alike [ozone mass flow rate]. *(Studied) gas*
(Ozone) mass flow rate	*Gas phase flow*
(Pollutant) Ph	*Solution Ph*
Polyphenol	Do not understand. Ask for the formula, and for its quantity ask mass or volume. *Polyphenol molecular formula/mass*
Molecular mass	It is necessary to define the molecular mass of what? *((?) Molecular mass*
Impurity tenor	Very specific, by knowing the purity degree you calculate the impurity tenor. *Purity degree*
Aqueous solution	The aqueous repeats the solution idea. Is necessary to define the type of solution. *Cleaning solution / Acid solution*
Pollutant particle size	*Particle size*
Remaining Accepted Descriptors	Atmosphere conditions; Total carbon; Total organic carbon; Reagent; Oxidant agent; Chemical element; Control solution; Photocatalytic activity; Solar light intensity; Sample crystallite size; Sample pore volume; Sample reference; Sample centrifuged amount; Sample drying temperature; Adsorbent area; Adsorbent ash tenor; Adsorbent particle size; Adsorbent molecular formula; Particle removal technique); Surface area measurement technique; Electromagnetic radiation measurement Instrument; Ozonisation reactor instrument; Ph measurement instrument; Absorbance measurement instrument; Light radiation instrument; Light intensity measurement instrument; Surface area measurements instrument; Catalysts analysis instrument); Photocatalytic reaction vessel instrument; Ozonation time, Light intensity measurement time; Suspension stirring time
Inorganic carbon	Do not use the concept but I recognize the concept.
Gas superficial velocity	The same as gas phase flow.
Remaining Rejected Descriptors	Ozone partial pressure (gas phase); Ozone interfacial concentration; Catalyst wavelength; Spectral measurements instrument; Sample diluted centrifuged amount; Absorbance
Acceptance percentage	52/60 - 86 per cent

Table 10: Results from the photovoltaic generation template evaluation

Curator input	**Researcher comment / recommendation**
Optical transmittance	*Transmittance*
Dielectric constant	*Dielectric constant* real part / imaginary part*
Aborbent layer production technique	*Absorbent layer manufacturing technique*
Remaining Accepted Descriptors	Method; Chemical compound; Band gap; Deposition potential; Semiconductor type; Potential rage; Complexing agent; Reaction type; Bath configuration; Characterization technique; Deposition time; Gap energy; Refractive index; Extinction coefficient; Compound yield; Compound absorption coefficient; Compound physical state; Sample drying; Sample drying temperature; Sample drying time; Substrate type; Substrate dimension; Substrate cleaning method; Substrate temperature; Working electrode; Electrode reference; Electrode counter); Annealing time; Annealing temperature
Electrical resistance	I cannot find the exact resistance.
Spraying time	It is for a technique I do not use.
Remaining Rejected Descriptors	Compound viscosity; Compound boiling point; Reagent; Solution matrix; Photon energy; Cathodic sputtering source; Radio frequency; Sample power; Temperature stabilization time
Acceptance percentage	45/56 - 80 per cent

Table 11: Results from the nanoparticle synthesis template evaluation

Curator input	**Researcher comment / recommendation**
Sample coat	Not the most suitable term.
Sample concentration	*Sample mass*
Laser pulse width	*Pulse energy*
Sample heating time	*Deposition time*
Remaining Accepted Descriptors	Sample; Sample producer; Sample coat dimensions; Stabilizer; Particle size; Solution; Instrument; Optical properties analysis instrument; Sample synthesis instrument; Radiation emission instrument; Pulse duration time; Synthesis temperature; Synthesis method; Characterization technique; Atmosphere conditions; Substrate
Passivation molecule concentration	Does not make sense
Reducing agent	I understand but I do not use it in my experiments.
Remaining Rejected Descriptors	Milli-Q resistance
Acceptance percentage	20/23 - 86 per cent

Table 12: Overall results

Descriptor evaluation	**Sustainable Chemistry**	**Photovoltaic Generation**	**Nanoparticle Synthesis**
Directly accepted	38	42	18
Revised	(15) e.g. Interfacial area - Interfacial distance	(3) e.g. Optical transmittance - Transmittance	(3) e.g. Laser pulse width - Laser pulse
Suggested	Solution concentration	Dielectric constant (real part) (imaginary part)	
Not needed	Catalyst wavelength	(11) e.g. Radio frequency	Reducing agent
Repeated	(3) e.g. Chemical demand (oxygen) = (Ozone) mass flow rate		
Not understood	(3) e.g. Spectral measurement instrument		Passivation molecule concentration

8 TRAINING RESEARCHERS IN METADATA CREATION: A USE CASE WITH MIBBI IN THE BIOMEDICAL DOMAIN

Recent initiatives in RDM recognize that involving researchers is a challenge and that taking into account the practices of each domain can ease this process. In this chapter I describe an experiment in the adoption of data description by researchers in the biomedical domain. In this exploratory study the aim was to facilitate the adoption of metadata tools by researchers from the biomedical domain. This scenario enable to explore a complementary approach to the data curator workflow, since it dwells from the adoption of a mature, yet complex, metadata standard which was adopted to involve several researchers from the same scientific institution in data description.

This case study took place in I3S[1], a large institute for health sciences and technologies in Porto, bringing together researchers from different backgrounds in the biomedical sciences. The motivation was the fact that RDM was not yet a concern for most researchers at I3S and by the diversity of disciplinary requirements in this institute. This work was carried out by two students who were conducting their master thesis in the TAIL context.

A generic lightweight ontology based on the Minimum Information for Biological and Biomedical Investigations (MIBBI) standard was developed and presented to the researchers. In this context 7 interviews and 4 data description sessions using Dendro were performed. The feedback from researchers showed that this intentionally restricted ontology favours an easy entry point into RDM but does not prevent them from identifying the limitations of the model and pinpointing their specific domain requirements.

To complete the experiment, extra descriptors suggested by the researchers were collected and compared to the full MIBBI. Part of these new descriptors can be obtained from the standard, reinforcing the importance of common metadata models for broad domains such as biomedical research.

The approach was designed with a focus on training researchers by means of tools for the production of metadata, and involved the adoption of a simple and comprehensive ontology for the biomedical domain. This led to the consideration of existing domain metadata standards, namely MIBBI [111], that can take into account the metadata requirements of groups at I3S. A subset of the standard was selected to account for the recognized difficulty of researchers in adopting complex standards [88].

Figure 14 shows that this work has started with an analysis of metadata standards and their adoption in data repositories in the biomedical domain.

[1] https://www.i3s.up.pt/

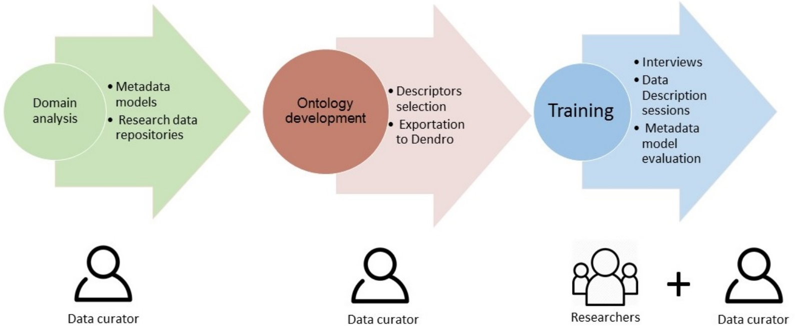


Figure 14: General Approach to train researchers in metadata

The next Section elaborates on the selection of a suitable set of descriptors for this case study. In Section 8.2 the creation of a lightweight ontology based on the MIBBI and its implementation in Dendro is described. In Section 8.3 the RDM perspectives of the I3S researchers and their feedback regarding metadata elements are outlined, while Section 8.4 deals with the data description session that took place in I3S.

This work is described in the following publication:

- Marcelo, S., Ferreira, A. L., Castro, J. A. and Cristina, R. (2019). Training Biomedical Researchers in Metadata with a MIBBI-Based Ontology *Metadata and Semantic Research* [102]

8.1 METADATA MODELS FOR BIOMEDICAL DATA

The research endeavour in biomedical sciences involves a multidisciplinary approach to the understanding of human health and diseases. This domain includes the study of human anatomy and physiology, cell biology, biochemistry, genetics and genomics, pharmacology and molecular biology [25]. With the current diversity of experimental and analytical techniques, the management of experimental data is not straightforward and to understand its context one must have access to a range of background information [111]. Hence, in recent years, recommendations regarding metadata for different kinds of experiments in this domain have appeared and several community-developed standards and recommendations are available [103].

The searching for an appropriate standard to train I3S researchers in metadata production, started with a list of standards in the life sciences featured in the Metadata Directory at DCC (Section 4.1.1), where the Genome Metadata, ISA-Tab and the MIBBI stood out as the more suitable ones.

The Genome Metadata consists of 61 metadata fields with a focus in Genetics and Genomics[2], while ISA-Tab is a general metadata tracking framework that facilitates standards-compliant collection, curation, visualisation, storage and sharing of datasets [49]. The ISA-Tab framework focuses on the description of the experimental metadata and builds on the Investigation, Study and Assay categories. The metadata in these categories is kept in

[2] https://docs.patricbrc.org/user_guides/organisms_taxon/genome_metadata.html

three tab-delimited files. An Investigation file maintains metadata on the context of the project and links to one or more study files. A Study file describes a unit of research, including the subjects of study and how they are obtained. Those subjects are then used in one or more Assay files, which in turn describe analytical measurements.

Checking the alternatives, MIBBI was considered more promising for the requirements of I3S and an accessible entry point for researchers to get into metadata creation. MIBBI consists of a set of guidelines for reporting data derived by current methods in the biology and biomedical domains [111]. Following MIBBI ensures that the data can be easily verified, analysed and clearly interpreted by the wider scientific community and promotes transparency in experimental reporting. There are 39 checklists in the MIBBI Portal[3] divided according to the experiment and its related biological science—e.g. the Minimum Information About a Microarray Experiment (MIAME) checklist is related to the use of (micro)arrays and analysis of the data they generate [17] and The Minimum Information About a Proteomics Experiment (MIAPE) checklist comprises modules for reporting the use and interpretation of data from various analytical techniques, such as mass spectrometry, gel electrophoresis or liquid chromatography [112]. Therefore, this metadata standard covers a wide range of disciplines, such as Genetics, Proteomics, Cell Biology and Bioengineering.

Table 13: Research data repositories in the biological sciences

Repository Name	**Metadata Standards**
European Nucleotide Archive (ENA)	Minimum Information about any (x) Sequence (MiXs)
Array Express	Minimum Information about an ENVironmental transcriptomic experiment (MIAME/Env), Minimal Information about a high throughput SEQuencing Experiment (MINSEQE) and Minimum Information About a Microarray Experiment (MIAME)
PRoteomics IDEntifications database (PRIDE)	Minimum Information about a Proteomics Experiment (MIAPE)
PubChem	Minimum Information about a RNAi Experiment (MIARE)
FlowRepository	Minimum Information about Flow Cytometry (MIFlowCyt)
European Genome-Phenome Archive (EGA)	Minimum Information about any (x) Sequence (MiXs)
Metabolights (MTBLS)	Core Information for Metabolomics Reporting (CIMR) and ISA-TAB
Sequence Read Archive (SRA)	Minimum Information about a MARKer gene Sequence (MIxS- MIMARKS) and Minimal Information about a high throughput SEQuencing Experiment (MINSEQE)

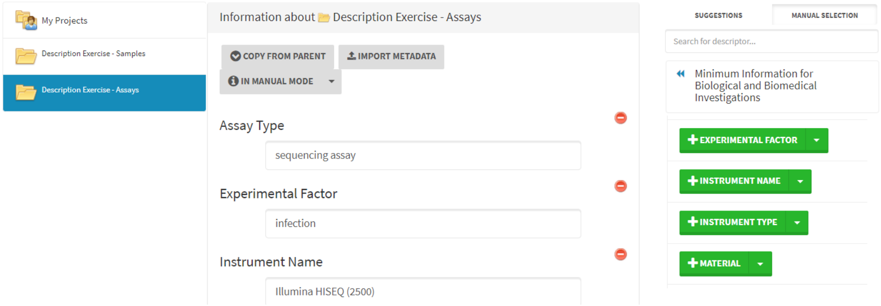
Table 13, based on the FAIRSharing community standards[4], shows that MIBBI checklists have been widely adopted by data repositories specialized in the biological sciences [103]. The number of FAIRSharing repositories is more extensive so the decision was to make a broad coverage of disciplines corresponding to the research groups in I3S.

8.2 MIBBIUP ONTOLOGY DEVELOPMENT

Given the usefulness of ontologies in resource description, a number of scientific communities have been working to establish domain-oriented ontologies [71]. For example, ontologies are widely used in biological and biomedical research where their success lies in the combination of four features: standard identifiers for classes and relations that represent the phenomena within a domain, vocabulary for the domain, metadata describing the intended meaning of classes and the machine-readable axioms and definitions [55]. An example of a widely used ontology is the Ontology for Biomedical Investigations (OBI) that provides terms with precisely defined

[3] https://fairsharing.org/collection/MIBBI

[4] https://fairsharing.org/communities



Information about Description Exercise - Assays

Minimum Information for Biological and Biomedical Investigations

Assay Type: sequencing assay

Experimental Factor: infection

Instrument Name: Illumina HISEQ (2500)

Figure 15: MIBBIUP ontology implemented in Dendro

meaning to describe all aspects of research in the biological and biomedical domains [12]. The OBO[5] and BioPortal repositories[6] provide an overview of ontologies in the biomedical sciences domain.

Considering the research diversity at I3S and the availability of domain standards, the first activity was a domain analysis selecting the descriptors most likely common to the disciplinary requirements at I3S.

The criterion was to have a set of descriptors aligned with the principles of simplicity and sufficiency identified as metadata goals for scientific data [128]. Moreover, there was a need to ensure that the proposed set of descriptors would allow any researcher, regardless of the type of experiment or data produced, to engage in data description.

Table 14: Selected descriptors from the MIBBI checklists

Category	Descriptors
Sample	Organism, Disease, Organism Part, Age, Sex, Ethnicity, Developmental Stage, Tissue, Cell Line, Cell Type, Sample Size, Molecule, Sample Type
Methods	Assay Type, Collection Date, Measurement, Method, Sample Collection Protocol, Treatment Protocol, Temperature, Study Design
Materials	Material, Drug Usage, Reagent
Technology	Instrument Name, Instrument Type, Software
Others	Experimental Factor, Environmental Factor, Study Domain

The analysis of biomedical metadata standards and their adoption by data repositories has led to the selection of a set of 30 descriptors and to include them as data properties in a lightweight ontology, MIBBIUP. The selection of the descriptors was based on the frequency of each one (related to samples, experiments or equipment) in the MIBBI checklists and their use to catalog datasets in the biomedical data repositories [92]. Table 14 provides the list of descriptors captured in the ontology and their MIBBI categories.

The ontology was operationalized in Dendro. Figure 15 depicts the data description interface in Dendro and part of the MIBBIUP ontology. Although MIBBI checklists also propose the use of metadata for the title, date and people affiliated with the research, these concepts were not included to avoid duplication of concepts in Dendro. Descriptors for this purpose are already present in more generic ontologies, namely DC.

[5] http://www.obofoundry.org/
[6] http://bioportal.bioontology.org/

8.3 THE PERSPECTIVES OF RESEARCHERS ON RDM

Between March and May 2019, seven researchers from four research groups were interviewed: three researchers from Genetic Diversity, two from Epithelial Interactions in Cancer, one from Glial Cell Biology and one from Differentiation & Cancer.

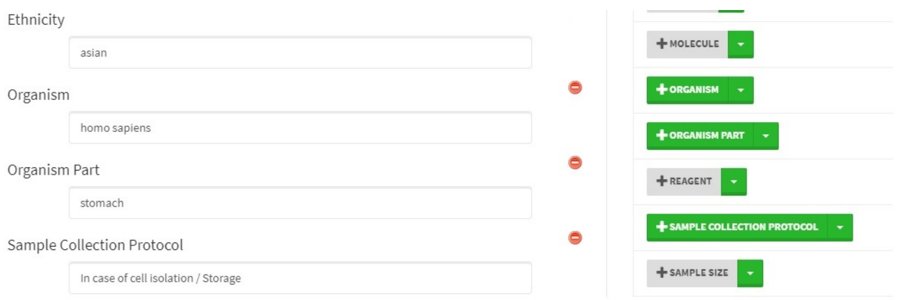
Sharing is mostly done with members of the group or external collaborators. Although the participants visit data repositories regularly and are interested in accessing data from other projects, they are still reluctant in doing it themselves. All the interviewees are familiar with data repositories and have positive experiences, occasionally resulting in data reuse. Although most researchers seem content with their present organization and storage of data, most believe their current methods can be improved. Moreover, they are aware that the risk of data loss increases as the number of files grows. The lack of established methods to manage data, even simple ones as systematic file naming, is regarded as an extra effort. From the collected feedback, it seems that there is room to improve the organization practices. Saving their time can be an encouraging factor to further the engagement of researchers in RDM.

When it comes to data description, the concept of metadata was not common knowledge and the researchers were not familiar with descriptors, except for two who are in charge of data deposit in repositories. This lack of knowledge about metadata was also verified in the user behavior and patterns of metadata usage in the *LabTrove* study [129]. Furthermore, data description is perceived as a burden and is not a priority for some. The concept of metadata was equated with database by one of them and with meta-analysis by another. However, after a short background explanation, the participants showed a better understanding of metadata benefits and were more open to considering them, even mentioning themselves some advantages of their use, such as facilitating the search for data. The researcher from the Differentiation and Cancer group is used to describing the sample and techniques adopted to produce their data and showed preference for a small number of descriptors. During the meeting, the researcher from the Glial Cell Biology group opened a disciplinary data repository to facilitate the discussion about metadata.

Researchers were also introduced to the concepts captured in the MIBBIUP ontology and most descriptors were unanimously accepted, except for *Drug Usage* and *Development Stage*, which got the agreement of only one of the researchers. Other descriptors were considered redundant: one researcher mentioned the *Environmental Factor* as a synonym for *Experimental Factor*, while many assumed that *Tissue* has the same use as *Organism Part*. Moreover, none of them would use a descriptor for the *Temperature* and *Reagent*, considering them part of the *Method* and *Material* descriptors.

8.4 DATA DESCRIPTION SESSIONS AT I3S

Data description sessions were carried out between April and May 2019, using the Dendro platform with 4 participants, two from the Genetic Diversity group and two from the Epithelial Interactions in Cancer group.



Ethnicity

asian

Organism

homo sapiens

Organism Part

stomach

Sample Collection Protocol

In case of cell isolation / Storage

Figure 16: Data description session in Dendro

After a brief demonstration of Dendro, participants were advised to use domain-specific descriptors from MIBBIUP, but they were also told that they could pick any other of the available descriptors. Moreover, DC descriptors were suggested to enrich the metadata, although these were not considered useful by the researchers. Figure 16 shows an example of some descriptors selected and filled by one of the researchers.

Participants from the Genetic Diversity group were PhD students involved in the study of human population and diseases. Researcher 1 was involved in a project focused on gastric cancer, while researcher 2 was analysing the exoma, microbiome and metabolome of african samples. Both researchers had no difficulty in exploring the Dendro interface or selecting the appropriate descriptors for their data. Both have selected metadata elements to contextualize biological samples used in their experiments. They also provided detailed information regarding the descriptors selected for the experiments, namely study design, materials and all the equipment used to generate the data. Overall, researcher 1 selected 21 descriptors, whereas researcher 2 selected 18 descriptors.

The Epithelial Interactions in Cancer group researchers were PhD students involved in different projects—researcher 3 studied the functional and molecular characterization of gastric cancer cells with stem-like cells, while researcher 4 was focused on the cellular and molecular mechanisms by which the Helicobacter pylori bacteria promotes the development of gastric cancer. Researcher 3 was particularly comfortable exploring the Dendro interface and selected a total of 16 descriptors from MIBBIUP. Researcher 4 found description more difficult and asked for help from the data curator. Nevertheless, this researcher filled in a total of 20 descriptors.

All 4 researchers stated that depending on the type of experiment and data they are likely to need a different set of descriptors, some of them not yet available on MIBBIUP. Moreover, researchers 2 and 4 would be interested in having the flexibility to create descriptors on the fly. However, they understood the importance of adopting normalized descriptors and agreed that the generic experimental metadata elements selected would be enough to help other researchers contextualize their data.

8.4.1 Overview of results

Results from the data description sessions showed that researchers understood the descriptors presented to them, although they suggested some new ones. The number of descriptors used ranged between 16 and 21 with an average of 18. The average duration of the experiments was 24 minutes.

Table 15 shows the 9 descriptors that were used in all the data description experiments. They are mainly generic ones about the technology used to generate data (instrument names and software), methods and materials used in the assays as well as information about the samples of the studies and the diseases they were targeting.

Given their studies about human population, Genetic Diversity researchers added more descriptors to the sample information such as the *Age*, *Ethnicity*, *Sex* and *Developmental Stage* of the subject. Other descriptors such as *Assay Type*, *Cell Line*, *Cell Type* and *Experimental Factor* were used by 3 researchers.

Table 15: Overview of the metadata created during the sessions

Descriptor	Researcher 1	Researcher 2	Researcher 3	Researcher 4
Disease	gastric carcinoma	hypertension	gastric cancer	cancer
Instrument Name	Illumina HISEQ (2500)	Illumina	Ion Torrent Sequencer (Thermofischer, City, Country)	Flow Cytometer
Material	Trueseq	whatmann paper	RPMI and Bovine Serum	collection tubes
Method	Protocol Reference	Protocol Reference	Stop infection Remove medium Wash 2x with RPMI medium Add new medium - 200 uL R10 Add RTK inhibitors - 2uL per each 96 well (dil 1:1000)	Staining for immunofluorescence
Organism	Homo sapiens	Homo sapiens	Homo Sapiens	Human
Organism Part	stomach	blood	stomach	gut
Software	GraphPad	Sequencher	GraphPad v8 (statistical analysis) — IDEAS software v3 (imaging analysis)	FlowJo
Study Domain	Disease susceptibility	Genetic Diversity	Oncology	Stem cells and cancer
Recommended descriptors	Replicate Count, Replicate Type, Country of Origin and Study Type	Sample Identifier, Instrument Manufacturer, Study Type and Protocol	Clinical Trial Description, Clinical Trial Phase, Clinical Trial Type, Collection Site	X

There were differences in the values for the descriptors used in common by the researchers. For instance, when referring to the *Instrument Name* used to produce data, researcher 1 and 2 named the same instrument, but researcher 1 also added its version. Also, researcher 3 pointed out it was important to record the instrument manufacturer, while researcher 4 misunderstood the *Instrument Name* and wrote its type. A similar situation was observed in the organism definition in which three researchers followed the NCBI taxonomy and wrote it in Latin, while one used English. Finally, the descriptor *Material* was interpreted by 3 researchers as auxiliary tools, while the other got it as mediums and chemical reagents used during an experiment.

After the data description sessions, researchers were also asked for more descriptors they might consider useful to describe and contextualize their data. Except for researcher 4, all suggested some new descriptors. Overall, the recommended descriptors are already implemented in the metadata checklists of the MIBBI. Researcher 1 from the Genetics group suggested two descriptors from the MIAME checklist [17] which provide additional information for the interpretation of a microarray experiment. Researcher 2 suggested a descriptor named *Protocol*, which value can either be a name or a reference to an external object, according to the MIBBI. The recommended descriptors by researcher 3 are not available in the MIBBI checklists. They seem to be motivated by the repository that the researcher uses to deposit their clinical trial studies.

During the data description sessions, it was observed that researchers may have preferences regarding the interface to enter the metadata. The unstructured metadata representation in Dendro was valued by researchers for making data organization easier during a project, yet with limitations if the goal is to deposit data in some disciplinary repository. Hence, researchers seem to prefer to record metadata in a tabular form. In this case, the ISA-TAB standard can be interesting to solve interoperability issues.

8.5 CONCLUSION

The motivation for this work was the need to train biomedical researchers in RDM, particularly by increasing their metadata skills. To this end the MIBBI standards was adapted as a top-down reference on the ontology-design approach. However, a bottom-up component with the involvement of researchers, with the potential to make the MIBBIUP grow according to their specific needs, was considered.

An iteration involving ontology design, test with researchers and check for additional descriptors was completed. The continuation with more case studies with researchers from this domain will provide new feedback to improve the MIBBIUP ontology. The data description sessions were productive, resulting in detailed metadata records in a short time period. This work laid the foundations for future work with groups of researchers from the same domain at University of Porto, although it is necessary to address issues such as the use of taxonomies and metadata quality.

Part IV

Multi-domain data description sessions

9 METHODS AND PROCEDURES

This chapter details the methods and procedures used in the multi-domain data description sessions. These include interviews with researchers, data description sessions and a follow-up questionnaire. I start by providing details about the interview script and coding categories, then I explain the set-up of the data description sessions in Dendro, together with the metadata categories that further enable the evaluation of the quality of the resulting metadata. To conclude this chapter I outline the follow-up questionnaire questions and the measuring scales used.

9.1 INTERVIEW

The approach to engage researchers in data management starts with a semi-structured diagnostic interview based on the Data Curation Profile Toolkit, Interview Sheet [19]. This interview sheet is designed to develop the data curation profile of specific projects. The sheet is structured in modules for specific stages in the data life-cycle, as well as about researchers practices and perspectives to guide the conversation.

Since the Data Curation Profile Toolkit is heavily structured, I edited the script in order to streamline the conversation with the researchers. The edition of the interview sheet takes into account the experience accumulated in interviews performed over time (see Section 8.2). During the definition of the final script version I took into account questions that work best and those that do not, as well as how to pose them. Another determinant was the selection of the most pertinent questions from the Data Curation Profile Toolkit to this work and the addition of new ones. I also checked the questions elaborated in the metadata related studies by Holly White [121] and Matthew Mayernik [72].

The script was originally written in Portuguese and includes 29 questions. The following is the structure of the interview script with some example questions:

Background Question

- I would like to know more about the research project associated with the data that will be addressed during the interview. What research project are you involved in?

Demographic information

- How often do you work with research data?

- Do you use data created by others in your research?
- Are the data you produced usually accompanied by metadata?

The dataset and its life-cycle

- What kind of research data are you working with (experimental, observational, quantitative, qualitative, etc.)?
- Is the data dependent on temporal, spatial context, sample definition, or any other type of conditions under which they are produced?
- Where does the research stands at this moment (collection, analysis, publication of results)?

Data sharing and organization

- Is the data shared, or is there an interest in sharing the data with people outside the research group? How? Why or Why not?
- Do you think that the data may have reuse value for researchers from different fields? In what sense?
- What activities do you carry out to organize the data produced during the research process?

Data annotation and publication

- Do you annotate the data as you produce them? How?
- Is the annotation of data an activity that you consider relevant? Why?
- Is there any information, complementary to the data, that you consider relevant so that others can interpret and use the data?
- Have you ever reused data produced by third parties? Was this experience positive (did you manage to reuse) or negative (did you not manage to reuse)? Can you tell me more about it?

The script was sent to the researchers beforehand so that they could have a better knowledge of the objectives and points to discuss. Some of them printed the script or had the document open on their computer during the interview.

Depending on the previous answers some questions were not asked if I thought that they would not add up to the conversation; also, new questions arose to follow up on some topic of interest raised by the researchers.

At the end of the interview researchers were asked to provide feedback about how the interview went. For instance if they felt that too many questions were asked, if there was any important subject that they wanted to mention and was not asked, if the duration of the interview was adequate, and most importantly, in which way the interview could be improved. The researchers signed a consent form so that the conversation could be audio recorded. The audio files were erased as soon as the interviews were transcribed.

The interviews were transcribed in Portuguese, and coded using the ATLAS.ti[1] software. Six main code categories were defined to highlight relevant information. The categories are as follows:

[1] https://atlasti.com/

- **Demographic Information**: Provides information to describe the study participants;
- **Awareness:** Statements that shows that the interviewee has awareness, or lack of, to a given RDM topic, either motivated by own personal experience or gained during the interview;
- **Share**: Statements that indicate interest in data sharing and issues related to sharing data;
- **Organization Practice**: Statements that describe tasks applied by the participants to organize their data; both for problem-solving activities and perceived issues;
- **Annotation Practice**: Statements that encompass activities to document data, from ad-hoc annotation to standardized metadata;
- **Reuse Perspective**: General statements concerning data reuse potential and data reuse experiences;
- **Other**: Codes created on the fly during the coding phase to capture important information.

Table 16 shows the relation between the defined categories and the code list.

Table 16: Transcription code list

Category	**Code**
Demographic	Professional title, Data usage frequency, Data repository usage, Metadata experience
Awareness	Acknowledge benefit, Raised awareness, (In) reuse; sharing; data description; organization; publication
Share	Perspective, Issue, No sharing, Practice
Organization	Organization Activity, Organization Problem
Annotation	Relevant, Not relevant, Benefit
Reuse	Data reuse, Positive experience, Negative experience
Other	Important

9.2 DATA DESCRIPTION SESSION SET-UP

The second moment of the researchers' engagement is the data description session. For each session I had to make sure that the participants had appropriate descriptors available for their corresponding domain. Otherwise, the same conditions would not be met for everyone, which would mean that a general analysis of the results would be undermined. Therefore, sometimes there was an extended period of time between the interview and the data description session, not just because of the availability of the researchers, but also due to the time dedicated to creating and importing new ontologies in Dendro.

To make sure that this condition was satisfied I took into account the interview, particularly the answers related to the dataset and its life-cycle to gather insight of domain specificities and the type of data usually created by the researchers. In case the ontologies already available in Dendro did not meet the metadata requirements, an existing ontology was updated with a set of new descriptors, or a new one was created from scratch for the session. The approach followed is in line with the processes developed to engage researchers in the data curator workflow, detailed in Chapters 6, 7, and 8.

All the researchers have accessed Dendro with my credentials. I created a project for each session in advance. First I wanted to spare researcher the additional step in the workflow, and most importantly all the sessions were kept under the same account for further analysis. All projects were kept private. This was explained to the researchers and they could change any information if they wanted to.

When scheduling the sessions I asked researchers to choose a dataset to describe, if possible a dataset mentioned during the interview, of an ongoing project or a recent publication.

I started the description sessions by introducing researchers to Dendro with a brief demonstration of its features. The researchers were then asked to create a folder and upload their datasets. After this step I explained in detail the choices that could be made in the vocabularies panel, by providing an overview of the available descriptors in Dendro, with emphasis on the most appropriate for the domain and the type of data of each session. Researchers were also introduced to DC as a complementary vocabulary to enrich the metadata.

During the session, the selection of descriptors was mostly up to the researchers - I interfered only upon request to explain the meaning of some descriptor, or when a researcher let me know they were looking for a specific type of descriptor. Moreover, when realizing that a researcher was stuck in the task I made suggestions on how to proceed. Exceptionally, I asked researchers if a given descriptor was suitable to contextualize their data.

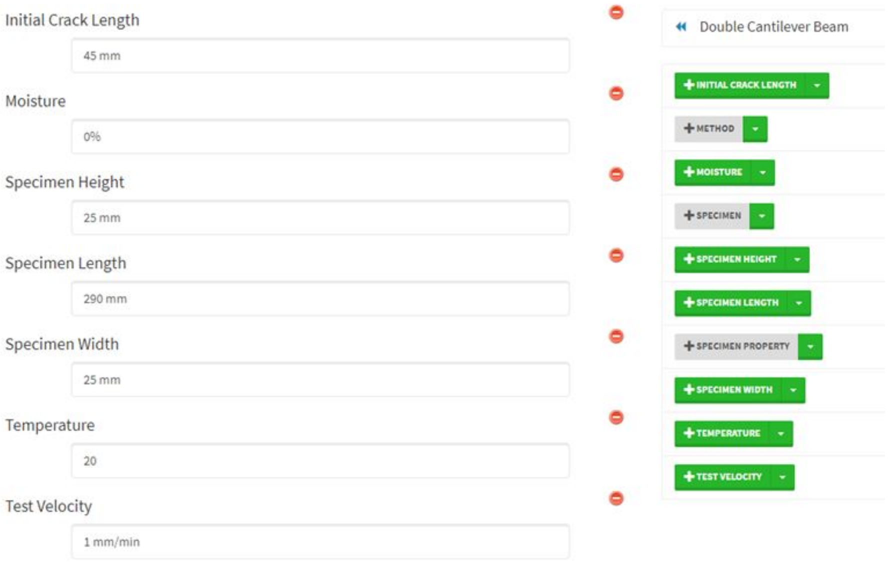
The researchers were given as much freedom as possible in their choices so that the experience was similar to a real-world scenario. With the exception of vocabulary recommendations, there was no pressure for researchers to opt for a particular descriptor, because this would compromise the subsequent analysis of results. When the researchers were finished with the description, I asked if they were sure they wanted to finish the session.

Sessions audio was recorded with consent and were deleted after the transcription of relevant events and comments during each session to complement the analysis of the metadata produced. The audio was also used to mark the moment the researchers started and finished the description, in order to ascertain the session duration.

Figure 17 represents a portion of the metadata created by one of the researchers and the selection of descriptors from a suitable vocabulary.

9.2.1 Data description results evaluation

For the assessment of the sessions results with regard to the quality of the metadata created by the researchers, I take as a reference the categories used by Jian Qin and Kai Li [89]. I adopted all of the categories proposed by these authors, except for the General metadata category, which did not fit with the nature of the metadata created in Dendro. Moreover, I defined the Experimental category as a new one, to represent the many environmental and sample properties, along with other aspects that do not necessarily fit into the Context category. Table 17 lists all the metadata categories and their definitions.



Initial Crack Length
45 mm

Moisture
0%

Specimen Height
25 mm

Specimen Length
290 mm

Specimen Width
25 mm

Temperature
20

Test Velocity
1 mm/min

Figure 17: Data description session

9.3 FOLLOW-UP QUESTIONNAIRE

A few weeks after the description session I asked the researchers to fill out a brief online questionnaire to get additional feedback, namely the perceived usefulness of data description for the research process and their assessment of the data description activity.

Moreover, I embedded an image with the metadata created during the session in the questionnaire form shared with the researchers, for them to judge whether the metadata was sufficient. The follow-up questionnaire ended with one question about their degree of interest regarding RDM activities and what they think are the most important factors for RDM engagement.

A first version of the follow-up questionnaire was tested with 8 colleagues, with adequate sensibility to interpret the questions, before the final version was sent to the participants in this study. It was verified that 5 minutes would be enough to complete it. The questionnaire was written in Portuguese and the following structure is translated to English.

9.3.1 Follow-up questionnaire structure

1. How do you evaluate the degree of data description usefulness for your research process?

The researchers´ attitude towards data description usefulness was measured with a 7-point Likert scale [66], from *irrelevant* to *important*. This question was accompanied by a mandatory question for researchers to explain *why?*

The second question regarding data description asked the researchers to finish the sentence:

2. Data description is a [adjective] activity

Table 17: Metadata categories

Category	Definition
Administrative	Meta-metadata, i.e. information about the metadata record, standard used, responsible party, rights for the metadata record,etc; Information about data archive/repository
Descriptive	General attributes about what the resource is and when it is possible, released, or made available; Related resources of the resource that is described
Context	Information about study/project design, model and population under study; data collection methods, instruments and constraints; analysis method used
Geospatial	Geographic names; Geospatial coordinates; Aerial maps and/or data
Identity	The name of an entity that is used to identify the entity understood by human users; A unique ID either in the form of some code or of a string following an identification system
Semantic	Subject terms describing the content of data; Subject or classification categories; Taxonomic classes
Temporal	Measurements of time; Temporal coverage of the content of data; Temporal criteria for data segmentation, processing
Technical	Parameters, models, measurements used in the dataset; Software-. system-, and format-relate attributes
Experimental	Experimental parameters such as environmental (temperature, solar light intensity), sample properties (specimen weight) and other (laser wavelength, substrate dimension).

To complete this sentence the researchers were given a semantic differential scale (SDS) with 5 pairs of adjectives. For each pair a 7-point linear scale was created. The SPS is a scaling tool, devised by Osgood, Suci and Tannenbaum [81], to measure social attitudes. By norm the scale is a 7-point bipolar rating scale using adjectival opposites, although it can also be used with 5 or 6-point scales, or more than 7 [3].

The 7-point scale enables a neutral choice (4) when compared to the 6-point scale, and a finer grade of judgement than a 5-point one. For participants responding (3) and (4) it can be said that they think that the activity is *a little* or *moderately* [adjective]; the (2) and (6) means that the participant is *quite* [adjective]; and the (1) and (7) means a very positive or negative attitude towards a given feature.

I defined a set of features to characterize the data description activity, and then set adjectives for the negative and positive poles of each feature.

The 5 features and the corresponding adjectives are:

Stimulation: *boring / interesting*
Motivation: *demotivating / motivating*
Difficulty: *hard / easy*
Duration: *time-consuming / fast*
Practicality: *impractical / practical*

3. Is the information [in the metadata record] sufficient to provide context for your data?

In order to enable the researchers to answer this question I attached an image to the question with the corresponding metadata record. To enable this, I did not share the questionnaire link with more than one researcher at a time. Only after a researcher had finished filling in the questionnaire I edited the image for the next. The options to this question were *Yes*, *No* and *Maybe*.

4. Do you think that more information is needed?
This is a follow-up to the previous question. The possible answer options were the same, and the researcher could say yes, even if they considered that the information was sufficient.

If the researcher thought that more information was needed, or maybe so, the type of information could be written through an optional free text.

5. According to the activities developed, what is your degree of interest in the management of research data? This question is not directly related to the assessment of the data description session, and therefore is not key to the scope of this study. Nevertheless, it was conceived as an opportunity to get more insight from the researchers perception to support general conclusions. A 7-point Likert scale, from *none* to *immense* was applied to measure the degree of interest in RDM. Through an optional free text answer the researcher could explain *why*?

6. I would have greater interest in the management of research data if... Here, the researchers had to complete the sentence with the selection of exactly three check boxes, from the following list:

- more tools are available for doing so;
- it gives greater visibility to my institution;
- it enhances communication and data sharing with close collaborators;
- it contributes to my scientific evaluation;
- my institution provides training;
- it provides contacts with other researchers and partnerships;
- appropriate data description templates are available for my project;
- I have to respond to funding agencies mandates;
- it gives more visibility to my work through data citation;
- it allows me to reuse my data in the medium and long term.
- other.

The options have been configured to appear randomly to each researcher, to avoid any bias effect on my part when sorting the check boxes. Like the previous question, this one was also outside of the main objectives of the study.

7. The activities I participated can be improved if... This is an optional free text question so researchers could provide additional feedback on how to improve the activities in which they participated. Their answers may show where the experience may have been unsatisfactory, and provide suggestions for improving further engagement with researchers.

10 PARTICIPANTS

A total of 13 participants completed this study. In this chapter I explain how the participants were recruited and then I provide a demographic description of the group. The participants are characterized by their professional title, frequency of data and repository usage, as well as their metadata experience. The research context of each participant is also briefly presented, as well as the type of data produced in each case. The domains are labeled as the researchers describe them during the interviews. Moreover, their sharing and reuse perspectives, organization and metadata practices are summed-up.

10.1 RECRUITMENT OF PARTICIPANTS

The recruitment of participants for this study was the result of the application of two sampling techniques. Most of the participants made up a sample of convenience and others were recruited via snowball sampling. The sampling is, therefore, non-probabilistic.

The nucleus of participants that constituted the convenience sampling was a group of people contacted spontaneously or through initiatives of the TAIL project (see Section 11.3.2). During the course of TAIL, contacts were made with researchers from faculties or research institutes in the University of Porto, to hold sessions to disseminate the project to researchers. Some of these contacts have been welcomed and I had the opportunity to explain to an audience of researchers the activities developed in the context of the data curator workflow and how I had been involving researchers up to that point. During these sessions I left my contact and an open invitation for anyone who wanted to participate in this study.

One of these sessions took place by the end of 2017, in the Faculty of Engineering with a group of researchers affiliated with the Faculty of Psychology and Educational Sciences. Further contacts were established with two researchers from the Family Psychology domain, and another from the Clinical Psychology. Another TAIL session took place at the Faculty of Arts and Humanities mostly with researchers from the Sociology Department, by the end of 2018. After this session I was contacted by the Consumption Sociology and Organizations Sociology researchers. The engagement of I3S researchers described in Chapter 8 was also motivated through this kind of session, but their cases were not included in this study.

The recruitment of the Nutrition and Cultural Studies researchers stem from contacts that did not lead to the organization of a general session, but still reached some researchers. The contact with these researchers occurred in December 2018.

The contacts with the Structural Adhesive Joints, Work Psychology, Services and Health have all different contexts but are also part of the convenience sample. Due to a small number of participants from the Engineering domain, I contacted the Head of Information Services at the Faculty of Engineering library, in February 2019, to ask for recommendations on possible researchers to engage. Three new contacts have been established. One of these contacts recommended a close collaborator working with Structural Adhesive Joints data.

Another participant from the Faculty of Engineering was a researcher from the Department of Industrial Engineering and Management, working in the Services domain. In this case the researcher was disseminating a survey via the institutional mailing list for their own research. Upon becoming aware of this ongoing research, I responded by requesting a short meeting in July 2018.

The Work Psychology researcher was an opportune contact. In November 2018, while I was waiting for the Clinical Psychology researcher to arrive to the data description session, I explained the nature of my work to this PhD student. During the conversation I asked if there was a willingness to participate in this study.

The Health researcher was an exceptional case, since this contact was started by a research group dedicated to active and healthy ageing, starting a Horizon 2020 project for the management and reuse of research data. A member of this group sent an email to the TAIL team, in March 2019, to find out about the existence of guidelines for data sharing in Portugal.

The remaining researchers were engaged via snowball sampling, where the participants provided contacts of other participants for the study. The liaison at the root of this recruitment pathway was a researcher that requested my support to solve a task. In this context I have solicited the contact of another researcher. This is where my first participant from the Faculty of Sciences comes from. The contact with the Sustainable Chemistry researcher was established near the end of 2017. After the interview with this researcher, I asked if there were researchers who could be suggested and, if so, if it was possible to probe their availability on my behalf. Thus, the communication with the Magnetic Materials researcher started in December, 2017. The latter has mediated contact with the Magnetic Dynamics researcher, which happened in July, 2018.

10.1.1 Sampling limitations and restrictions

The most obvious limitations associated with the used sampling techniques is sampling bias, that may hamper the possibility to generalize the results. A sample this size can hardly represent the universe of researchers at the University of Porto.

Within the scope of this study it can be fairly argued that it is difficult not to be biased in the selection of participants, since their availability to participate may imply that they were already motivated individuals for RDM. This is true, yet the focus of this study is the description of data, and even though a few researchers showed sensitivity to some kind of metadata production, none has ever done it using a tool like Dendro. In this sense, I could not identify preconceived ideas of the researchers that jeopardize the validation of results.

Another limitation of snowball sampling is that, by being in the contact range of the person making the recommendation, participants can share similar perspectives. In this case the results show that the researchers have actually shown different attitudes towards data description.

In order to minimise bias, I adopted a number of measures which restricted the enlargement of the number of participants. First and foremost, I made sure not to involve people within my range of personal contacts.

Therefore, I ensured that I had no prior knowledge of the people who made up the sample, their current perspectives and practices. Moreover, I excluded potential participants, who were easy to reach, who were in any way cooperating with TAIL activities, i.e., in the definition of a DMP, publishing data, or other collaborations built over time. This also excludes researchers that collaborated directly with me in the definition of the proposed data curator´s workflow. Hence, the number of researchers with whom I have had contact during this work far exceeds the number of participants in the study.

Another matter to consider was the control over the sampling size. Given that the data description setup implied the availability of descriptors for the domain of each participant, it was not feasible to extend the sample without meeting this condition. Another aspect considered is the representation of domains. As such, I deliberately ensured that the representation of domains was not disproportionate, thus some contacts were excluded if they were too close, and in some cases direct collaborators, with the people already recruited.

Overall, I opted in general for a casual posture in recruiting participants. Nevertheless, the number of participants here is not short in comparison with studies with a similar setup. This recruitment occurred in a context where no institutional service to facilitate access to researchers exists and where few researchers take the initiative to find out more about RDM. I did not provide any kind of incentives to encourage participation, although this is a common recruitment practice.

Before formalising the participation, I briefly met with most researchers to disclose the general objectives of the study and the different procedures, without mentioning the expected outcomes.

10.2 participants' area of work and demographics

Family Psychology Two researchers participated in this interview, the Principal Investigator and a post-doc researcher who works more frequently with data. The focus of their work is to understand the dynamics of people that have to respond to many demands because they are workers, parents and full-time loving companions. Data is collected from families with the objective to understand the impact of conciliation challenges, or the enrichment that derives from doing various things in the development of children. These researchers have created a longitudinal database with data collected through questionnaires. At the time of the interview they were still in the data collection phase, yet analysis of existing data and publications from this longitudinal study were happening simultaneously. The researchers explained that they have used some dimensions of interest at a given point, which does not imply that these dimensions will not be used again with a different purpose. There was manifested interest in knowing more about metadata creation since there was an opportunity to share data with external colleagues and they would like the data to be underpinned in an"instructional book".

Sustainable Chemistry The research project of the Sustainable Chemistry participant consisted on advanced oxidative processes for water treatment. The synthesis and characterization of catalysts enable to assess pollutants in aqueous contexts. This researcher was in the writing phase of the doctoral thesis. The produced data is experimental. The researcher started by taking reagents and producing semiconductor oxides. After characterizing the produced material, this is used in reactions for the oxidation of organic compounds. The several data collected allows to asses the reaction efficiency. The experimental conditions are taken from preconceived conditions, either local or from the literature. In the laboratory researchers use an act as a diary, for the recording of working conditions and if something unforeseen happened. This information is then taken into account when it comes to processing the data. The experimental content is important since factors like the temperature can interfere with the experience, so the researcher must be aware of factors that may alter the constant temperature.

Clinical Psychology The researcher from the Clinical Psychology is a Professor, working on general psychological well-being. More precisely this researcher works with psychometric, that has to do with the measurement of psychological variables. For instance, through questionnaires the symptoms of depression can be measured and then the researcher can tailor-made an instrument for each person. Data can be quantitative or qualitative. From the quantitative point of view it is necessary to transform the items into dimensions. Items have to be added or averaged to have the dimension, which is generally used in analysis. The qualitative data is the text. The string variables that were in the questionnaire are entered into a database. To measure reliability, the coding is blindly performed by the researchers. The current study results were published, but there were others studies at the beginning. The researcher does not create any kind of metadata except for keywords.

Magnetic Materials The researcher from the Magnetic Materials domain is a PhD student with a thesis subordinated to confinement effects. The tasks involve the study of a well known family of materials, chemical doping, the implementation in polymers to make composites and nanoparticles. The data is experimental and one of its constraints is the atmosphere. Since chemical elements have different oxidation levels, the atmosphere interferes with the measurement of the nanoparticles. Therefore, the researcher explained that with time the oxidation grows and cannot achieve the initial measurements, so the chemical elements have to be prepared as close as possible to the time of measurement. The data stays with the researcher thru their complete life-cycle. First the researcher manufactures the sample, then performs an X-ray, followed by the magnetic measurements. Depending on the composition and the type of material they end up doing more or less analysis. At the time of the interview the researcher was finishing some analysis in order to conclude the thesis work. When it comes to contextual information the researcher noted that the instrument already generates several experimental parameters.

Services The researcher in the Services domain, is a Professor that was conducting a collaborative survey with an Australian colleague. They were collecting data by the time of the interview. The aim of the survey was to find out whether people use fitness trackers, why they do, and what the advantages and benefits are from their use or potential use. The fitness trackers survey produced quantitative data, mostly with closed questions. For these data the researcher thinks that no spatial or temporal dependencies exist, due to the fact that to validate the results the study was not limited. However, the Services researcher has also experience working with qualitative data from interviews and in this case the data are dependent on the segmentation of certain groups of people. This researcher does not have the habit of creating any sort of metadata.

Consumption Sociology The Consumption Sociology is a post-doc researcher who was developing a project regarding collaborative consumption and mobility, with the aim to influence urban policies and strategies to reduce carbon emissions. In previous work the researcher already created a database over car sharing practices that was considered underused. During the interview the researcher told that was preparing a questionnaire to be applied in schools to identify the consumption behaviours of the students. The resulting data needs to take into account the students characterization, who are students of a specific school at a specific time, thus temporal and spatial information is important to provide the data context. The Consumption Sociology researcher acknowledged that they went to find out more about metadata before the interview, and that without knowing it, always have produced them. However, these metadata correspond to a code book to cross variables at different levels.

Organizations Sociology The project of the Organizations Sociology researcher is subordinated to unemployment and social inequalities. It was originated by an interdisciplinary working group to diagnose the dignity at work. At the moment of the interview the group was outlining a strategy to reach the organizations. The project had already worked with primary and secondary organizational data. Primary data was obtained with questionnaires, interviews, focus groups and direct observation. Secondary data are extant statistics on social entrepreneurship and on the business reality. For data interpretation the researcher thinks that it is essential to have some knowledge about the economic conjecture, so as to contextualize the data in the geographic and temporal space. The data is not independent of the policies for financing business activities. To support data analysis the project prepared an interpretative report of the data.

Nutrition This researcher was on a project commissioned by the World Health Organization, to characterize street food in cities in Central Asia and Eastern Europe. Its aim was to observe what kind of food was available in the vendors' stalls and take samples to check a number of parameters from a nutritional point of view. In addition, through the observation of trained locals who collaborate with the project, the team can describe what the people were buying. The project started with the development of research protocols. After data collection, analysis and publications, the last stage was the elaboration of final objects to disseminate the project results. At the end of the project the database is destroyed. For the data collected in the markets, seasonality and the religion context are important elements to interpret the data. As for the analysis the researcher does not know the specificities of the laboratory work, but believes that they have to obey to parameters such as temperature and humidity. The Nutrition researcher did not perceive the meaning of metadata and said that it is a concern they do not have.

Magnetic Dynamics The Magnetic Dynamics researcher was in the PhD final year by the time of the interview, but was still collecting data with a spectrometry technique to check the dynamics in magnetic materials. The objective was to verify how the material behaves at a nanoscale, how it interacts with light, for future technological applications, memories and sensors. This researcher produced experimental, quantitative data. For each statistically defined value, working at a very small scale, many statistical points have to be acquired. The conditions of the room itself affect the measurements, since variations in temperature and humidity have an impact on the performance of the laser. A protocol is set up before the experiment to ensure that the parameters are maintained through the experiments.

Cultural Studies The background of the interview with the Cultural Studies researcher was an ongoing PhD project. The motivation for this project was to analyse a specific literary corpus, through the lens of utopia, diet and gender, using a set of specific questions in order to deconstruct power dynamics. The researcher was still collecting data (corpus reading), indexing by questions and topics. The following phases would include data analysis, creation of infographics and thematic maps. The data is both quantitative and qualitative. They include citations, keywords, notes, number of occurrences and other features. In order to be interpreted the data is mostly dependent on their literary context. The researcher only captured bibliographic metadata, for the authors, documents, editors, among others.

Work Psychology During this interview the Work Psychology researcher, that was starting a PhD, talked about a finished recent project. This project consisted in measuring how personality types influence the job satisfaction in an industrial place. The researcher worked with demographic information, such as age and years of work, collected with questionnaires or information provided by the factory. Most of the data was quantitative. The researcher recognized being a little inexperienced when asked about which kind of information would be necessary to interpret the data, stating that had only used one type of software for working with data.

Structural Adhesive Joints The Structural Adhesive Joints researcher was leading a project for the automotive industry, but also conducting fracture mechanics studies, among other collaborations with companies. Therefore, the projects were in distinct stages. Most of the data generated in these projects is quantitative, chiefly data processed by the acquisition systems and fine sensors, e.g. force, displacement, temperatures, deformations, accelerations, among others. Complementary data include high speed video footage and photographs. The researcher also works with data from simulations. Each assay type have their own context, yet the research group has standardized internal rules and the software are configured to provide information about the test specimen conditions, operator name and test date. This information is accessible long after the assay takes place.

Health At the time of the interview the Health researcher was a member of a research group working in the validation and reliability of a self-assessment of frailty survey based on a mobile application. The survey was producing quantitative data about self-perception of health, nutrition, medication, psycho-social cognitive status, time management and more. According to the researchers it is important to have information about the sampling procedure to interpret the data. The researcher is not familiar with metadata and the research group does not have established procedures to describe the data produced.

Table 18 provides an overview of the participants demographics, with respect to their professional title, frequency of data and repository use, and metadata experience. In the Family Psychology case two researchers were present at the interview.

Table 18: Participants demographics

Domain	Professional title	Data use frequency	Repository use	Metadata experience
Family Psychology	(1) Professor; (2) Post-doc	(1) low (2) regular	(1) never (2) rarely	(1) low (2) none
Clinical Psychology	Professor	low	rarely	low
Sustainable Chemistry	Student	regular	never	average
Magnetic Materials	Student	frequent	never	average
Services	Professor	frequent	never	none
Consumption Sociology	Professor	regular	rarely	low
Organizations Sociology	Professor	regular	rarely	average
Nutrition	Professor	occasional	never	none
Magnetic Dynamics	Student	frequent	never	average
Cultural Studies	Student	regular	rarely	average
Work Psychology	Student	regular	never	none
Structural Adhesive Joints	Post-doc	frequent	never	high
Health	Post-doc	frequent	rarely	none

The domains represented in this study can be divided into two groups. One group is associated with the natural sciences and is represented by 5 participants: Sustainable Chemistry; Magnetic Materials, Magnetic Dynamics, Structural Adhesive Joints and Health. The remaining participants can be classified as belonging to the range of social sciences. This classification is coarse, since the domain boundaries are sometimes fuzzy, while some projects are inter or multidisciplinary. The most represented science is Psychology with three participants, followed by Sociology and Physics with two participants each. Regardless of the proximity in the respective scientific fields, the research projects have all different realities. Thus, I consider the sampling of participants both balanced and diverse.

In terms of professional, or academic experience, the sample is also well-balanced. The participants can be divided into two main groups; one with more substantial experience, made up by Professors, and the other composed of doctoral students. The group of Professors gathers 6 participants, while that of the students includes 5. The remaining three participants (including the second participant in the Family Psychology interview), are Post-docs involved in research projects. With the exception of the Work Psychology researcher, all doctoral students were at an advanced stage of their projects.

Most researchers work regularly, or frequently, with data. Naturally, the PhD students and the Post-docs with current projects are the ones with a more intense contact with the data, either in production or analysis. The Services researcher, who is a Professor, also worked with data frequently, and was active in collecting data for the fitness trackers project. The Work Psychology and the Cultural Studies researchers are doctoral students working with data on a regular basis. Professors who work regularly with data are the Consumption Sociology and Organization Sociology researchers, mainly for the analysis of data collected over time. The Nutrition researchers works with data occasionally, especially in phases of light workload. The Family Psychology and the Clinical Psychology are project coordinators, so their contact with data is low, and it is centered in data analysis and in the writing phase of publications. The second researcher from the Family Psychology domain has a regular contact with data and is actively involved in data collection. For this reason, this researcher was the participant engaged in the data description session.

The participants do not have experience in using, or assessing, data repositories. Many have never used or are unaware of their existence, whilst only 5 researchers said that they access data repositories rarely. These researchers have sporadically used statistical database services, such as EUROSTAT, or the National Statistical Office.

The Nutrition, Work Psychology, Services and Health researchers do not create any kind of metadata and demonstrated that they were not familiar with the concept. These researchers asked for a definition of metadata. For the Nutrition researcher metadata creation was not a concern. The researchers from the Family Psychology domain have no experience in creating metadata, but seemed to interpret the concept well, mentioning that it was something that they were interested in learning.

The Clinical Psychology and Consumption Sociology have low practice in the production of metadata. The first said that usually only records the keywords, the second searched for the meaning of metadata before the interview, tried to give some examples of metadata but the distinction between the metadata and the data itself was not clear.

Other researchers have regular contact with metadata of some kind, although the Sustainable Chemistry, the Magnetic Materials and Magnetic Dynamics researchers inquired about what was meant by metadata. These researchers use at least one procedure for the documentation of the experimental conditions, either by means of a minute, or the information is generated by the instruments themselves. The Organizations Sociology and Cultural Studies were already acquainted with the metadata concept. The Organizations Sociology researcher creates documentation to explain the data, the other regularly makes bibliographic records.

Only the Structural Adhesive Joints seemed to be fully aware of the concept and showed great deal of experience with metadata. As already mentioned, this researcher works with instruments configured by the research group itself to generate metadata. According to him it is information that can be consulted one or two years later to understand how the raw data were processed.

10.3 SHARING PERSPECTIVES

As a whole the participants did not exhibit a culture of data sharing. Only one researcher explicitly stated that they already shared data with third parties. Most of the participants shared with close collaborators, and some have never shared data to support the work of others.

Two researchers mentioned having no experience in data sharing. One only has sent data to a fellow researcher to get some help with data analysis. This researcher said it would be interesting to share data but explained that sometimes is not even worth it because others will not have access to the sequence used in the measurements and, therefore, data cannot be replicated. Another prefers to work on the data alone at the beginning of a project and would only consider showing the data to others to solve specific issues.

Another researcher thinks that data sharing outside their immediate collaborators is very unlikely to happen since the policies for sharing sensitive data in the country of collection are very strict and complex. This researcher has already faced some issues working with data shared by team members, by email. Sensitive data implies replacing participants identification by codes and it is a challenge to cross variables when combining two databases from different sources. Some data have already been discarded and in other cases the researcher had to spent a considerable amount of time to make sense of the data.

Data sensitivity was also mentioned by another researcher, who agreed that people have to obey demanding rules in order to share personal data. This researcher only considered the possibility of sharing data after reading the interview script I send beforehand, and showed openness to share data only if fully aware of the intentions of others, to enforce correct use of the data and the provision of credit. Moreover, this researcher is worried about ethical issues and the idea of someone using the data for financial returns was mentioned as a concern. However, it also recognized a sense of ownership over data collected with their effort. This feeling was also mentioned by another researcher, who argue that the research challenge is to acquire the data. The latter would consider data sharing if someone requested the data after the publication of results and the data had been fully explored.

Some researchers are more at ease with data sharing. In two cases the researchers were getting prepared to share data in international networks, with people with common interests who will have access to databases created in different projects. In one of these cases the researcher identified a sharing issue due to the fact that survey instruments are imported from other countries and cultural adaptations are required. The interpretation of data is also dependent on cultural factors. Another participant sees no problem in making data available in an Open Access platform to allow others to have different perspectives on the data.

A participant who is also open to data sharing mentions that they usually struggle to revisit data from recent projects. At a given point there were many people collecting data and a collaborative cloud platform was used to share the data. Yet, the group felt access problems and a difficulty to keep track of updated versions of the data. At the time of the interview the data were still in the cloud and the project coordinator was in charge of managing the project archive, being the person responsible to provide the data upon request by email.

In another case, the participant said that data sharing has never been considered by the research group, but in order to do so the funding entity would have to be consulted, and the agreement of all team members would also be a requirement. However, this participant remembered a situation where the data was shared with a researcher from an American university who requested the data.

Finally, the researcher that has experience in sharing data with third parties, told me that when people ask for the data their research group does not have any problems in sharing them, although most people ask for help about the process, post-processing, and how the data was obtained, rather than the data itself.

10.4 REUSE PERSPECTIVES

Similarly to the data sharing perspectives participants were not familiar with data reuse, however, most (8/13) agree that their data has reuse potential, three do not have clear positions and two do not find much potential in the reuse of their data.

Some participants have identified opportunities to reuse their own data, or think that their data can be applied to support related projects. A common argument is that the data can be analysed from different angles and so the reuse potential is immense. One of the participants regularly find themselves using legacy data to compare the data between projects. Hence, the data has to be archived and easily accessible. Some participants also mentioned that their data can be used in the development of software to support research in their scientific fields, or even to improve services for the general public. For this reason a few have disclosed their data to public entities. One researcher who was not convinced of the reuse potential of data already has contributed with data on material properties to a database of a startup company coming out of the research group. In this database users can find the materials that best fit their research objectives.

When asked about the potential of data reuse in different domains the participants had some difficulty in answering. If data sharing and reuse in several cases is not yet a concern, the opportunities for reuse in different domains is a topic the participants had not thought of previously. For one of them I was proposing a funny exercise, therefore raising awareness. Researchers from the social sciences mostly refer other disciplines in the social sciences where the data could be reused, although one mentioned that the data could support work in the field of architecture and mathematics. This participant sees theoretical interest in reuse in these fields but does not know how this could be operationalized. Another common answer is that the data can be explored by statistical experts which is a dimension not addressed by them.

As for the participants who were not sure, or believe that their data hardly has reuse potential, one argued that others will be more interested in the parameters of what has been measured to assess if the material or the type of structure is useful for their work, and not so much in the raw data. Another researcher has a similar opinion but was more assertive in saying that no one will be interested in the raw data, only in the generated tables and figures to compare with the literature. Another participant claimed that top tier journal will not accept a paper if the data has already been exploited in another study. This participant thinks that data reuse is not common in their domain.

10.4.1 Data reuse negative experiences

Although data reuse is not yet a pressing issue in the participants' reality, 6 of them have revealed a situation in which they have had considerable difficulties in reusing data, or where it was not possible to reuse the data at all. These negative experiences are related to inherent difficulties to replicate experimental data, lack of data documentation, cultural differences in data provenance, variables inconsistency and data being outdated. Their accounts have been freely translated into English.

"I have tried to recreate data published in a paper but I could not reproduce it. It

is natural not to get the same conditions. It is always hard to make sure you can get the same conditions from lab to lab. Since it is experimental work we will never have the same behavior. If there is lack of information we contact colleagues and usually we get a response."

"I have this negative experience of not being able to put together databases, or waste a lot of time, because I cannot identify people and the data has to match. In two or three cases I could not match the data. I lost some data, I lost a lot of time putting the two databases together and then I could not do the analysis. For example, I had trouble with databases taken from two different centers, which for the same variable had different names, and I was not sure if the variables match. So, combining databases is a problem. It does not have [contextual information], and if it did...I think it is necessary and would help if there are standardized checklists easy to fill in."

"My issue regarding the use of international databases is the culture. How can we integrate knowledge, that is specific to a culture, in the interpretation of data? I think this raises a lot of problems by creating misunderstandings in the interpretation of data that come from different cultures. Some of them may be solved with sophisticated scientific procedures, to see if there is equivalence in the way people understood the questionnaires, taking into account certain dimensions in different cultures. But I do not think this is enough, especially in comparisons where some have this and others have that. I find it complicated to centralize data, especially data that are sensitive to social, cultural, political and economic issues. I think this requires collaboration between researchers, it cannot be centred on one. There are many risks from that point of view."

"It was a headache to work with the survey data, and with income and expenditure data. I was collecting data from the Statistics Portugal web portal, which was going through an update, and I was limited to only using the data that was already converted to other formats. The analysis of longitudinal data was impossible, since during data collecting the variables have change over time, even their names. There I experienced various challenges. I even include in my thesis a number of suggestions to solve this kind of issues."

"Well, I have had those kinds of experiences. The darkest and most negative issues arises when the data are outdated. For example, during a doctoral project I collected data, and I felt like going back to it, but the social-economic point of view was already very different. I could go back to the data but I could not inquire the same people, so the purpose of revisiting the data was not exactly achieved in this respect. I cannot reuse the data because it was too dated."

"In one of the first surveys I worked on, an old one, the methodological terms had many flaws and later we could not take into account the variables because they... I do not remember if they were not collected or if they were not made available at all. We really had this gap. We knew that we should not analyze the data since we did not have the material for proper analysis. These were variables that should be made available. Basically, the file was incomplete."

10.4.2 Data reuse positive experiences

Successful experiences of data reuse were mentioned by 5 participants. In general, good data documentation, the description of variables, and in one case close collaboration, were the factors deemed to influence reuse.

"To support the simulation part in our publications we have to use properties, which can be curves of material behaviour, and much of this information we don't have, so we have to consult external sources that have proven to be quite useful. We have been able to put many models to work based on that."

"My experience in working with data collected by others is positive. It is usually a job that is done in cooperation with the person who knows the database very well. Of course, there have been doubts and issues, but since it is a collaboration these issues are always solved quickly. Thinking in a more comprehensive sense, like having to get an external database, I do not have such experience and I admit it could be different."

"The documentation was a determining factor for the reuse of data. In the data analysis software the variables were well described. This is very important. One thing that is important is that each item variable has to have the name of the question."

"Recently, I revisited data from a chapter I wrote with a colleague and also the data from an individual project. I returned to the data, not so much for its content, but to the context of data collection, its metadata, to make an analysis about the methodology of scientific research. Basically I took advantage of a research I had already done in order to review the difficulties of data collection and wrote a paper on that in the area of scientific research methodology. After this successful experiment, I would say that all data can be reused for very different purposes. I never thought that after five years I would be able to revisit the data for a completely different purpose."

"Yes, I reused data from a national health survey. I worked on the data and the experience was positive thanks to well-documented data manuals. Each of the variables were described so that other researchers could reuse the data. I had no problems."

10.5 ORGANIZATION PRACTICES

As expected, all participants use their personal computers to store their data, in some cases exclusively, in others combined with other media. Cloud platforms, external discs and portable storage devices were listed as solutions to maintain copies of the data to prevent loss. In experimental domains the data is stored in laboratory computers shared by group members. For one of the researchers it is important to have multiple copies of the project folders spread over several platforms. The redundancy of the folders works as an insurance for any kind of damage that may occur with the data. This researcher also send emails to their own personal account for file versioning. The use of the personal email account was also acknowledged by another participant as a way to access the data from anywhere.

Typically, researchers follow a file and folder structure approach to organize and access the data. Most stated that the files are named with easy to recognize labels. The main folder is usually the project identified by its name. The internal structure of the projects has different approaches depending on the nature of the research and data. One researcher prefers to organize the second folder level by the location of data collection, identified by the place name and the date of collection. If the folder was reviewed by any collaborator it also has its initials. For each location folder there are sub-folders for field work, for the publications and for the protocol. For another participant the folder is structured first by experiment, and then by sub-folders for each technique applied. Usually the file names refer to the type of material or date, and sometimes even to the working conditions, to identify a specific one when working with several conditions. Therefore, the file can be identified by the combination of the material, date and a specific parameter, like the milligrams in experiments where the mass was varied, which results in long file names. In some other cases folders are organized by date, subject, variable or measurement depending on the type of data.

A few participants also refer approaches to make sense of the data. In this case, personal experimental logbooks and laboratory notebooks are traditional supports for the registration of the scientific activity, but the participants also use conventional, unstructured notepads, to make notes. One of the researchers uses text documents to write notes warning colleagues about possible limitations in the data. Spreadsheets are also applied by one researcher to make notes that help to interpret numerical values after obtaining conclusive final results, which then produce a simplified report. This researcher has the habit to print and archive the annotations since they prefer to read them in paper. Likewise, in another case the instrument produces an attached file to print, although the researcher prefers to maintain the digital version. The use of slides to copy and paste relevant data was another approach pointed out as a support always at hand when some details about the data need to be discussed. Nevertheless, this participant is also trying to keep a journal to log the activities that were carried out on a given day and tasks that are important to redo.

Sensitive data also require special care by some researchers. When study participants are easy to identify, two researchers create a separate file with the name and number of the study participant, while the data analysis software only contains the assigned number to match the data, to ensure anonymity and confidentiality.

The said approaches and strategies to manage data are mostly *ad-hoc*, therefore, few researchers recognized some issues related to their practices:

"In times of greatest tiredness documents are hard to find and make sense of. There is always the challenge of remembering what has been done in past experiences, even more so if some structures are not optimized. Often is not possible to interpret the data after some time has passed, or if the data was collected by another group member."

"It is absolute chaos. I need someone to help me. It is horrible. I never know what the latest version is. We have the raw data, then more data comes in and you change it with a name that looks good. If I look at the directory I have several versions and then I spend endless time to know what is the most recent one."

"Sometimes it's very confusing because it is a lot of data. For example in a week when I did a series of experiments and I did not organize the data, in the end it was all confusing. So I have to get the minutes to make sure what happened. I have seen colleagues have to repeat experiments because they did not know what the results were. Organization is essential."

"Recently, I discovered a set of data that I had not yet analyzed, I was extremely

confident that I had analyzed the data. I did not find annotations of what I had done anywhere."

"Where are my projects?"

"When using a cloud service there was so much information. I wanted more specific things and getting to the specific was hard. I clearly feel that there must be a best way to organize the data, because I lose a lot of time. At a certain point the coordinator sent me documentation and I was lost because it had been a few years since I worked directly with that information and I had to ask for a filter "Just send me the folder of this, otherwise I'm here and I spend hours until I find what I want". This difficulty is also probably due to a lack of practice on my part, although it is of a more technical nature, but also to the disorganization of the data."

"It takes time to find because of bad organization. I feel that if I organize better I can get to the data faster. That happens, sometimes it means opening all the folders to find something, or not."

"I lack organization and waste time looking for the data. That is why I am always saying that I have to find a better method, which is not the one I have described, which does not always work. I basically only know where the recent data are. The other I cannot remember, I have to keep looking for them. It's not very nice to say this, but it's true."

10.6 METADATA PRACTICES

All the researchers considered data annotation a relevant activity. Yet, this is not a systematic activity for most. In some cases researchers described practices consistent with the production of metadata without being aware of it, that is, the creation of metadata is more intuitive than formal, and comes naturally as part of the research process. The practices can be considered mainly as *ad hoc* or personal, but there are also metadata resulting from guidelines or based on instrumentation.

10.6.1 *Ad hoc* and personal metadata practices

An activity to annotate data commonly reported is the creation of additional columns in databases, particularly if the database is shared with other people. These annotations are chiefly free text. Text documents, even written in paper, is something one of the researchers asks from subordinates, which are regarded as "*hard discs*" and memory. Another researcher makes several annotations in paper but can also have an open document to make more casual annotations. The latter is concerned with annotating the days the surveys are launched and closed. Some events, challenges or strategies to surpass challenges are also annotated.

Notebooks accompanying researchers throughout the research were also mentioned occasionally. These notebooks can be used to code variables or sort contextual information organized by date. One of the researchers has the habit of documenting in audio, video or photo, which support the information that is poured into reports. An output of a project was a documentary that has been shown in classes and served as a starting point to explore several data from the project.

10.6.2 Guidelines and instrument-based metadata

Certain participants made reference to data annotation procedures that follow soft rules. One researcher has to follow a laboratory minute, which is kept as a diary. As the researcher explained, group members need to record work conditions, what has been done or whether anything unexpected happened. All these details are considered when it comes to processing the data. As the experimental data are collected the researcher writes down the information to make a report, which is then transcribed on the computer. There is flexibility to edit as required but there is a basic standardized set of information that must be recorded, although not everyone follows this rule. Another participant's group takes the relevant notes from a notepad to a spreadsheet with previously standardized columns, while also attaching pictures of the experimental setup to the data.

The methodology or technique adopted in the research can also impose certain data documentation norms or rules. In certain processes it is not possible to escape the protocols underlying each research technique, since these ensure the accuracy of data collection. However, one researcher noted that there may be adaptations that derive from the specificity of each project, namely the dynamics of the team, the subjects, whether it is a more applied or fundamental research. There are baseline protocols that can be adapted for each reality, according to another researcher's practice.

The software used to analyse the data also influences the way some participants document their data, particularly when encoding. A researcher who showed awareness on the software influence in metadata creation noticed that when it comes to statistics, the data are very easy to understand, but other types of data are not intuitive and must be contextualized for the people who are not from the project. In other domains the instrument itself already generates parameters as metadata. One research group configures the experimental program to have a clear reference about the test plan and the instrument automatically generates the required information.

10.6.3 Annotation benefits

Researchers, independently of the approach used, consider data annotation to be an important practice in the research process.

Data annotation has been assessed as pivotal for data reuse and reproducibility of results. For two participants data documentation is directly related to the quality of the data itself. Both described a scenario in which it is important to document occurrences that are out of the ordinary in study participants behaviour, so the data can be analyzed properly. One of these researchers show interest in learning new methodologies and processes to document data, despite an overall feeling that so far the projects were successful in that respect.

The importance of data documentation to good recordkeeping was also considered by some. An argument was that although data may not be physically lost, the confidence to reuse may be limited if essential details are forgotten. The inability to remember can also mean that researchers have to go back to the source material which means significant time losses.

Well-documented protocols are also critical for the success of the research projects. Protocols are the reference document to guide the research and, according to one participant, if these are not well designed it can lead to bypass some procedures and people may have different interpretations in the collection of field data. In one case the research group writes in a file what went wrong and shares this file via email, however this information is frequently lost and therefore the researcher assumes that this is a procedure that has to be improved.

11 DATA DESCRIPTION SESSIONS RESULTS

In this Chapter I describe the 13 data description sessions that were carried out between January 2018 and September 2019. I start by presenting each individual session in a chronological order. For each session I start with indicators regarding the number of descriptors completed by the researchers and the amount of time researchers spent in data description. Moreover, I present the most pertinent details observed during the sessions, some related to the researcher´s behaviour and others to their remarks. A brief analysis of the descriptors chosen is also included, particularly by comparing them to the metadata categories proposed by Qin et al. [89]. This analysis yields preliminary results on the quality of the metadata created by each researcher. To conclude, the results from the follow-up questionnaire are presented to discern personal opinions and perspectives about the sessions.

Then I present an overview of the data description sessions results. The total number of descriptors filled in and the total amount of time spent by the researchers, as well as the average values, are estimated. Another aspect explored is the descriptor count, which gives indications on the most commonly used descriptors and the metadata categories that researchers are most likely to fill out. This Chapter ends with the overall results from the follow-up questionnaires to assess how researchers evaluate the data description activity.

The datasets that support the results from the data description sessions[1], and for the follow-up questionnaires[2] are publicly available.

11.1 INDIVIDUAL SESSION RESULTS

Session 1: Family Psychology, Faculty of Psychology and Education Sciences U.P., 31 January, 2018

In the Family Psychology session a descriptive statistics dataset concerning children's emotions regulation, and parents' work-family conflict and psychological availability, was described. The researcher that participated in this session was accompanied by a colleague that had collaborated in the collection of data. Together they have created 13 key-value pairs in 30 minutes. None of the researchers had any data description experience but they became familiar with the proposed task quite easily. They talked to each other during the session to discuss the meaning of some descriptors. The pair was very careful in the selection of descriptors and with the information provided, given the perspective of subsequent data publication. The researchers have selected metadata elements for methodological information, such as the *Sampling Procedure*, *Time Method* and *Sample Size*. In this

[1] Castro, J. A. (2019). Multi-domain data description sessions data. INESC TEC research data repository.https://doi.org/10.25747/gc4j-vm58

[2] Castro, J. A. (2019). Multi-domain data description sessions follow-up questionnaires. INESC TEC research data repository.https://doi.org/10.25747/x9ak-5w15

case they considered the DDI subset convenient for their data, especially because the concepts are familiar to the terminology they regularly adopt.

The metadata record produced included elements from six different categories. It included Administrative metadata (2 elements), Descriptive metadata (4 elements), as well as Technical and Semantic metadata (1 element from each category). The *Abstract* was used to describe the content of the dataset and its objectives.

Almost half of the descriptors used is scientific oriented, 4 elements are for Context metadata purposes and 1 Temporal element. Geospatial information was provided in the *Sample Size* to limit the geographic area where the data was collected. Therefore, the record contains information about the responsible party, related publication, subject, variables and minimal study design information. The researcher that filled in the follow-up questionnaire expressed the opinion that the information recorded is sufficient but may perhaps be improved, with no further recommendations on how this can be achieved.

The researcher found data description 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 the researcher it is possible to make use of existing databases, thus not overload participants with new questionnaires.

There is a moderate degree of interest regarding data management that can be improved if metadata models are made available in their projects, with increased work visibility through data citation and even more if it enables data reuse.

Session 2: Clinical Psychology, Faculty of Psychology and Education Sciences U.P., 23 November, 2018

The metadata recorded in this session pertains to the validation of three assessment tools in the area of the psychosocial impact of genetic testing for cancer risk. A total of 17 descriptors were filled in about 30 minutes. Although short and very objective, the information provided was easy to interpret for a layperson in the domain, and realistic, as actual values were used, without any abbreviations.

Although the *Description* was already provided for the description of the project, the researcher used this element again to inform about the type of data. This is not rigorous metadata but shows that the researcher differentiated the description of the project from the description of the dataset content. The *Deviation From Sample Design* was also filled, showing consistency to what has been said during the interview regarding the need to document the contingencies in the research process. I could observe some anxiety during the description session, namely in relation to the time that had to be dedicated to the task. Two constraints may have limited the production of more detailed metadata. First the data description session took place amidst a busy schedule; second, as the follow-up questionnaire would reveal, the perception about the benefits and the objectives of metadata creation were not yet fully clear.

Nevertheless, the metadata record that resulted from this session was complete and diversified in terms of categories used. Five key-value pairs correspond to Administrative information for *Access Rights*, people involved and to which *Audience* the data is intended for. On top of that the researcher filled in 3 Descriptive elements and one for Technical metadata. Moreover, 7 key-value pairs provided the context for the production of data and type of study performed, while 2 different Geospatial elements were recorded, though with redundant information about the country where the study took place. No Temporal metadata was provided, something that the researcher said was not relevant for the objectives of the study. The researcher thinks that the metadata is sufficient and that there is no need for more information.

In the follow-up questionnaire the researcher did not identify particular usefulness in data description, since was not able to fully understood how the knowledge resulting from describing the data can be applied. After the participation in the session and the production of a detailed metadata record in the process, the researcher still thinks that this is an important subject, yet still an overly abstract activity. Data description was perceived as a very boring activity, however, in contrast to being a very motivating one. Metadata production was also seen as slightly difficult and time-consuming. As for RDM, the interest is very high since it helps to organize, store and reuse the data. The availability of more tools for data management, work visibility via data citation and data reuse are the main motivators for this researcher.

Session 3: Consumption Sociology, Faculty of Arts and Humanities U.P., Sociology Department, 12 December, 2018

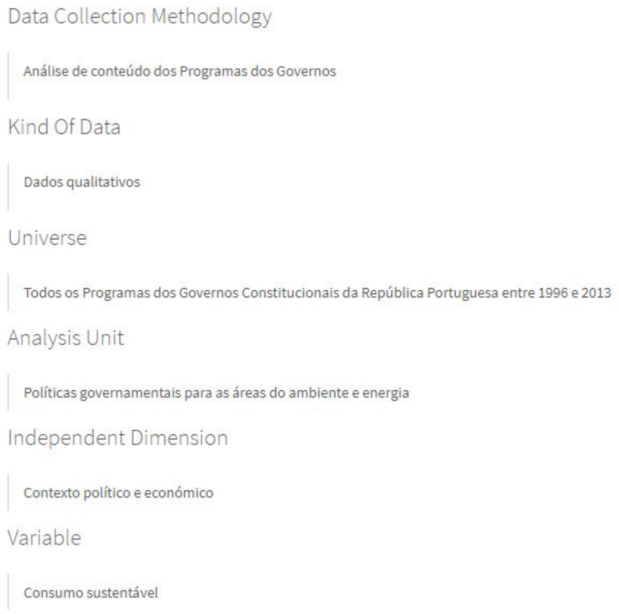
During this session the researcher described a dataset about the role of political intervention and new social movements in sustainable consumption. This dataset was produced by performing content analysis over governmental programs.

A total of 21 descriptors were filled in a session that took 75 minutes. The length of the session can be explained not only by the number of descriptors used, but also by the fact that some fields required more text. Most of the time the researcher also explained the choice of particular descriptors while browsing the vocabularies in Dendro. Moreover, the description was supported by documentation that the researcher consulted to confirm the accuracy of the metadata provided.

The result was an extensive metadata record, with details such as the *Table of Contents* and the *Accrual Periodicity*. A broader set of metadata categories was explored in this case. Technical, Semantic and Identity elements were selected to refer to the *Analysis Unit*, the *Subject* and an identifier for the funding entity. Combined, the researcher used 11 elements for Administrative and Descriptive purposes. The Context metadata was also rich to inform the objective of the study and the methodological steps, with a total of 4 descriptors in this category from the DDI subset, with two more for Geospatial and Temporal information (Figure 18). According to the researcher, there was no need for more metadata.

At the end of the session the researcher asked if it would be possible to have access to the metadata record created for future reference. This request, together with the focus in metadata creation, suggest a commitment and interest in a previously unknown activity.

The answers to the follow-up questionnaire are consistent with this interest and with an increased RDM awareness. For instance the concept "metadata" was used in a comment in the follow-up questionnaire. In the researcher opinion data description is a useful activity that allows not only data to be stored with all relevant information, but also allows data to be reused. Data description was considered to a small extent both interesting



Data Collection Methodology

Análise de conteúdo dos Programas dos Governos

Kind Of Data

Dados qualitativos

Universe

Todos os Programas dos Governos Constitucionais da República Portuguesa entre 1996 e 2013

Analysis Unit

Políticas governamentais para as áreas do ambiente e energia

Independent Dimension

Contexto político e económico

Variable

Consumo sustentável

Figure 18: Example of the metadata created in the Consumption Sociology session

and practical, but also a little discouraging. The researcher did not find the activity either difficult or easy, time-consuming or fast.

The degree of interest in RDM was deemed moderate due to the fact that it prevents the loss of data and avoids forgetting relevant information. The systematization of information also allows its reuse. Like the previous researchers the motivation for RDM depends mostly in credit and reuse. In this case enabling contacts with other researchers and partnerships is also value.

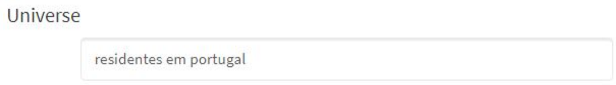
Session 4: Services, Faculty of Engineering U.P., Department of Industrial Engineering and Management, 10 January, 2019

In the Services case the researcher opted to describe a recent quantitative dataset regarding the adoption and usage of fitness trackers. A total of 11 descriptors were used in a session that lasted for 25 minutes.

The researcher started by exploring DC elements. The descriptor *Access Rights* was considered important but not relevant for the session since the data and its description would not be made available for others, so the researcher opted not to provide this metadata. Then the researcher proceeded to the exploration of DDI elements. The *External Aid* descriptor was considered for the type of studies the researcher did, since it allows for the capture of any material, such as text card or images, and the instruments that supported data collection. However, for this specific dataset it was not considered pertinent. Moreover, the researcher considered the *Data Collection Software* as irrelevant, stating that it would rather have a descriptor for the representation of the software for data analysis, information that was later recorded under *Software*.

The production of metadata was done with ease and interest by scrolling patiently through the list of DDI elements, explaining the preference or the motive to reject some of them. However, the detail and accuracy of the final metadata record is limited, with only minimal information in most fields.

The metadada is mostly Context, with 5 elements used. Three elements represent the Descriptive category, while one element was selected for the representation of Administrative, Technical and Temporal metadata. No Geospatial elements were selected, yet the *Universe* metadata capture the necessary information about the study boundaries (Figure 19). The metadata record is balanced taking into account several comments suggesting that the metadata would be populated if the goal was to disseminate the data. The researcher also recognized that although the metadata may be sufficient more information can be added, for instance by including metadata about the end of the data collection.



Universe

residentes em portugal

Figure 19: Universe metadata to provide study geographical boundaries

On the other hand the research showed awareness by recommending that the data should be accompanied by documentation, such as protocol scripts that can somehow replace the metadata and provide additional information and assist in the reuse and production of new data, which can promote citation. Finally, the researcher suggested that vocabulary terminology in Dendro should be changed considering that the designation "Dublin Core" and "Friend of a Friend" do not convey any meaning.

The degree of data description usefulness was evaluated as important since data reliability and access are critical to the quality of the research. As for the description activity the researcher considered it relatively interesting and easy but also a little impractical. The degree of interest in RDM is also moderate and the researcher seems to be more interest in improving the communication and data sharing with close colleagues, in obtaining credit via data citation and in the possibility of extending contacts and partnerships.

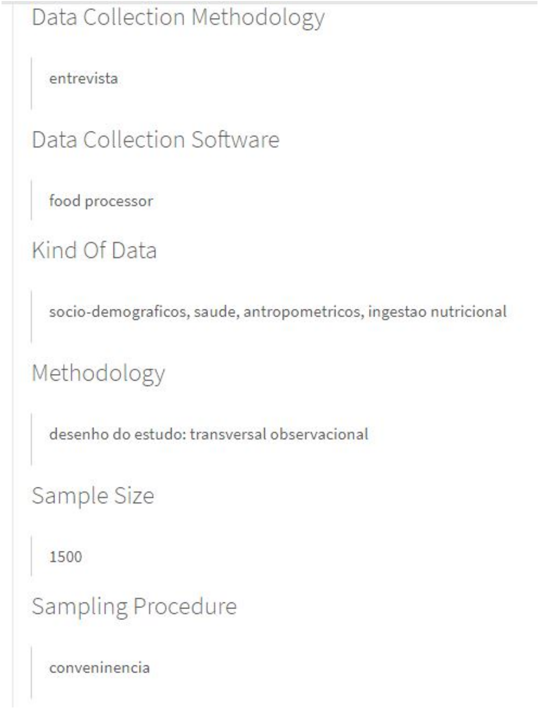
Session 5: Nutrition, Faculty of Nutrition and Food Science U.P., 18 February, 2019

The Nutrition researcher created a metadata record with 11 descriptors in a 25 minutes session to describe a dataset for the assessment of the nutritional status of people with dementia.

Although the *Abstract* was selected, the information it contains was a random selection of characters with no meaning. The researcher justified that in the context of this session a full description of the abstract was not needed. The participant only wanted to show that it is a descriptor of interest that would be used in a real scenario. The same happened with the *Date Created* and *Data Collection Date*, in this case the participant left the default values, *i.e.* the date of the session.

The researcher only viewed descriptors from recommended vocabularies, DC and those related to DDI. The descriptors of interest were added one after the other and filled in the end. The target *Audience* was identified, while the *Kind of Data* information was provided as a list of subject terms that might facilitate access to the dataset if properly indexed. Moreover, no information was added for the location, or geographical boundaries, of the interviews, despite the fact that the naming of folders by location is one of the strategies used for personal data organization.

This effortless description was compensated with the explanation of the reasoning behind the choice of some descriptors, such as mentioning that the descriptors and their meaning are close to what is specified in the research protocols. The researcher created 4 key-value pairs for Context metadata, by shortly describing the *Methodology* and other related descriptors, as shown in Figure 20. The *Data Collection Software* information was added, since it makes sense to inform others about the software for data analysis.



Data Collection Methodology

entrevista

Data Collection Software

food processor

Kind Of Data

socio-demograficos, saude, antropometricos, ingestao nutricional

Methodology

desenho do estudo: transversal observacional

Sample Size

1500

Sampling Procedure

conveninencia

Figure 20: Methodological metadata provide by the Nutrition researcher

In the response to the follow-up questionnaire the researcher indicated that the metadata was sufficient and that no further information would be needed.

At the end the researcher asked if it would be possible to use Dendro and requested the link, denoting an interest in the functionality of organizing project data for better data search. At the end of the description session the researcher used the term *Dublin* to refer to the initial set of descriptors selected showing awareness to the task.

Data description was perceived as moderately useful. For this researcher it allows for a better organization of information and process control. The activity was characterized as slightly interesting, easy and fast, but on the on the other hand a little impractical and demotivating. There is some interest in data management by the fact that data can be managed more easily. Interest in RDM can increase if metadata models are available for their projects, and if RDM helps to improve the communication and data sharing with colleagues and to comply with mandates.

Session 6: Work Psychology, Faculty of Psychology and Education Sciences U.P., 28 February, 2019

This session generated the shorter metadata, with a number well below the average number of filled descriptors, with 4 in 15 minutes. This could have something to do with the fact that this participant was the youngest among the researchers and also an international student just starting their PhD at the date of the session. Moreover, during the interview the researcher also revealed some confusion about the concept of metadata. The communication with this participant was always made in English, not the native language for both, which may have hindered the communication.

The described dataset consists of interviews conducted with factory workers to assess job satisfaction based on personality types. The researcher created metadata for *Subject*, *Kind of Data*, and short references to *Sampling Procedure* and *Data Collection Methodology*. The 15 minutes session duration is justified by the need to explain the meaning of some descriptors for which the researcher showed curiosity while not finding them relevant to fill in. The resulting metadata record is, therefore, very short. However, the researcher considered that it was enough information and that there was no need for additional information. What can be inferred is that the researcher only consider the metadata for personal organization of the data since the moderate level of data description usefulness was associated to data saving benefits.

Data description was viewed as an impractical and discouraging activity, although somewhat interesting. There is some interest in RDM since it helps to manage data more easily, in the researcher´s own words. The interest can be reinforced with the availability of tools, by enabling more contacts and improved scientific evaluation.

Session 7: Structural Adhesive Joints, Faculty of Engineering U.P., Department of Mechanical Engineering, 06 March, 2019

The session with the researcher working with Structural Adhesive Joint was focused on the description of a dataset about the design methodology for impact resistant bonded multi-material automotive structures. The researcher took the initiative to create three folders in Dendro to exemplify how the folders and files are usually structured. One folder was dedicated to Material Characterization Tests files, such as the one described in the session and two others were created, one for Structural Tests, another for Simulation Models, as shown in Figure 21.

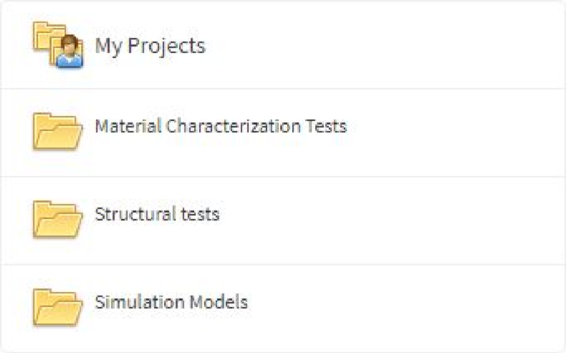


Figure 21: Folder structure created by the Structural Adhesive Joints researcher

The metadata record was completed with 14 descriptors in 28 minutes. The metadata provided was of high-quality and the researcher showed a very high level of awareness about metadata production. The use of metadata categories was both balanced and diversified. There were a high number of Experimental elements selected, with 7 key-value pairs created in this category, mostly to describe specimen parameters and environmental conditions. Two more Context elements were used, one for the type of study and another for the instrument used. Furthermore, two descriptors were chosen for the Administrative, Descriptive and Identity, and one for the Temporal, Semantic and Technical categories. Given the nature of the data the use of Geospatial metadata was not applicable.

The researcher commented most of the choices made regarding the selection of descriptors. The descriptor *Conforms to* was used to register the assay standard. The information captured in this field is an identifier that points to a resource that includes several parameters. Thus, there was no need to use some of the available descriptors in Dendro, such as the *Sample Dimensions*. However, the researcher noted that if an assay does not follow a standard it is interesting to have the necessary fields, even to compare an assay that follows a standard procedure to another who does not. Even at the risk of redundant information the researcher acknowledge that it may be useful to have additional fields to capture the same information provided by the assay standard.

After the selection and filling of DC elements the researcher stated that this was information his group was already accustomed to register. Moreover, the registration of materials involved in the test was recommended, and therefore *Instrument* was considered an important field to register the test machine. When asked about if the descriptors in Dendro tailored for this domain were already part of the dataset files it was observed that only partially. For instance the *Temperature* and the *Test Velocity* values are not included in the dataset. Most descriptors are parameters that are usually registered manually and not an intrinsic part of the assay. At the end the researcher reread the information to confirm satisfaction with the metadata. For the production of metadata for the Simulations component the researcher noted that it would be useful to have information about the reference model and mentioned that perhaps a series of fields should be created for this purpose. As a consequence, the researcher outlined what would be the necessary fields and possible values. Then asked if a user could create the fields directly in Dendro.

The degree of usefulness of data description was classified as high due to the need to carefully describe all the test conditions in order to be able to replicate results or understand unexpected results. The researcher thinks that data description is somewhat an interesting activity, easy, fast and very practical. The metadata record was assessed as complete. The interest in RDM is also high since the participation in a number of projects, consultancy and other research activities leads to a substantial increased in the amount of data generated that can only be correctly interpreted, stored and reused if there is a data management policy. The researcher would have greater interest if having to comply to mandates, if the communication and data sharing is improved with colleagues and if it brings more contacts and partnerships.

Session 8: Organizations Sociology, Faculty of Economics U.P., 07 March, 2019

The researcher from the Organizations Sociology domain created 12 key-value pairs in 45 minutes in order to describe a dataset resulting from 7 organizational case studies of social entrepreneurship. In this session some technical failures were experienced, thus the duration of the session is inflated. Moreover, it was not possible to save the information recorded in Dendro by the researcher, therefore, it was necessary to copy the metadata to a text sheet and after the session I inserted the metadata again in Dendro.

Project documentation was used to support the description of data and the researcher mostly resorted to DDI elements to describe the dataset. The researcher started by exploring the DC terms and noticed that the concepts were not related to their data, stating that the concepts for the "documental sciences" are not intuitive. Nevertheless, there were DDI descriptors that have caused some doubts, especially the *External Aid*. In this case the researcher was hesitant between the media to capture the audiovisual content or its output, and decided to include information about the latter. Another concept that raised some doubt was the *Time Method*. This descriptor although considered was not added. The researcher differentiated between the *Methodology*, as a route to follow that includes several techniques that are operational instruments, and the *Data Collection Methodology* to describe strictly the techniques, but would prefer a data collection techniques field for this metadata. Yet, assuming that this preference could be a scientific preciosity because in some areas the methodology covers what were the methods, in sociology the method is identified as intensive and extensive analysis, then the instruments are the interview techniques, direct observations, among others.

There was some difficulty in providing the description for the *Sample Procedure* since it was not possible to remember how the sample was constituted, even after consulting the document, which led to the conclusion that - *"Since I do not have Dendro to organize the data, the data are out there"*. A fictitious value was provided to make a point about the relevance of the *Sample Procedure*. Some limitations were identified regarding the usage of controlled vocabularies in Dendro. For instance in the description of the project language the researcher needed to insert both Portuguese and English, yet it is only possible to select one value. The use of a calendar to register the *Temporal Coverage* was not seen as practical since the researcher needed to go back a few years in the calendar and asked for a more practical way to insert this information. Overall, the researcher admitted being a bit hesitant in the choices that were made, recommending the use of more clear definitions.

Despite the aforementioned issues the metadata record combines 4 Descriptive with 5 Context elements, with two more elements from the Temporal and one from the Semantic categories. There is no Geospatial metadata but that its not essential for exploring the dataset since this information is easily identifiable by its content. The researcher believes that the metadata does not yet suffice, lacking detailed information on the profiles of the organizations surveyed, the people interviewed and observed. This indicates the willingness to develop higher quality metadata in the future. The degree of data description usefulness was evaluated as high for more rigorous production, management and use of the data. As for the description itself, the activity was found to be practical and relatively interesting and easy, even though a little discouraging and time-consuming. This may be related to the technical problems in Dendro during the session that made it longer than would have been necessary if these had not occurred.

Given the perceived benefits of RDM for the dynamic of the research process the degree of interest in RDM is very high. The availability of metadata models, enabling contacts with other researchers and partnerships, and reusing the data in the mid to long term were selected as the top motivators.

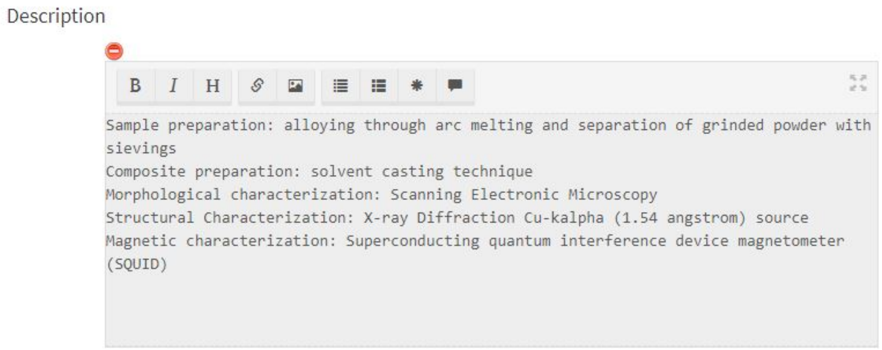
Session 9: Magnetic Material, Faculty of Sciences U.P., Department of Physics and Astronomy, 28 March, 2019

This session was the most productive in terms of descriptors filled in, 28 in total, in a short period of time, 16 minutes. The dataset described was collected in the context of a PhD thesis regarding the study of multifunctional compounds.

The researcher prioritized domain-specific metadata elements and followed a logical order to produce the metadata, that is, structured the description in a hierarchical order, starting with the *Sample Composition* and proceeded with finer metadata. As such, the researcher explained that the *Characterization Technique* is a sub-field of *Method* and should be described in its sequence.

Likewise, it was suggested that the descriptors layout should be structured to ease the description and not as independent elements within the same vocabulary. Furthermore, the researcher commented that the available descriptors in Dendro cover the entire research workflow, from the representation of nanoparticles to composite information, and part of the information is transverse to all the research process. All the Characterization information is standard and includes, in this order, Morphological, Structural and finally the Magnetic Characterization.

Most descriptors used correspond to Experimental parameters, with 16 key-pair values created in this category. This explains the amount of descriptors used in such a short time, as the metadata are very fine grained and the values introduced are mostly numerical. The researcher selected 13 descriptors that were identified via the content analysis technique (detailed in section 7) and added fields that were not available in Dendro, namely the *Sample Preparation*, *Composite Preparation*, *Morphological Characterization*, *Structural Characterization* and *Magnetic Characterization*. In order to do so the *Descriptor* field was adapted as shown in Figure 22.



Sample preparation: alloying through arc melting and separation of grinded powder with sievings
Composite preparation: solvent casting technique
Morphological characterization: Scanning Electronic Microscopy
Structural Characterization: X-ray Diffraction Cu-kalpha (1.54 angstrom) source
Magnetic characterization: Superconducting quantum interference device magnetometer (SQUID)

Figure 22: Adaption of the *Description* element to include unavailable descriptors in Dendro

Some other descriptors were adapted so that the information that the research found relevant was not excluded. This was the case of the use of the *Specimen Property* to introduce values regarding the *Mass*, *Diameter* and *Height*, although a *Specimen Height* descriptor is available on Dendro. Yet, sample is the preferred term for the kind of experiments conducted rather than the specimen.

As concepts from outside the domain were explored the researcher questioned, after I explained the aim of the DC vocabulary as a whole, whether this vocabulary is for keywords. The researcher asked if the *Abstract* should be described, but gave up the idea because this information was not relevant to the data. The researcher also asked about the objective of the *Accrual Method* descriptor.

The use of metadata categories was diversified, with six descriptors for Context metadata and three for Temporal, like the *Deposition Time* and *Annealing Time*. For the Administrative, Technical and Descriptive categories one descriptor was used for each. The result is a rich metadata record that has limitations in terms of quality since it include several parameters and study design information but short information to provide access. Some of the non domain-specific descriptors were also filled in with little accuracy. The researcher seems to agree with this remark since has assessed the metadata as enough but observed that some more information might be added.

The usefulness of data description was highly held due to the fact that the way data is obtained can change some properties or lead to errors. Therefore, it is very important to know the parameters from the production to the characterization of magnetic samples because of possible contamination or oxidation - in case of metallic samples. The opinion about the data description is that it is rather interesting, motivating, easy and very practical activity.

The degree of interest in RDM is moderate and the top motivators are the support to personal scientific evaluation and work visibility by data citation, as well as enabling more contacts and partnerships.

Session 10: Sustainable Chemistry, 02 April, 2019

This case was exceptional because it was not possible to held the session in person, despite the participant best efforts to find a date for the session. When there was greater schedule flexibility, technical issues did not allow having suitable descriptors available for the sustainable chemistry domain. Without appropriate descriptors, conditions similar to the other sessions were not guaranteed so I opted to postpone the session. The solution found was to explain in detail to the researcher, via email, how to navigate in Dendro and required confirmation when all the instructions were understood. Among other aspects, the researcher was requested to pay particular attention to the concepts in the vocabulary related to chemistry and to the DC list. The researcher was also asked to measure the time spent in the activity.

The Sustainable Chemistry researcher described a dataset about oxidative processes for water treatment, that supported the PhD thesis, on the 2nd of April, 2019. The metadata record produced consists in 13 key-value pairs produced in approximately 10 minutes, according to the researcher. Although the metadata is heterogeneous in the type of categories used all the descriptors were chosen from the same vocabulary, the ontology based in the sustainable chemistry domain (section 7). The most used category was the Experimental parameters with 8 fields, including information about the experiment set-up (e.g. *Solar Light Intensity* and sample *Degree of Purity*). Two key-value pairs for Context metadata, the *Absorbent Measurement Instrument* and the *Photocatalytic Reaction Vessel* are also part of the metadata. The researcher created metadata regarding the model used for the *Absorbent Measurement Technique*, and the *Suspension Stirring Time*, corresponding to

Technical and Temporal metadata. The description also contain a *Sample Reference* identifier. Despite the richness of scientific-oriented metadata, there is no Administrative or Descriptive elements, therefore this metadata record cannot be regarded as having quality, as there is a lack of information to facilitate access and the discovery of the data.

The researcher mentioned that has consulted the DC elements but did not find any of relevance to the data being described. The metadata provided was considered sufficient, yet it could be improved to contain average absorbance values, the researcher said. The usefulness of data description was assessed as moderate particularly to support data management and to archive information. Moreover, data description was considered an easy, interesting, practical and very motivating activity, but it has to be taken into account that only domain-specific elements were used. Although only spending 10 minutes in the task the researcher did not find the activity to be either fast or time-consuming.

As for the interest in data management, this was seen as very high. The interest is justified by the influence that well documented experiments may have in further work. The most motivating factors for engaging in RDM for this researcher are the availability of metadata models for personal projects, enabling contacts and partnerships, as well as the possibility to reuse data in the medium to long-term.

Session 11: Cultural Studies, Centre for English, Translation, and Anglo-Portuguese Studies, 04 April, 2019

The Cultural Studies dataset is the result of a systematic survey of articles of the periodical “O Vegetariano”. At the time of the session the researcher was under pressure on the agenda, and this seems to have contributed to a more rushed session when compared to the others. In this context it is likely that the lack of available time influenced the number of descriptors filled in. The session duration was 16 minutes and 8 key-value pairs were created.

The researcher only selected descriptors from the DDI subset and showed no particular interest in using complementary vocabularies. With a greater availability the metadata could evolve into a comprehensive and accurate record.

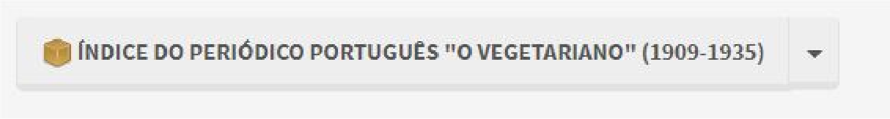

ÍNDICE DO PERIÓDICO PORTUGUÊS "O VEGETARIANO" (1909-1935)

Figure 23: Temporal coverage information represented in the project name in the Cultural Studies session

Nevertheless, the metadata have information on the date of data collection and the total number of articles surveyed, although information about the temporal coverage is only explicit in the project name created in Dendro (Figure 23). The metadata also have information on support materials and the entity from which the data was requested. The metadata contains three key-value pairs from the Descriptive and Context categories, one descriptor for Temporal and another one for Technical metadata.

In the follow-up questionnaire the researcher expressed the opinion that the metadata were sufficient to represent the dataset and that no further information would be needed. The usefulness of data description was seen in relation to data conceptualization and structuring the data more efficiently. Even though the session only lasted 16 minutes the activity was considered time consuming, a perception that may be related with the hurry to go to other activities on the day´s agenda. On the other hand, data description was perceived as moderately interesting, motivating and practical.

The degree of interest in RDM is high because it can lead to different insights and readings over the same topic. The availability of tools for RDM, improved communication and better data sharing with close collaborators, as well as enabling more contacts with other researchers are the top motivators for the Cultural Studies researcher.

Session 12 : Health, Faculty of Pharmacy U.P., 22 May, 2019

Two researchers have participated in this session, the researcher that was interviewed and a PhD student who was working on the data chosen for the description session. This student was in charge of registering the metadata in Dendro and had the data and associated documentation open on his personal computer. A dataset about self assessment of fragility, captured to make the validation of a survey based on an application, was described in 20 minutes in a collaborative effort between the two researchers. The final metadata record includes 16 key-value pairs and is diversified in the categories used. However, only descriptors from DC and DDI were selected.

Although the researchers' domain is the Life and Health Sciences, the dataset was created via questionnaire so I recommended the researchers to start with DDI. The researchers briefly discussed the selection or rejection of available descriptors. They quickly understood the concepts represented in the DDI vocabulary, which made it clear that this was a suitable vocabulary for the creation of metadata and that the researchers were well acquainted with the concepts. The same is true for their assessment of DC elements, yet they were quicker to reject most concepts. The meaning of *Coverage* caused some doubt and was included in the metadata after a short explanation. Halfway through the session they asked if the metadata should be made in English or Portuguese and were told that it would make no difference in the session context but this doubt made it opportune to provide insight on the advantages of describing data in English in the future. Another aspect that shows their awareness and commitment to the task was their intention to include all of the survey questions under the descriptor *Question*, however since they had the questionnaire in a file, the file was uploaded upon recommendation (Figure 24).

The Descriptive metadata has information on the survey objectives and benchmark on which it was based, while the Administrative metadata concerns the people involved in the project and its target audience. Moreover, 6 descriptors were filled in in order to provide Context metadata, such as the *Methodology* and the *Data Collection Methodology*. On top of that the researchers created metadata for the *Sample Size*, *Universe*, *Sampling Procedure*, *Instrument* and *Collection Mode*. The metadata also has one descriptor for the Semantic (*Subject*), Technical (*Format*) and Geospatial (*Coverage*) categories. There is no Temporal metadata, although this information could be usefull, bearing in mind that the assessment of fragility may be linked to a specific economic and social context. Overall, this record has metadata that can support search and access to the data, has a balanced description with the use of standards that promote interoperability and sufficient study design information that may ease reuse. Thus, this record may have FAIR metadata potential.

The PhD student who filled in the metadata was the person who responded to the follow-up questionnaire. In their opinion the usefulness of data description is very high because it helps to systematize everything in a simple and more correct way. The metadata was considered sufficient with no need for more information. As for the data description activity the researcher found it slightly easy, fast and practical, but a little discouraging. The interest in data management is also high with no further comment, and the interest can increase if it helps with the communication and data sharing with close collaborators and brings greater visibility to their work via data citation. The other motivator selected by this researcher is the possibility of having to comply with funding agencies mandates.

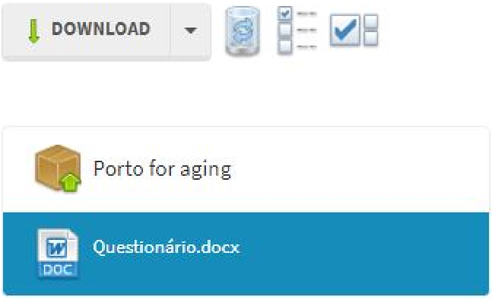


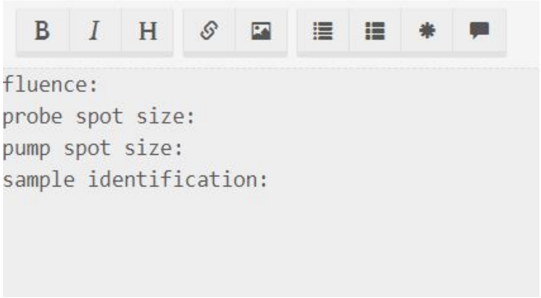
Figure 24: Upload of questionnaire file

Session 13: Magnetic Dynamics, Faculty of Sciences U.P., Department of Physics and Astronomy, 18 July 2019

The duration of this session was 31 minutes and the researcher filled in 7 descriptors related to ultra-fast magnetic dynamics. The apparent imbalance between the session duration and the number of filled descriptors is a result of the detailed explanation given by the researcher behind the selection of most descriptors. The session was more of a talk about what factors would be determinant for the adoption of tools like Dendro in this field of expertise. Likewise, the researcher assessed that the available descriptors in the physics ontology make perfect sense for colleagues in the same lab, working within material science, not necessarily in the same project, but not for personal purposes.

The 7 key-value pairs produced are part of the Experimental parameters category, with three elements, two elements for Context metadata, and also two elements for the Temporal category. The researcher started by introducing the name of the *Characterization Technique*, and stated that the *Pulse Duration* and *Pulse Energy* are adequate for this context. When talking about descriptors related to the substrate used in the experiment the researcher mentioned that if the samples were not provided in the first place and they had to produce them, then they would have to use the available descriptors for the type of substrate and how the substrate was cleaned, perhaps its dimensions and at what temperature should the substrate be for the deposition. This shows that suitable descriptors were available on Dendro for this specific domain, however the conditions of the experiment that generated the data to be described did not make them relevant for this session.

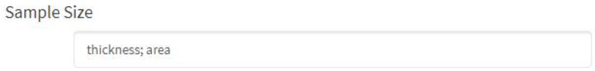
During the session the researcher also recommended four descriptors that were not available on Dendro. The DC *Description* field was then used to make a note about these elements but no values were given (Figure 25). These prospect descriptors are the fluency, probe spot size, pump spot size and the sample identification. The researcher mentioned that this kind of information is often needed and is recorded in the file name or in the form of a commentary file, but may not be useful to others. Such information has value and has to be maintained for further processing of data, but when discussing with colleagues what happened in the system there is no need to tell how the calculations were made.



```
fluence:
probe spot size:
pump spot size:
sample identification:
```

Figure 25: Adaption of the *Description* field to propose new descriptors

Overall, the researcher showed preference for starting with a few generic descriptors and then unfold to a finer description. The type of technique would be the first metadata entry and, depending on the technique, the methodology could be specified, followed by the experimental setting. From here the metadata could have numerous details. Based on the preference for generic descriptors, I asked the researcher to see if DDI descriptors would be of interest. The answer was positive as the *Data Collection Date*, the *Data Collection Methodology* and *Sample Size* have a sufficiently generic quality to allow the description to proceed. Of these, only the *Sample Size* was not filled in. In this case the researcher added information in the Sample Size to illustrate that it should contain the thickness and sample area (Figure 26).


Sample Size

thickness; area

Figure 26: Sample size metadata should include thickness and sample area information

A specific doubt that arose during the session was whether Dendro makes it possible for the researcher to add new descriptors, so that metadata creation is not limited. The researcher recommended that, except for standardized fields, there should be a way to add new fields, as opposed of going through a large list of descriptors where it can take a considerable time to find the one that fits the expectations. In the researcher opinion, as it is, Dendro almost requires a data curator for each researcher, which can lead to its rejection by researchers.

The metadata produced in this session is short but the researcher showed awareness throughout the session and if Dendro met some conditions it could easily have created better metadata, even though DC elements were promptly discarded. The researcher agreed that the metadata is not sufficient and that more information has to be added for the identification of the medium sample, measurement parameters related with the equipment for the detection and measurement acquisition, spot size of the laser on the sample surface, as well as the magnetic field applied. Some of these data may be available as part of the data files, but having access to them without having to open the files can be helpful. Therefore, the researcher highly valued the usefulness of data description for a quick identification of the measured and recorded data to support analysis. The ease of access to the data and to the measurements parameters was also recognized as a benefit of data description.

The researcher thinks that data description is time-consuming, somewhat boring and discouraging. The degree of interest in data management is moderate, yet it seems that this opinion is focused in the session developments rather than RDM as a whole. The researcher noted that the interest depends on the extent to which the tool (Dendro) is easy to adopt in everyday life rather than an additional work and time spent filling in data to be recorded. As such, the researcher chose the availability of tools, and of metadata models for personal projects, as motivators for greater interest in RDM. The improvement of communication and data sharing with collaborators was also selected as a possible motivator.

11.2 OVERALL SESSIONS RESULTS

Overall a total of 178 descriptors were filled in by the researchers in a total description time of 351 minutes. This corresponds to an average of 13,6 descriptors and 27 minutes per session.

There are, however, clear outliers. In domains where the metadata are mostly numerical, we may expect a high number of descriptors filled in within a short period of time. This was the case of session 9, in which the researcher filled in 28 descriptors in only 16 minutes. This was the session that produced the most metadata and its duration is much shorter than the average. In contrast, the researcher from the Consumption Sociology domain, created a metadata record with a number of descriptors well above average, 21 in total, in what by far was the longest session, taking 75 minutes. Here, the metadata has a substantial volume of text and the researcher also took time to consult documentation and explain some of the options made during the session.

On the other hand, the researcher in session 6, Work Psychology, has a very low number of descriptors filled in, only 4, and in line with this, the time dedicated to the task is also well below average (15 minutes). In the other cases in which less metadata was produced, session 11 and session 13, with 8 key-value pairs each, the time dedicated to the task is also disparate. The Cultural Studies researcher took 16 minutes to create the metadata record, while for the same number of descriptors the researcher from Magnetic Dynamics, in session 13, took 31 minutes. The differences between session 6, session 11 and session 13 are related to the researchers' attitude. In the case of Session 6, the Work Psychology researcher had some difficulties in understanding the objectives of metadata, the researcher in session 11 was in a rush to do other activities, and in the case of Session 13, the duration of the task had to do with a very critical and reflexive attitude of the researcher towards the task. Finally, session 8 also had an above average duration of 45 minutes, explained by a few technical problems during the session.

Nevertheless the median of 13 descriptors in 25 minutes is very close to the average. From a qualitative perspective these numbers seem realistic for estimating the length of a typical data description session and the metadata produced. If only the sessions in which quality metadata were produced are considered (except the aforementioned outliers), which I will discuss further ahead, the minimum duration was 20 minutes (session 4) and the maximum was 30 minutes (session 2). Session 4 has the minimal amount of key-value pairs registered for a metadata record with quality, with a total of 11, while session 2 has the highest count with 17.

The average and median values could be slightly higher if the researchers were in a situation where they had more pressure to produce higher quality metadata, and would certainly be higher if all the metadata is of good quality. On the other hand, the different requirements per domain and the data typology itself may result in disparate values between domains.

11.2.1 Descriptors count

A total of 89 unique descriptors were filled in in the 13 sessions. Of these, 56 (more than half), occurred only once. On the other hand, 33 descriptors cooccurred, 16 of which only twice. The 10 most commonly used descriptors correspond to 38 per cent of the total, the last of them with 5 occurrences. It took only 15 descriptors to reach 50 per cent of the total number of descriptors used. The fifteenth most commonly used descriptor has only three occurrences.

Figure 27 shows the descriptor occurrence frequency. It only considers the 33 elements that were filled in more than once. The overall distribution of occurrences is skewed in a long-tail.

Three metadata elements were completed 8 times, in 61 per cent of the sessions, the *Abstract*, *Sampling Procedure* and the *Data Collection Methodology*. Only two other descriptors occurred in more than half of the sessions, namely the *Kind of Data* and the *Sample Size*. The frequency of use of these metadata elements across the sessions is not surprising given that they can be adapted to most research scenarios.

The Magnetic Dynamics researcher, although working in a very specific experimental setup showed interest in more generic metadata. However, what is surprising is the low frequency of several descriptors that are domain-agnostic and are important for any metadata record. For instance in session 2, 7 and 12 the researchers included both the *Creator* and the *Contributor*, while the *Creator* is also part of the metadata in session 1. This means that the metadata from the remaining 7 sessions does not include responsible party information. Likewise, the *Title* was only filled in in three sessions.

In the list of descriptors that were used more than once only 4 are not from the DC not the DDI vocabularies. These descriptors are the *Specimen Height*, the *Characterization Technique*, *Substrate* and the *Pulse Duration*. These were all filled in in the Magnetic Materials and Magnetic Dynamics sessions, except the *Specimen Height* which was registered in the Structural Adhesive Joints session and not in the Magnetic Dynamics session.

Therefore, DC metadata and elements from DDI were more frequently used, as Figure 28 depicts. Key-value pairs based on DC were created 62 times, corresponding to 34 per cent of all metadata, whereas DDI elements account for 66 occurrences (37 per cent of all metadata). The remaining de-

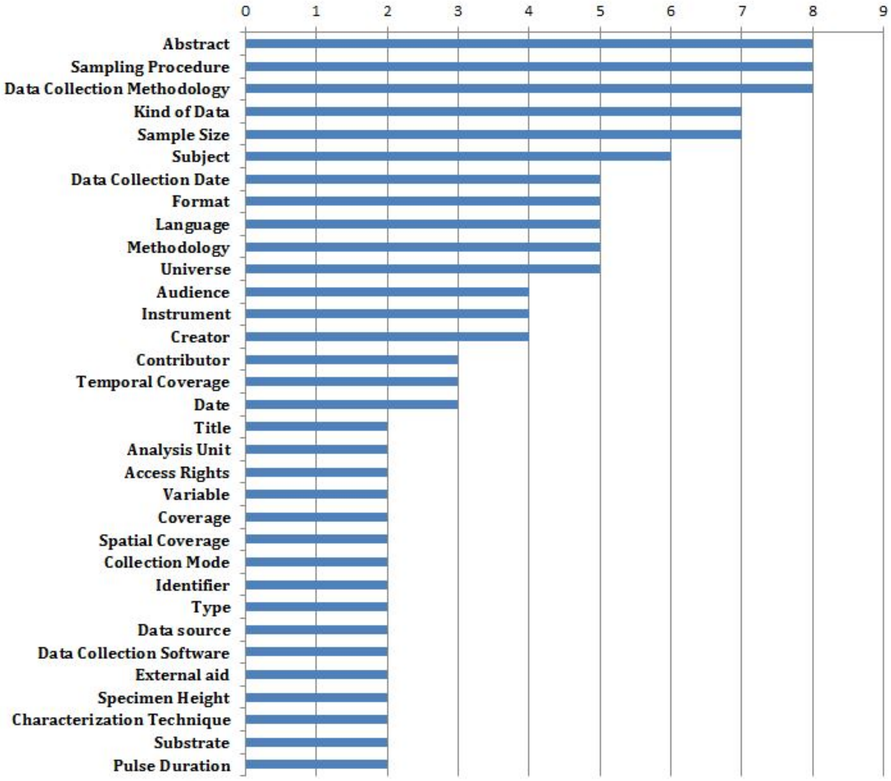


Figure 27: Descriptors occurrence

scriptors, those created with specific-domains in mind (see Chapter 6), make up for 28 per cent of total occurrences, with 50 key-value pairs completed. Together, DC and DDI account for 72 per cent of the metadata produced in the 13 sessions.

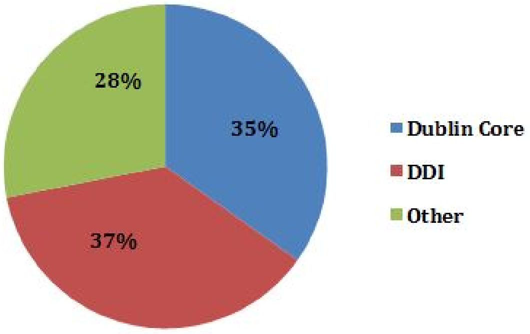


Figure 28: Frequency of vocabulary used in the 13 sessions

As for the contribution of the number of unique descriptors, showed in Figure 29, DC adds up 24 descriptors, DDI 19, and the other domain-specific ontologies 46 descriptors, thus, the latter make up for most of the descriptors in the long-tail.

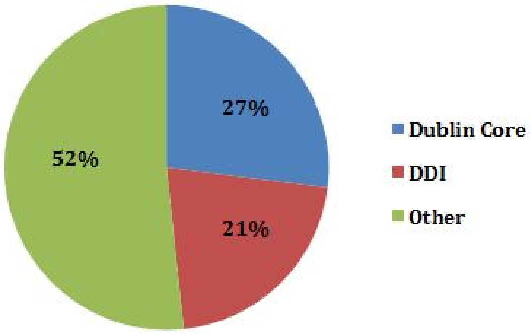


Figure 29: Distribution of unique descriptors used in the 13 sessions, by vocabulary

11.2.2 Frequency distribution of the metadata categories

Considering metadata categories (see section 9.2.1), Figure 30 shows that Context metadata was by far the most used category, with a total of 52 key-value pairs created, which is equivalent to 29 per cent of all metadata produced. This category was also the only that was used in every single session. This can be attributed to the fact that it is scientific metadata of a more general purpose. In other words, regardless of the scientific domain or research type, all datasets have an associated context, methodological processes, instruments and other aspects involved in their production and analysis. Thus, concepts such as the *Sampling Procedure* and *Methodology*, are not only meaningful, but also have a widespread application across domains, even when the terminology used in the metadata elements is not the preferred one to convey the same notion in a specific domain.

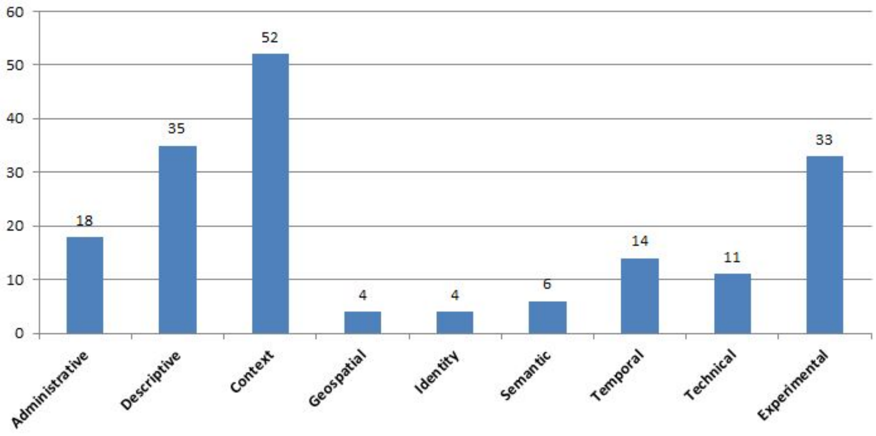


Figure 30: Distribution of the metadata categories

The Context metadata is followed by Descriptive metadata with 35 occurrences and by Experimental metadata, with 33 occurrences. The use of Experimental metadata is substantiated mostly by session 9, in which the Magnetic Materials researcher created 16 key-value pairs of this category. In the case of Descriptive metadata, it was frequently used in session 3 (8 times) and regularly distributed between sessions (3 to 4 occurrences). Yet, the researchers in session 9, session 10 and session 13, revealed no interest in this kind of descriptor.

In the Context category 21 different descriptors were used, 11 of them at least twice. Most of the Context descriptors are among the most frequently used, with prominence to the *Sampling Procedure* and the *Data Collection Methodology*. The Context descriptors that co-occurred are all from the DDI subset, except for the *Characterization Technique*, which is a descriptor more suitable in the physics domains and that was used by the researchers in the Magnetic Materials and in the Magnetic Dynamics sessions. The remaining 10 Context descriptors, those chosen only once, are rather fine-grained descriptors that could not be widely applied. This is the case of the *Absorbent Measurement Instrument* registered by the Sustainable Chemistry researcher and the *Sample Synthesis Instrument* by the Magnetic Materials researcher. The exception is the *Independent Dimension* used in Consumption Sociology, since it has the potential of wider used across the social sciences, and the *Method*, which is a very generic descriptor. The *Method* descriptor was used by the Magnetic Materials researcher to provide additional information to the *Characterization Technique* metadata. Moreover, the *Methodology* descriptor was widely chosen, making the usage of the *Method* redundant.

The Experimental metadata is represented by a set of descriptors with a fine granularity that relate to a very specific research context and are therefore concentrated in the long-tail of the descriptors count. Among the 33 occurrences in this category only 1 descriptor was used more than once, the *Substrate* in the Magnetic Materials and Magnetic Dynamics sessions.

Since Descriptive metadata provides general attributes for the data and related resources, a generalized use of this category was expected. A total of 15 different descriptors were used in this category, the *Abstract* and the *Kind of Data*, with 8 and 7 occurrences respectively, being the most popular descriptors in this category. Some descriptors in this category were picked only once, usually to relate to other resources, such as the *Bibliographic Citation* of the paper that resulted from the data, or the *Source* material on which the data was *Based on*.

Administrative metadata was regularly used with a total of 18 occurrences. This does not mean, however, that it was used in every session. This category was left unused in 5 sessions. The researcher in session 2, dedicated to Clinical Psychology, filled in 5 Administrative descriptors, the maximum for a single researcher. This researcher created *Access Rights*, *Audience*, *Creator*, *Contributor* and *Language* metadata. Only one other researcher defined the *Access Rights* to the data. The *Audience* descriptor appeared 4 times, which demonstrates a possible interest in data dissemination. The *Language* had 5 occurrences, and other most used Administrative descriptors were the *Collaborator* and the *Creator*.

Figure 31 represents the percentage distribution of the categories. Besides the expected highest incidence of the categories analyzed so far it is important to notice that Geospatial metadata, usually of scientific interest, have been underused, corresponding to 2 percent of the total metadata.

Considering the type of data described in the sessions the frequency of Geospatial metadata was expected to been higher, although in five of the sessions this information would not be essential for the metadata. Geospatial metadata was only used 4 times and repeated by the Clinical Psychology researcher with different descriptors to provide the same information. On the other hand, Temporal metadata was frequently used, 14 times in total.

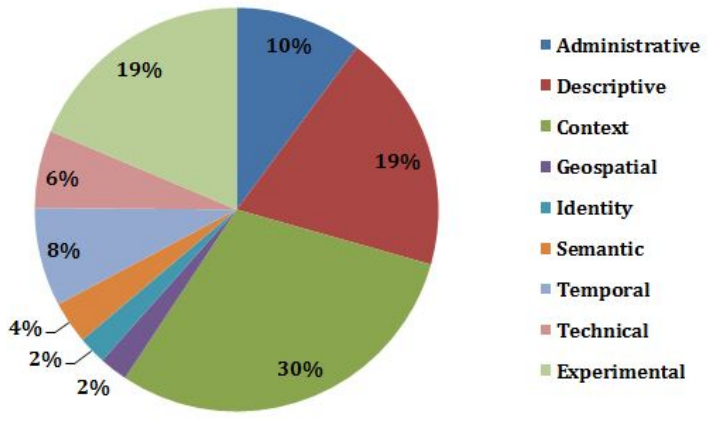


Figure 31: Metadata categories frequency percentage

In three sessions Temporal metadata would be pertinent but was not used. There were 3 occasions in which this kind of metadata was not properly applied. In these cases the researchers clearly understood the meaning of the *Data Collection Date*, but either did not remember the exact date of data collection or avoided the navigation of the available calendar in Dendro, and chose to record the date of the session instead, to exemplify their interest in this descriptor. In one case, both information about the study *Temporal Coverage* and *Data Collection Date* were provided. The Magnetic Materials researcher was the one that used the most temporal descriptors, with 3, to provide fine metadata about time parameters in relation to the material used in the experiments, such as the *Deposition Time*. The researchers from the Structural Adhesive Joints, Sustainable Chemistry and Magnetic Dynamics also provide the temporal parameters for their experiments.

The Technical category accounts for 6 percent of the metadata, with 11 key-value pairs completed. This category was absent in 4 sessions and was used to convey the dataset *Format* information by 5 researchers, although in one case the information it is not accurate, since the researcher registered that the format is the "doctoral thesis". The *Software* and *Data Collection Software* were also selected once, and the same happens with the *Summary Statistic Type*.

The Semantic category was also poorly used (4 percent) bearing in mind that it is a suitable category for all domains. A single descriptor was adapted to convey semantic information, the *Subject*, in 6 occasions. Among others, a *Keyword* descriptor was also available, but in general this category is under-represented compared to the others. Moreover, since Dendro displays the descriptors in an alphabetical order, it means that the *Subject* is almost at the bottom of the DC list and some researchers did not proceed through the list when they found that the first set of descriptors did not matter to them. The researchers that have chosen to fill in this metadata did it correctly to enter keywords. The Health researchers in session 12 entered a total of 10 keywords, others only the main topic or the type of experiment.

Finally, Identity metadata was used 4 times, 2 percent of the total, and was repeated by one researcher. The researcher in the Structural Adhesive Joints, in session 7, used the *Identifier* not only to identify the experiment code, but also ingeniously used the *Conforms To* field to relate the assay to a standard and thus redirect to more metadata in an external source. In session 3 the researcher included the *Identifier* of the project, while the Sustainable Chemistry researcher provided the *Sample Reference*.

Figure 32 depicts the distribution of metadata categories use by session. It shows that only the Context metadata was used in all of the 13 sessions. Descriptive metadata was used in 10 sessions with a greater contribution from session 3 with 23.5 percent of the 35 occurrences in this category. Temporal metadata was also recorded in 10 sessions, yet only once in 7 sessions, and 3 times in session 9. This category was not used in 3 sessions where temporal metadata could be adequate to describe the research project context, as the datasets in these sessions were collected over a period of time.

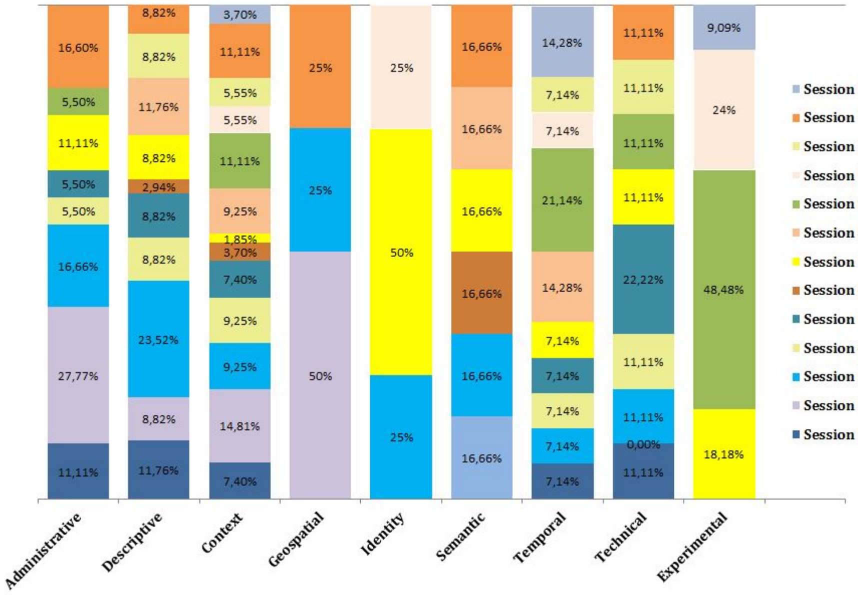


Figure 32: Distribution of metadata categories frequency by session

Technical metadata was recorded in 9 sessions and repeated in the Nutrition session. Administrative metadata was captured in 8 sessions, and the session that generated more metadata in this category was the dedicated to Clinical Psychology, with 5 occurrences, which produced 27,77 percent of the metadata recorded in this category (a total of 18 key-value pairs created).

The remaining categories were poorly distributed across the sessions, that is, they were used in less than half of the sessions. However, for some of the categories their absence in most sessions has greater implications for the quality of metadata produced. Like mentioned above, the Geospatial metadata has been poorly represented and it was only captured in 3 sessions, although in 4 of the sessions the data was produced in a controlled environment and no spatial metadata was required. This means that in 5 sessions were Geospatial metadata should be applied it was not.

Identity metadata was only used in 4 different sessions, but in this case there was no particular expectations regarding a generalized used, as this metadata is concentrated in a few specific descriptors and is only directly related to the context of data production sporadically.

The low distribution of Semantic metadata and Experimental metadata should be interpreted distinctly. Semantic descriptors were used in more sessions that the Experimental ones, 6 and 4 respectively. However, Semantic metadata can be applied in all domains, while the Experimental metadata was used in all the domains in which it was supposed to be applied. Therefore, despite used in a small number of sessions, the rate of use of the Experimental metadata is high. In session 9, related to Magnetic Materials, the researcher created 16 key-value pairs in this category (48.4 percent of all the Experimental metadata), the most a researcher produced for a single category. The researcher in session 13, Magnetic Materials, has only completed 3 Experimental fields, which is a very small number compared to the researcher in session 9, but these researchers working in similar domains had different approaches to the way metadata were created.

11.2.3 Metadata quality

The criteria for assessing the quality of the metadata produced in the sessions takes into account:

1. The number of descriptors filled in by the researchers;
2. A balanced distribution of metadata categories in the session;
3. The usage of descriptors that are expected to be included in the submission to a mainstream data repository;
4. Overall rigour in the information provided by the researcher.

The number of descriptors filled in by the researchers is considered from a qualitative point of view. The assessment does not follow any specific metric, such as the completeness of the record. Establishing a common metric to evaluate the metadata quality for all the metadata produced in this work would be impractical since the metadata requirements are very different from domain to domain. Setting a minimum number of descriptors per session to certify the completeness of each metadata record would also be not useful for assessing metadata quality across sessions with the same yardstick. To the best of my knowledge there is no recommendation to evaluate the completeness of metadata records across domains, or even for specific domains.

In order to assign quality to a metadata record with regard to the number of descriptors filled in, and taking into account the effort required from the researchers, I consider metadata records to have quality if they have at least 10 descriptors filled in. This is a vague and not very informative metric that has to combined with the other criteria.

The balanced distribution of metadata categories takes into account the diversity of descriptors from the different categories used in a particular session (Figure 33). This does not mean that all categories have an equal contribution to ensure the quality of the metadata, since it is not expected that a metadata record has the same number of Temporal, Identity and Context descriptors. However, if a record has 20 descriptors filled in, where 18 are Context metadata and the remaining are Descriptive it is probably not a balanced description and can hardly be considered high quality.

A more immediate comparison can be made with the metadata available in mainstream repositories (see Section 3.2). If the metadata captured in these sessions have the minimal information found in those repositories then the metadata can be considered of good quality. The analysis of data documentation on 5 different repositories [8] provides a baseline to verify

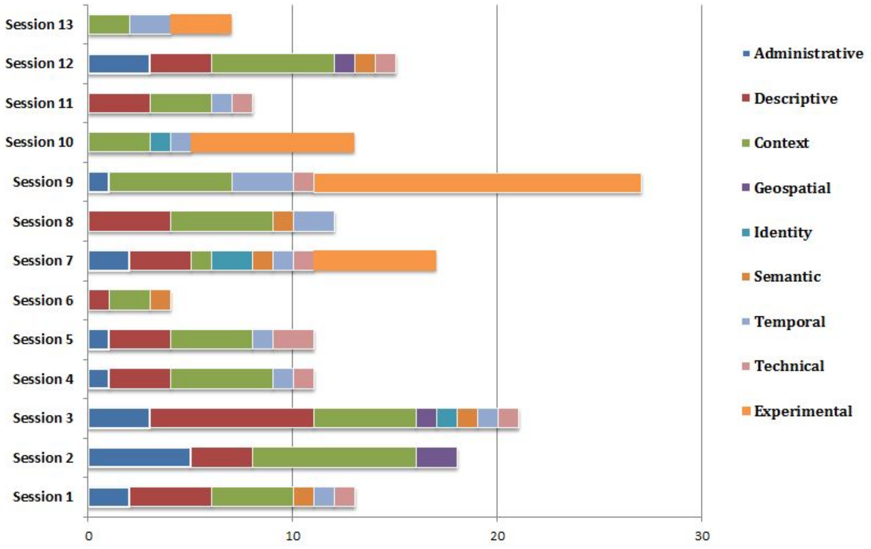


Figure 33: Use of metadata categories in each session

if the metadata created in these sessions would be fit for deposit in such repositories.

Finally, the metadata content, i.e. the value inserted by the researchers in a given field should also be considered. If the researcher enters values that are easily verified as unrealistic, like a future date to refer to the *Data Collection Date*, this has implications for the overall quality of the record. Moreover, it is also important to take into account the opinion of the researchers in the final questionnaire as to whether the metadata are sufficient and whether further information needs to be added.

Note that, I do not assess the accuracy of the metadata, first because I lack the necessary specialized knowledge in the diverse domains and secondly, I have not had enough detail about the experiences to validate the metadata produced. As such, a record that is deemed to be of good quality here may fail a more rigorous assessment, or be perceived as lacking in rigor by a domain expert in a possible reuse scenario.

Good metadata quality

With respect to four criteria 5 of the 13 metadata records are considered of high or very high quality. The metadata produced in session 1, session 2 and session 12, from the Family Psychology, Clinical Psychology and Health domains respectively, are rich and of good quality. The metadata produced by the researchers in session 3 (Consumption Sociology) and session 7 (Structural Adhesive Joints) have very high quality. Considering only these 5 sessions the researchers created an average of 17 descriptors in 35 minutes, although the average duration is overly influenced by the duration of session 3 (75 minutes).

The Structural Adhesive Joints researcher in session 7 has created 17 fields and resorted to descriptors from all categories, except for Geospatial metadata because it did not apply to the type of experiment conducted in his domain. The metadata record is very balanced and as shown in figure 33, has sufficient Experimental parameters information. Likewise, the Consumption Sociology researcher (session 3) provided 21 key-value pairs and only

Experimental metadata is absent since it was not appropriate for the data described. This researcher has provided detailed Descriptive metadata, as well as Context metadata. There are also three Administrative descriptors. In both cases the metadata far exceeds what is expect for deposit in a mainstream repository, except for the lack of a title (explicit in the file name) and access rights information in the Structural Adhesive Joints case. The values provided by both researchers are accurate.

The researchers in session 1, Family Psychology, produced a record with 13 descriptors, with 6 represented categories. The metadata, albeit concise, has several information to enable data to be findable and has sufficient information regarding the research methods. There is also metadata about the publication where the corresponding table is included. The record could be enriched with more detail about the *Temporal Coverage* and the *Format* information is not accurate. Nevertheless, apart from the *Access Rights* the metadata record complies with the deposit requirements of several repositories.

In session 2, Clinical Psychology, the researcher filled in 18 fields with a combination of 5 Administrative descriptors, three from the Descriptive category and 8 for the Context, in what can be seen as a comprehensive and detailed metadata record. It complies with the main requirements of data repositories. The downside is that some information is not precise but would easily be adjusted in case of publication. Yet, Semantic and Temporal metadata are missing. The two Geospatial descriptors were used to convey the same information.

The Health researchers created 16 key-value pairs and used 6 different categories, with prominence to Context metadata. The values fit the selected descriptors and are realistic. The information provided makes it easy to enable search, for instance by including 10 subjects. However, there is no Temporal metadata and the information is in Portuguese. This metadata record was produced in 20 minutes.

All the researchers that created good quality metadata believe that the metadata is sufficient and that there is no need for more information, except in the Family Psychology case, where the researcher thinks that maybe more information can be provided.

Except for the Structural Adhesive Joint researcher, who showed great metadata awareness and skills working in Dendro, all the other researchers had no prior experience with metadata. Between the interview and the description session, the Consumption Sociology researcher sought to know more about metadata, which is reflected in the high quality of the metadata.

Satisfactory metadata quality

In 4 sessions the metadata record produced cannot be assessed as having good to high quality but, considering that researchers had little experience with metadata, the resulting metadata is satisfactory and has the potential to become good quality with minor changes. These satisfactory records were produced in the Services, Organizations Sociology, Magnetic Materials and Sustainable Chemistry sessions.

In the first two cases the researchers created 11 key-value pairs (Services) and 12 key-value pairs (Organizations Sociology). The Organizations Sociology researcher, in session 8, resorts to 4 categories, namely Temporal, Descriptive, Context and Semantic. The quality of metadata in this session is limited by the lack of responsible party and spatial coverage information and the *Subject* metadata is mostly generic and does not promote search-ability. The same is true for the metadata in session 5, where the data producers are absent and the spatial information was recorded as part of the study *Universe* description. Also, there is no *Subject* or *Keyword* information. Both records have sufficient methodological information which is an important factor for scientific metadata and such information is not usually displayed in several generalist data repositories.

As for the metadata produced in the Sustainable Chemistry and Magnetic Materials it contains several elements to inform about the experimental set up conducted by both researchers. The researcher in the Sustainable Chemistry session (a remote session) filled in 13 descriptors from 5 different categories, of which 8 are Experimental and 3 are for Context metadata. Therefore the metadata is purely scientific in nature and lacks Descriptive, Technical, Semantic and Administrative elements. The values entered are precise. The Magnetic Materials researcher completed 28 elements, also using 5 categories, 16 of which are Experimental metadata and 6 are Context metadata. This metadata record includes Administrative metadata, yet it lacks Descriptive elements. Some of the information provided is not rigorous, which does not contribute to the quality of the metadata. In both cases the metadata regarding the experiments and data context of production is very rich and the number of descriptors used, particularly in the Magnetic Materials session, seems more than sufficient, but the records do not have the minimal information expected in repositories. Nevertheless, if the data went through a deposit process much of this missing information, in some cases mandatory, could be added without much effort.

All the researchers that produced satisfactory metadata agree that more information can be provided, although the Sustainable Chemistry and the Magnetic Materials researchers find that the information is already sufficient. The Organizations Sociology researcher thinks that the metadata is not enough and the Services researcher thinks that the metadata may suffice. This feedback is consistent with the assessment made of the quality of the metadata.

Poor metadata quality

The metadata produced in 4 sessions has not been assessed as being of satisfactory quality. Still, the quality of the metadata in these sessions has in some cases the potential to evolve into a higher quality record.

In session 13 for instance, the Magnetic Dynamics researcher only filled in 7 descriptors, of the Experimental, Context and Temporal categories, but the feedback given throughout the session made it clear that if the descriptors were displayed in a different fashion the description would be richer. As such, the researcher showed great awareness and interest in the task despite the final metadata showing otherwise. In the follow-up questionnaire this researcher noticed that the information was not sufficient and that more information was needed. Accordingly, the Magnetic Dynamics researcher left a comment with an example of possible metadata fields missing.

The Cultural Studies researcher created 8 key-value pairs in only 16 minutes. As mentioned earlier this researcher participated in this activity among a busy agenda. The metadata record, however short, is balanced and has 4 different categories, with Descriptive and Context metadata being the most used, with 3 occurrences. The Descriptive metadata includes information related to provenance and related documents which is interesting for end-users, yet it lacks minimal information like the people involved, a short description or a title. Hence, it does not conform to metadata deposited in generalist repositories. The content is mostly correct. The researcher also believes that the metadata is sufficient and that there is no need for more information.

The researcher from the Nutrition domain, in session 5, created a similar record to the one created by the Cultural Studies researcher. It used more descriptors, 11, and 5 categories, though information like the *Abstract* is completely random and does not add value to the metadata. There are sufficient elements to cover the study methods, however the content is not detailed, and some of the information would be adequate within the *Subject*. Overall, with slight changes this metadata record could be ready for deposit, provided that a few descriptors are added and the description is more rigorous. However, the researcher has the opinion that the metadata is sufficient and that there is no need to create more metadata.

Finally the Work Psychology researcher, the less experienced researcher among the participants, produced a metadata record with only 4 key-value pairs, 2 for Context, and 1 descriptor for Semantic and Descriptive metadata. The *Subject* and *Kind of Data* information would provide good access points to the dataset, but the record lacks contextual information and there is no metadata to match the minimal information required by data repositories. Nevertheless the researcher thinks that the record is sufficient and that there is no need for more information.

11.3 RESEARCHERS' FEEDBACK

11.3.1 Data description usefulness

Overall the researchers have classified the data description activity as useful. As shown in figure 34, except for one, all have classified the degree of usefulness above 5, and the majority (9 out of 13) rated it with 6 or 7.

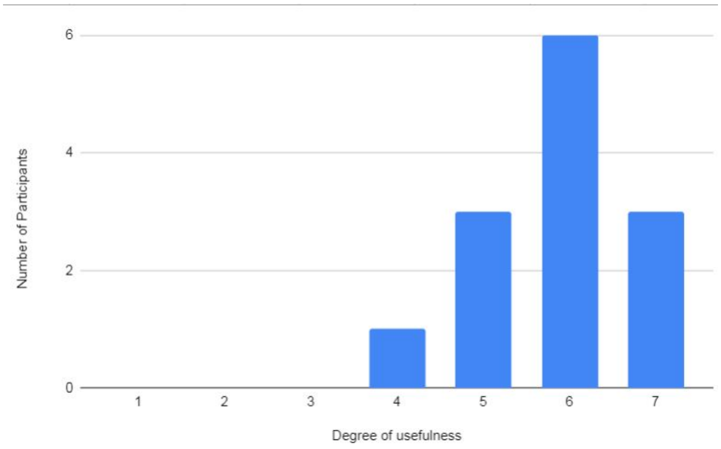


Figure 34: Degree of data description usefulness

The degree of usefulness was rated very high by the researchers from the Magnetic Materials, Health and Magnetic Dynamics domains. The metadata created by the Magnetic Materials and Health is of satisfactory and good quality respectively, while the metadata by the Magnetic Dynamics was considered of poor quality, although the researcher from the latter showed great awareness and provided valuable feedback on how to improve its own experience in the data description session. For the Magnetic Dynamics researcher data description is useful for a quick identification of the measured data for analysis purposes, as well as to provide easy access to measured parameters. The researcher from the Magnetic Materials is from a close domain and has a similar opinion. For this researcher the way data is obtained can change some properties or lead to errors, so it is very important to know specific experimental parameters. The Health researcher that answered the follow-up questionnaire seems to relate the usefulness of data description more with data organization, since it helps to systematize everything correctly and simply.

The researchers from Family Psychology, Consumption Sociology, Organizations Sociology, Services, Structural Adhesive Joints and Cultural Studies consider data description a highly useful activity. From these, the researchers from the Family Psychology, Consumption Sociology and Structural Adhesive Joints have produced good quality metadata. The Family Psychology researcher has the opinion that data description is useful to facilitate the dissemination of data to other researchers, because it allows the use of existing databases and thus does not overburden participants with new questionnaires. The Consumption Sociology researcher also has data reuse in mind when considering the usefulness of data description, since it not only allows data to be stored with all relevant additional information, but it also favours reuse. For the Structural Adhesive Joints researcher there is a need to carefully describe all the assay conditions in order to be able to replicate results or to understand unexpected results. As for the Services and Organizations Sociology researchers that created satisfactory quality metadata records, the first argues that data description is useful given that data reliability and access to data are critical to the quality of research, and the second thinks that the usefulness of description is related to organization purposes and rigor in the production, management and use of data. Finally, the Cultural Studies researcher who produced a poor quality metadata record wrote that data description allows one to conceptualize and structure data collection more efficiently.

The researchers that rated data description as useful are those from the Work Psychology, Sustainable Chemistry and the Nutrition domains. The metadata that resulted from the Sustainable Chemistry domain was of satisfactory quality, and the researcher has the opinion that metadata is usefull for data management and information archiving. The other cases produced metadata of less than satisfactory quality. For the Work Psychology researcher data description helps to save data with more details, while the Nutrition researcher stated that it allows for a better organization of information and process control.

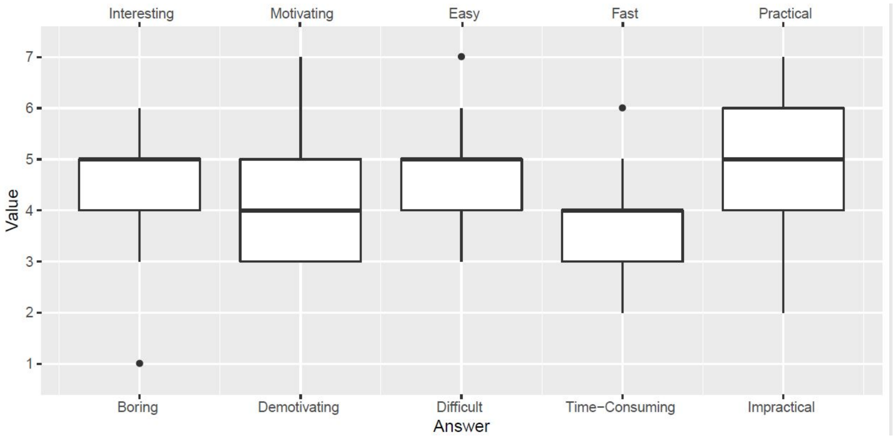
The Clinical Psychology researcher deemed data description as moderately useful. Despite having created a metadata record of good quality this researcher has not fully understood the application of the knowledge that resulted from the description of the data. For this researcher it is an important subject but still a somewhat an abstract activity.

11.3.2 Data description activity assessed by researchers

Taking into account the feedback provided by the researchers in the follow-up questionnaire, data description, as performed in these sessions, can be outlined as somewhat interesting, a little discouraging, slightly time-consuming, yet moderately easy and moderately practical.

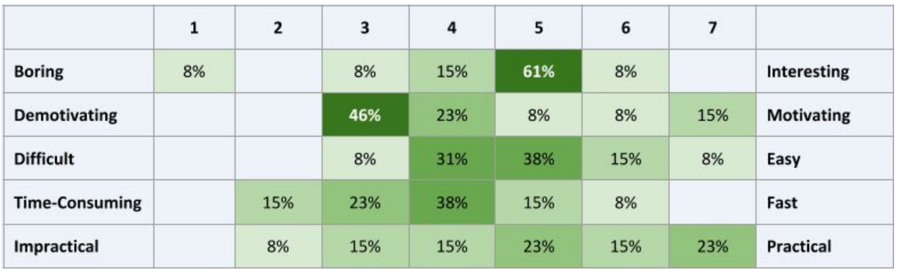
The distribution of data in table 19 shows that the median, for all the two bipolar adjectives featured in the semantic differential scale (see Section 9.3.1), is either 4 or 5. For none of the characteristics associated with data description can it be said that the researchers´attitude is extremely positive, but in none of the cases is the median below the neutrality point. Moreover, the lower quartile for the motivational factor and duration is at 3, so it is slightly tending to the negative pole.

Table 19: Researchers assessment of the data description activity



On the other hand, the percentage agreement displayed in table 20 shows that a higher number of researchers, 6 (46 percent), thinks that data description is a little discouraging, compared to the 31 percent who think otherwise. Moreover, 5 researchers are neutral regarding the duration of this activity, yet the number of those in the negative pole is higher than those in the positive pole, 5 versus 3, 38 per cent and 23 percent respectively.

Table 20: Semantic differential researchers agreement percentage



	1	2	3	4	5	6	7	
Boring	8%		8%	15%	61%	8%		**Interesting**
Demotivating			46%	23%	8%	8%	15%	**Motivating**
Difficult			8%	31%	38%	15%	8%	**Easy**
Time-Consuming		15%	23%	38%	15%	8%		**Fast**
Impractical		8%	15%	15%	23%	15%	23%	**Practical**

Three researchers had an attitude towards the description of data that tended to be more negative, namely the Clinical Psychology, Work Psychology and Magnetic Dynamics researchers. The first has produced a good quality metadata record but while very sure of the activity objectives, considered it a very motivating one. The other two created poor quality metadata records, so their negative attitude towards data description makes sense in relation to the metadata they have produced.

Three others have the opposite attitude and hold a more positive perception of the activity. The perception of the Sustainable Chemistry tends towards the positive end of the scale in every feature, except for the activity duration, although the time dedicated to data description was short (10 minutes). Also, this researcher does not think that the activity is particularly motivating. The Magnetic Materials also had a positive experience, according to the feedback. This researcher was only neutral in relation to the activity duration. These researchers which tend to make a positive assessment of the data description activity have produced metadata of satisfactory quality, and of good quality in the case of the Structural Adhesive Joints researcher.

Next, I will do a more detailed analysis for each semantic differential feature assessed by the researchers.

Stimulation

The levels of enthusiasm of the participants in this study towards the data description activity was not high but it can be said that researchers lean more into considering it interesting rather than boring. The median for the stimulation feature stands at 5, while the lower quartile is at 4. Only two researchers consider the activity boring, the researcher from the Magnetic Dynamics who did not think that the descriptors were displayed according to own expectations, and the researcher from the Clinical Psychology who assessed data description as a very boring activity.

Most researchers gave a score of 5 to this activity, and the Sustainable Chemistry researcher gave it a score of 6. It is important to notice that three researchers who consider data description as somewhat interesting did not create a record of satisfactory quality. In particular, the Work Psychology researcher's feedback needs to be contextualized, since this researcher only created 4 key-value pairs and, therefore, the experience was different from what is expected in a real data description scenario. The researchers that assess the activity as neither boring, nor interesting, are from the Health and Family Psychology domains, and these have created good quality metadata.

Motivation

The results regarding the motivational level of data description show that this activity is a little more discouraging than motivating, although few researchers assessed it with a very positive score. There was agreement of 6 researchers (46 percent) as to whether the description of data is a slightly demotivating activity. Three others have a neutral attitude.

For this feature there is no relation between the quality of the metadata record and a positive evaluation. There are researchers that produced satisfactory and good metadata records who evaluate the activity as a little demotivating. However, 3 of the 4 researchers who have created poor metadata also classified data description as a little demotivating activity. The exception was the Cultural Studies researcher who classified it with a 5. The Sustainable Chemistry and Clinical Psychology researchers found the description of data extremely motivating. Still, the latter had a general negative perspective of the overall experience and therefore, it is possible that this was a lapse in the evaluation.

Difficulty

With respect to the difficulty of data description, the researchers' attitude is consistent with the assessment that this was a moderately easy experience for them. The median stands at 5, while the lower quartile is at 4. Only the Clinical Psychology researcher found the activity lightly difficult.

Considering the researchers that had produced poor quality metadata records, since it can have a direct relationship to the difficulties researchers may have felt during their sessions, 3 of them were neutral in the evaluation of this feature, while the Nutrition researcher assessed the difficulty of data description as moderately easy. This makes sense since this researcher did not feel any challenge that would prevent the production of a better quality metadata record. The researchers from the Structural Adhesive Joints and Magnetic Materials considered data description an easy activity and for the Sustainable Chemistry researcher it was a very easy one.

Duration

The duration of the session was taken as time-consuming by the Materials Dynamics researcher and as somewhat time-consuming by the Family Psychology, Organizations Sociology and Consumption Sociology researchers. The latter had a session far beyond the duration of the others. For 5 researchers data description was neither time-consuming, nor fast, even for the Sustainable Chemistry and Work Psychology researchers, which sessions lasted 10 and 15 minutes respectively.

A general observation is that those who have a session that lasted for more than 30 minutes assessed the activity as time-consuming, while those who found it slightly fast, or fast, had a session duration below 20 minutes. The researchers from the Structural Adhesive Joints, Health and Nutrition domains had a more positive perception of the session duration.

Practicality

Finally, the median for data description practicality is 5. Moreover, the lower quartile is 4 and the upper quartile is 6. A total of 8 researchers have a positive perception of the practicality of data description. In contrast, 3 of them perceived the activity as impractical. Only the Health and Material Dynamics researchers were neutral.

The researchers who found the activity impractical also had a demotivating experience. This was the case of the Work Psychology and Nutrition researchers, who produced poor metadata quality records. The Services researcher considered the activity a little impractical. These 3 researchers had a general experience tending towards moderation, or tending towards somewhat positive (Nutrition), taking into account all the semantic differential features. On the other hand, those who highly assessed data description as very practical, had a general attitude also tending towards the positive pole. This was the case of the Structural Adhesive Joints, who produced a record of good quality, of the Magnetic Materials and of the Sustainable Chemistry researchers. The researchers from the Clinical Psychology and Magnetic Dynamics were neutral and they both have a general negative perception of the activity.

Part V

Conclusions

12 DISCUSSION

As research policies lead institutions and researchers to adopt RDM practices [41, 36], metadata activities are becoming embedded in the research routines. One reward of investing in metadata production is that it favors data reuse [115], which may promote data citation and in turn lead to more data being deposited and reused [83]. As long as there are clear incentives and adequate tools, researchers are likely to engage more and more in RDM tasks [114]. Researchers are domain experts and data producers, therefore they are well positioned to be key metadata producers as well [122], particularly considering the generalized lack of staff with the expertise in institutions. Taking into account the results from the survey on metadata-driven studies in Chapter 5 and the metadata practices of the participants in the presented study (Section 10.6) it can be said that metadata creation is something that they already do, although mostly supported in *ad-hoc* and *personal practices*. On the other hand, scientific-oriented metadata is often supported by complex standards that researchers may struggle to adopt [89].

The main purpose of this work was the definition of a data curator´s workflow focused on the development of domain-specific metadata models, by engaging researchers in the process, as detailed in Chapter 6. The development of these domain-specific metadata models, by capturing familiar concepts to reduce possible adoption barriers, is an essential task to engage researchers in RDM through metadata creation.

Many RDM activities stem from the creation of metadata. Among other aspects, metadata are an enabler of data sharing and reuse. If tackled conveniently, the production of metadata during the data life-cycle will address several RDM challenges at once. Therefore, I believe that engaging researchers in metadata creation is a practical way to introduce them to RDM. I carried out data description sessions with 13 researchers from a diversity of domains (see Chapter 11) to evaluate whether the proposed workflow enabled researchers to create quality metadata records, and whether the researchers held a general positive attitude towards data description once the sessions were complete.

I devised two main research questions in order to ascertain the potential of the data curator's workflow:

Do the metadata models available in Dendro enable researchers to create quality metadata?

How do researchers assess the data description activity, taking into account the collaborative process between researcher and curator and the domain-specific metadata available?

The combination of satisfactory answers to both questions means that this data curator´s workflow can be regarded as a promising approach to engage researchers in RDM, particularly in metadata creation.

12.1 QUALITY OF THE METADATA PRODUCED BY THE RESEARCHERS

The quality of the metadata records created by researchers was compared against four criteria that I have established (Section 11.2.3), namely the number of descriptors used, the metadata categories of the descriptors (Section 9.2.1), the usage of descriptors that are expected to be included during submission to mainstream repositories (Section 3.2), and the overall accuracy of the metadata records created. The metadata created in each session was classified as either poor, satisfactory or good.

The results showed that the metadata quality was satisfactory or good in 69 percent of the sessions. One of the metadata records considered poor was produced in the Magnetic Dynamics session. In this case the way the descriptors were displayed in Dendro were not in line with expectations of the researcher, who wanted a more structured presentation of the descriptors.

Overall the researchers were prompt to use Context and Experimental descriptors when needed. This reinforces the importance of having available descriptors to fit domain metadata requirements. On the other hand some metadata records lack Descriptive and Administrative metadata, as well as metadata from other categories.

From the 9 researchers that produced metadata records with a greater scope for improvement, i.e. those who have created metadata considered poor or satisfactory, only 3 find that the information is not or may not be sufficient. In contrast, only 3 think that there is no need for more metadata. The latter are researchers from the 3 sessions that created the poorest metadata. This suggests that for these researchers metadata awareness was still low after the description session, but this may also be due to a lack of interest in data description in general.

The general quality of the metadata records resulting from the 13 session has plenty of room for improvement. However, since the quality of 9 metadata records was considered satisfactory (4), and good (5), it can be cautiously concluded that the metadata models available in Dendro enabled researchers to create quality metadata. Particularly taking into account that the previous metadata experience of the researchers was low (Section 10.2) and it was the first experience creating metadata, in a platform like Dendro, for all of them.

12.2 RESEARCHERS ATTITUDES TOWARDS DATA DESCRIPTION

To verify whether data description was a positive experience for the researchers I probed their attitude by applying a data description follow-up questionnaire, detailed in Section 9.3. Researchers were asked to characterize data description through 5 pairs of adjectives, capturing their feelings regarding; stimulation, motivation, difficulty, duration and practicality. Moreover, I wanted to know their perceived degree of data description usefulness.

There is no support to assert data description as a stimulating activity for researchers (Section 11.3.2), but the results lead me to conclude that the description of data was considered a somewhat interesting activity. In this feature the median stood at 5 and only two researchers considered the activity boring.

Although not totally conclusive, from the gathered feedback it can be inferred that data description tends to be a slightly demotivating activity (Section 11.3.2), as there is agreement of 46 percent of the researchers in this matter, while 3 others remained neutral.

As for session duration, considering the distribution of data and that the median was 4 (Section 11.3.2), it can be concluded that researchers did not have a positive perception of the session duration and their feedback leads towards the time-consuming pole. Hence, data description can be defined as slightly time-consuming. Sessions that lasted for more than 30 minutes were considered time-consuming.

Overall, as there is only one negative opinion on the level of difficulty (Section 11.3.2), some neutral researchers, while some considered data description as moderately easy, it can be concluded with fair confidence that researchers have not perceived data description as a challenging activity. Since most researchers found data description a practical activity and the median for this feature was 5, data description was assessed as moderately practical (Section 11.3.2).

Regarding the researchers´ experience in creating metadata in Dendro, data description can be characterized as somewhat interesting, slightly demotivating, slightly time-consuming, moderately easy and moderately practical activity.

With respect to the perceived degree of data description usefulness, assessed in Section 11.3.1, as a whole the results suggested that those who created better metadata records tend to highly rate the usefulness of data description. Moreover, those who most valued the data description usefulness associated the benefits of metadata creation to data reuse and to the quality of the data itself. The remaining researchers mentioned benefits associated with data storage, management and organization. The lowest score for the degree of usefulness of data description was assigned by the Clinical Psychology researcher, with a 4, despite having created a good quality metadata record.

Given that data description was generally considered to be a useful activity, with the downside of being slightly time-consuming and slightly demotivating, it can be reasonably assumed that the description of data, as conceived in this study, is a realistic activity for researchers to perform.

All summed up, the proposed data curator´s workflow (Chapter 6) can be assessed as a promising approach to engage researchers in data description.

This data curator's workflow encompasses interviews with researchers (Sections 6.2), the development of domain-specific metadata models, formalized as lightweight ontologies (Section 6.3), and the involvement of researchers in data description sessions (Section 9.2). To streamline the communication with researchers, content analysis over domain publication was also recommended as an additional task, if needed, in this data curator´s workflow. This approach was detailed and evaluated in Chapter 7.

12.3 INTERVIEW

The final interview script, applied in the data description sessions, although rooted in the Data Curation Profile Toolkit, was refined as I gained more experience in interviewing researchers. Overtime, I encountered some limitations in the adoption of an overly structured interview form and I edited the script in order to fit my goals. When asked about their feelings on how the interview went, researchers noted some pros and cons, and also made some recommendations to improve it.

The silver lining of the interview, according to some, was that it helped to think about RDM issues never considered before. One researcher acknowledged that if the interview has taken place before the start of a recent project perhaps it would influence the way the data was organized, from a personal to a more sharing-oriented perspective. Moreover, the general thought was that the interview has a good balance between its duration and the number of questions. Sending the interview beforehand was also valued by some. The average duration of the interview was approximately 40 minutes.

On the other hand, it was pertinently observed that some questions can be very specific, leading to "yes" or "no" answers. Moreover, one researcher admitted not knowing if the answers corresponded to my expectations. After this comment, I tried to explain the following participants that I was not expecting any particular outcome from the interview, other than knowing more about their data and their practices. Some issues regarding the interview terminology were also pointed out, especially what was meant by "metadata". The use of some concepts that I assumed that could generate some doubt was, in some cases, premeditated. By asking researchers if their data was accompanied by metadata, their answers would show their understanding of the concept. Nevertheless I provided several examples during the interview to address any doubt.

One very interesting recommendation was to adapt the interview script to the domain context. This would mean having prior knowledge about the domains. However, it is likely that the script can be adapted to high-level domains. To make the most out of the interview it may also be useful to have a glossary for the most specific concepts, one researcher suggested.

The final version of the script is, I believe, a useful resource for data curators who want to gain an understanding of the domain and the metadata requirements researchers may have.

12.4 DEVELOPMENT OF LIGHTWEIGHT ONTOLOGIES

Throughout this work there was the opportunity to design lightweight ontologies for several domains, either from scratch or starting from existing standards, to establish a subset that reduces the complexity of the standards to a level more suitable to researchers. The design process has been discussed and instantiated in Section 6.3.2.

These lightweight ontologies do not in any way represent the domains in their entirety or aspire to a standard statute. Instead, they can be regarded as an operational tool to streamline the data curator's workflow by representing concepts tailored to the needs of the researchers with whom I, or other TAIL members, collaborated. Whenever possible, concepts from existent standards were reused, if already modeled on other ontologies.

The development of the domain-specific vocabularies was rewarded by the general quality of the metadata produced by the researchers in the data description sessions. The effort and need to develop, or extend, new ontologies has decreased as I have included new participants in the study.

The lightweight ontologies process described in [22], was taken as a pertinent topic and an interesting case study for the digital libraries community.

12.5 DATA DESCRIPTION SESSIONS

The data description sessions (Section 9.2) resulted in a very practical activity to engage researchers in RDM and raise their awareness.

The data description sessions are also convenient to gather insight on the choices researchers tend to make in terms of descriptor selection and the amount of time spent in this activity. These are tangible indicators that can inform future decision making in the development of services and tools to support researchers in daily RDM activities.

The results suggest that a basic set of metadata, of a more general nature, but with scientific objectives, can be devised with the combination of DC and DDI elements, and still be suitable for a diversity of domains, even to those disciplines closer to the natural sciences.

The descriptor count results in Section 11.2.1, provided the insight that a small number of descriptors cover basic metadata requirements in the domains represented in the sessions. However, the diversity of domains and the long-tail of unique descriptors used make it necessary to provide researchers with fine and domain-specific elements to enable them to provide metadata. To address the "black-canvas effect" [129], where researchers do not know where to begin the description, it would be usefull for tools like Dendro to provide users with a common set of domain-agnostic descriptors as a first step in metadata creation.

The results regarding the metadata categories in Section 11.2.2 indicate that some categories are underrepresented and that their use should be promoted among the researchers. The most obvious case is that of the Geospatial metadata. On the other hand, the distribution of Context metadata and the use of Experimental metadata in all cases where it applies shows the readiness of researchers to fill in descriptors that are targeted to the scientific context of data production. It seems that researchers are available to create this kind of metadata, which demonstrates the importance of having descriptors that fit the scientific context of each researcher.

12.6 CONTENT ANALYSIS IN THE DATA CURATOR'S WORKFLOW

As I gained experience through increasing contacts with researchers I became aware that my communication with researchers was at times hampered both by my lack of domain expertise and their difficulty to master concepts such as metadata. I felt the need to streamline the data curator's workflow by proposing manual content analysis of domain publications. As elaborated in Chapter 7 the content analysis approach has the potential to improve the communication between RDM stakeholders and build metadata models.

Comments provided in the peer-review of publications generated by this work recognized that domain publications certainly contain insight into comprehensive metadata possibilities. However, this proposal is not consensual since it entails an additional task to the data curator. It was remarked that the efforts undertaken by the curator may be better spent in other activities with the researchers. My point of view is that this additional task is more of a timely activity, which can make the most out of the other relevant activities with the researchers. I have experienced initial frustration in some contacts precisely because I was lacking sufficient domain knowledge.

12.7 LEARNING THROUGH COLLABORATION

The actions undertaken in this workflow can be seen as a learning process not only for researchers but also for the data curator. Several tasks were carried out by MSc students, since I found it relevant to verify whether the different activities would be successfully carried out by less experienced "curators".

For instance, the manual content analysis was carried out in the context of an MSc. thesis in information science [61]. As I closely followed the progression of this work I noticed that, after a slow start, the effort to achieve results diminished over time, as the skills of the curator improved, after the selection of the concepts from the first set of analyzed publications. I also had the opportunity to follow the evaluation sessions with the researchers (Section 7.3) and noticed good dynamics between the curator and the researchers.

Likewise, the work carried out in order to train researchers in metadata creation through the use of a subset of the MIBBI standard (Chapter 8), involved mapping the availability of biomedical repositories and standards, and was performed by another student. The outputs were then useful during the conversations with the researchers that tested the proposed metadata model [102].

In line with the RDA Libraries for Research Interest Group[1], I agree that one of the most impactful ways to engage researchers is to create awareness about the need for good RDM, and that direct training, although requiring substantial time and effort, is one of the most effective ways to make people aware of the importance of RDM practices [30]. However, I think that it is also necessary to train future curators to be sensitive to the challenges that researchers face. In that sense I envision the proposed data curator's worflow as a contribution to train both researchers and data curators through collaboration.

[1] https://www.rd-alliance.org/groups/libraries-research-data.html

13 FURTHER DIRECTIONS

The usefulness of thinking about RDM as a "wicked problem" was proposed by information practitioners as a way to address RDM issues [9]. Participants in a two day workshop identified features of wicked problems they felt more relevant to RDM. Among others, they agreed that RDM problems are complex and there are multiple possible intervention points, along with a great resistance to change. During this thesis work I faced these challenges, which led me to the adoption of a set of research techniques and a flexible conduct to interact with the researchers.

Setting up the multi-domain data description sessions to accommodate the metadata requirements and researchers expectations from several domains has proven to be a demanding task. As with any other study, there are design limitations which need to be addressed to improve further research and foster the participation of researchers in RDM.

Limitations concerning the sample size have been briefly discussed in Section 10.1.1. The sample size in this work does not fall short in comparison to studies with a similar design, yet more participants would be needed to draw more general conclusions. It would be interesting to open the study to other experts from the domains represented in this study, in particular for the purpose of assessing the quality of the metadata produced. Although this hypothesis has been considered, the recruitment process turned out to be a laborious task, thus my priority was to make sure I could finish all the activities proposed to the initial set of participants. During my participation in scientific meetings, curiosity about the recruitment process was common with colleagues. The general belief that current RDM is already good enough may prevent several researchers from participating in this type of study. Future work must take this into account, and I believe that it is essential to involve more researchers from the same domain to collaborate in the activities. The data description sessions in the Family Psychology and Health were successful, in part, also due to the exchange of ideas between the pair of researchers that participated in both sessions.

Another important aspect to consider is that the conclusions regarding the generation of metadata are tightly related to the usage of Dendro (Section 3.3). It could be speculated on where Dendro might have contributed to the general attitude of researchers to data description. It is possible that the attitude of researchers reflects more their experience of describing data in Dendro, than on the actual description of data. An interesting study would be the assessment of researchers' attitudes towards data description performed over different platforms. However, a more practical perspective is to consider whether the attitude of researchers and the overall quality of metadata could be improved. As already suggested, it would be useful for the researcher to have a default list of general descriptors presented to them as soon as they start the description of data, as in most cases they do not know where to begin. On the other hand the availability of a high number of descriptors entails more time spent in browsing through a list of vocabularies whose designation may not be familiar to many. Browsing descriptors according to the type of data or metadata categories would make it easier for researchers to select suitable descriptors. For instance, if the researchers are interested in describing experimental data they should be able to find the descriptors on a list previously curated to satisfy this need. The adoption of controlled vocabularies would also have the potential to shorten the duration of data description and improve the accuracy of the metadata records. Controlled vocabularies were not exploited in this work, although this approach has already been tested in activities related to the TAIL project [59]. More work has to be done to integrate controlled vocabularies in an extended version of the domain-specific metadata models.

The data description sessions showed that the definition of metadata models must balance generic and specific descriptors. Highly-specific descriptors were filled in by the researchers very quickly, while the absence of some specific descriptors apparently limits the metadata. The description provided by some researchers showed that fine-grained descriptors can make metadata production smoother. Therefore, a higher degree of specificity is desirable when metadata tools or platforms allow the combination of metadata elements with respect to the diversity of experiments, techniques and datasets a researcher may need to describe. My work with the researchers led me to conclude that a scenario where researchers do not have suitable descriptors should be avoided.

As long as researchers are provided with adequate tools and have clear, practical examples of metadata goals, data description can become an intuitive task for them. At the time of writing, researchers mostly captured metadata in an informal fashion, so the priority is to encourage them to adopt tools to make metadata creation more systematic. After that, the focus can shift to the next level, namely with actions to improve the overall quality of metadata.

By the end of the engagement activities the researchers have shown interest in RDM. For most of the researchers (7 out of 13) this interest is still moderate, while the remaining have a high degree of interest in RDM. Among the valued aspects for the interest in RDM are the systematization of organization practices to avoid forgetting important details, the fact that it brings fundamental dynamics for the rigor of the research and also that it enables new perspectives on data. For one of the researchers, a greater interest is dependent on the use of tools that are easy to implement on a day-to-day basis and not an additional work and time spent filling in data to be recorded. Based on their feedback their interest can grow if RDM enables to obtain more contacts and partnerships with others (8/13), if it gives more visibility to their work through data citation (6/13), if it enhances the communication and data sharing with close collaborators (6/13) and if allows data reuse of their data in the medium, long term (5/13). Moreover, some researchers would be more interested in RDM if appropriate data description templates are available for their projects (5/13) and tools are made available for RDM activities (4/13). Three researchers would be more interested in case they have to respond to funding agencies mandates and only 2 if it contributes to their scientific evaluation.

Some researchers provided recommendations on how to improve the activities in which they were engaged. One researcher pointed out that it would be helpful to see practical examples of the use of data description, so that the activity is not so abstract. Other mentioned the need to involve other researchers in a collaborative process. Two other researchers suggested training activities, either within collective research units or even through an institutional training plan, or small training activities with other researchers. 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 with their own expertise to RDM.

The engagement of more and more researchers in RDM activities is likely to encourage others to participate, as a virtuous circle. It was common during this work for researchers to ask for practical examples of data publication in their domains. Therefore, more datasets need to reach the publication stage in order to motivate others to join, so that RDM can become an established practice. Thanks to their involvement some researchers became quite aware of the importance of good RDM practices. In one case, the researchers included the publication of their data in B2SHARE[1] in the project reports, considering it a valuable result of the project [100]. Moreover, two other researchers have collaborated with TAIL to develop of a DMP for their projects. These activities are essential for researchers to become proactive in RDM and to benefit from the collaboration with data curators.

To conclude, researchers have shown aptness to choose and fill in scientific-oriented metadata. Overall, I was able to notice a general development of researchers' awareness and skills in RDM and metadata creation throughout this endeavour. I also see metadata production as a realistic activity from the researchers point of view, and that institutional support is critical for researchers to have confidence in adopting RDM tools.

[1] https://b2share.eudat.eu/records/7b3c66dfa4df4a7f9ba04fbc30cfb8bc

REFERENCES

[1] K. G. Akers and J. Doty. ≪Disciplinary differences in faculty research data management practices and perspectives≫. In: *International Journal of Digital Curation* 8.2 (2013). DOI: 10.2218/ijdc.v8i2.263 (cit. on p. 15).

[2] D Akmon, A Zimmerman, M Daniels, and M Hedstrom. ≪The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs≫. In: *Archival Science* (). DOI: https://doi.org/10.1007/s10502-011-9151-4 (cit. on p. 14).

[3] Al-Hindawe, J. ≪Considerations when constructing a Semantic Differential Scale≫. In: *La Trobe University Working Papers in Linguistics* 9 (1996) (cit. on p. 87).

[4] A. A. Alsheikh-Ali, W. Qureshi, M. H. Al-Mallah, and J. P. A. Ioannidis. ≪Public Availability of Published Research Data in High-Impact Journals≫. In: *PLOS ONE* 6.9 (Sept. 2011), pp. 1–4. DOI: 10.1371/journal.pone.0024357 (cit. on pp. 11, 15).

[5] R. Amorim, J. Castro, J. Rocha da Silva, and C. Ribeiro. ≪A comparison of research data management platforms: architecture, flexible metadata and interoperability≫. In: *Universal Access in the Information Society* 16.4 (2017). DOI: 10.1007/s10209-016-0475-y (cit. on p. 2).

[6] R. C. Amorim, J. A. Castro, J. Rocha da Silva, and C. Ribeiro. ≪LabTablet: Semantic Metadata Collection on a Multi-domain Laboratory Notebook≫. In: *Metadata and Semantics Research Conference Proceedings*. Vol. 478. 2014. DOI: 10.1007/978-3-319-13674-5_19 (cit. on p. 21).

[7] P. Arzberger, P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhlir, and P. Wouters. ≪Promoting access to public research data for scientific, economic, and social development≫. In: *Data Science Journal* 3 (2004) (cit. on p. 14).

[8] M. Assante, L. Candela, D. Castelli, and A. Tani. ≪Are Scientific Data Repositories Coping with Research Data Publishing?≫ In: *Data Science Journal* 15:6 (2016), pp. 1–24. DOI: 10.5334/dsj-2016-006 (cit. on pp. 2, 17, 19, 127).

[9] C. Awre, J. Baxter, B. Clifford, J. Colclough, A. Cox, N. Dods, P. Drummond, Y. Fox, M. Gill, K. Gregory, A. Gurney, J. Harland, M. Khokhar, D. Lowe, R. O. Beirne, R. Proudfoot, H. Schwamm, A. Smith, E. Verbaan, L. Waller, L. Williamson, M. Wolf, and M. Zawadzki. ≪Research Data Management as a ”wicked problem”≫. In: *Library Review* 64.4/5 (2015), pp. 356–371. DOI: https://doi.org/10.1108/LR-04-2015-0043 (cit. on p. 143).

[10] A. Ball. *Tools for research data management*. University of Bath, 2012 (cit. on p. 10).

[11] A. Ball, J. Greenberg, K. Jeffery, and R. Koskela. *RDA Metadata Standards Directory Working Group: Final Report*. 2016. URL: https://rd-alliance.org/system/files/MSDWG-Final-Report.pdf (cit. on pp. 2, 23).

[12] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, et al. ≪The ontology for biomedical investigations≫. In: *PloS one* 11.4 (2016), pp. 1–19 (cit. on p. 76).

[13] G. Bartha and S. Kocsis. ≪Standardization of Geographic Data: The European INSPIRE Directive≫. In: *European Journal of Geography* 22 (2011) (cit. on p. 26).

[14] T. Berners-Lee, J. Hendler, and O. Lassila. *The Semantic Web*. 2001. DOI: 10.1038/scientificamerican0501-34 (cit. on p. 28).

[15] C. L. Borgman. ≪The conundrum of sharing research data≫. In: *Journal of the American Society for Information Science and Technology* 63.6 (2012), pp. 1059–1078. DOI: 10.1002/asi.22634 (cit. on p. 9).

[16] C. L. Borgman, J. C. Wallis, and N. Enyedy. ≪Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries≫. In: *International Journal on Digital Libraries* 7.1-2 (July 2007) (cit. on p. 45).

[17] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, et al. ≪Minimum information about a microarray experiment (MIAME)—toward standards for microarray data≫. In: *Nature genetics* 29.4 (2001), p. 365 (cit. on pp. 75, 80).

[18] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt. ≪A Text Feature Based Automatic Keyword Extraction Method for Single Documents BT - Advances in Information Retrieval≫. In: *Proceedings of the 40th European Conference on Information Retrieval (ECIR'18), Grenoble, France*. Ed. by G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury. Springer International Publishing, 2018, pp. 684–691. ISBN: 978-3-319-76941-7 (cit. on p. 62).

[19] J. Carlson. In: (), pp. 0–19. DOI: 10.5703/1288284315651 (cit. on pp. 39, 82).

[20] J. A. Castro, R. C. Amorim, R. Gattelli, Y. Karimova, J. R. da Silva, and C. Ribeiro. ≪Involving Data Creators in an Ontology-Based Design Process for Metadata Models≫. In: *Developing Metadata Application Profiles*. Ed. by M. C. Malta, A. A. Baptista, and P. Walk. IGI Global, 2017. DOI: 10.4018/978-1-5225-2221-8.ch008 (cit. on p. 45).

[21] J. A. Castro, C. Landeira, and C. Ribeiro. ≪Role of Content Analysis in Improving the Curation of Experimental Data≫. In: 2 (1 2020). DOI: https://doi.org/10.2218/ijdc.v15i1.705 (cit. on pp. 5, 62).

[22] J. A. Castro, D. Perrotta, R. Amorim, J. Rocha da Silva, and C. Ribeiro. ≪Ontologies for research data description: a design process applied to vehicle simulation≫. In: *Metadata and Semantics Research Conference Proceedings*. 2015 (cit. on pp. 4, 141).

[23] J. A. Castro, J. Rocha da Silva, and C. Ribeiro. ≪Creating lightweight ontologies for dataset description. Practical applications in a cross-domain research data management workflow≫. In: *IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. 2014. DOI: 10.1109/JCDL.2014.6970185 (cit. on pp. 4, 45).

[24] J. A. Castro, J. Rocha da Silva, and C. Ribeiro. «Designing an Application Profile Using Qualified Dublin Core: A Case Study with Fracture Mechanics Datasets». In: *Proceedings of the International Conference on Dublin Core and Metadata Applications*. 2013, pp. 47–52 (cit. on p. 4).

[25] T. R. Cech. «Fostering innovation and discovery in biomedical research». In: *JAMA* 294.11 (2005), pp. 1390–1393 (cit. on p. 74).

[26] D. C. Centre and K. P. Ltd. *Data dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability: A comparative review based on sixteen case studies*. January. 2010 (cit. on p. 10).

[27] T. C. Chao. «Enhancing metadata for research methods in data curation». In: vol. 51. 2014. DOI: https://doi.org/10.1002/meet.2014.14505101103 (cit. on p. 62).

[28] T. C. Chao. «Identifying Description Indicators for Research Data from Scientific Journal Publications». In: 2014. DOI: 10.9776/14366 (cit. on p. 62).

[29] P. Cisar, D. Soloviov, A. Barta, J. Urban, and D. Stys. «BioWes-from design of experiment, through protocol to repository, control, standardization and back-tracking». In: *BioMedical Engineering Online* 15 (2016). DOI: 10.1186/s12938-016-0188-8 (cit. on p. 35).

[30] C. Connie, M. Cruz, E. Papadopoulou, J. Savage, M. Teperek, Y. Wang, I. Witkowska, and J. Yeomans. *Engaging Researchers with Data Management: The Cookbook*. Open Book Publishers, 2019. ISBN: 978-1-78374-797-9 (cit. on p. 142).

[31] O. Corcho. «Ontology based document annotation: trends and open research problems». In: *International Journal of Metadata, Semantics and Ontologies* 1.1 (2006), pp. 47–57 (cit. on p. 49).

[32] L. Corti, V. den Eynden, L. Bishop, and M. Woollard. *Managing and Sharing Data: A Guide to Good Practice*. SAGE Publications, 2014. ISBN: 978-1-4462-6725-7 (cit. on p. 22).

[33] A. M. Cox and S. Pinfield. «Research data management and libraries: Current activities and future priorities». In: *Journal of Librarianship and Information Science* 4 (2013). DOI: 10.1177/0961000613492542 (cit. on p. 23).

[34] A. Crystal and J. Greenberg. «Usability of a metadata creation application for resource authors». In: *Library & Information Science Research* 27.2 (2005), pp. 177–189. DOI: 10.1016/j.lisr.2005.01.012 (cit. on p. 46).

[35] R. G. Curty. «Factors Influencing Research Data Reuse in the Social Sciences: An Exploratory Study». In: *International Journal of Digital Curation* 11.1 (2016). DOI: 10.2218/ijdc.v11i1.401 (cit. on p. 15).

[36] Directorate-General for Research and Innovation (European Commission). *Turning FAIR into reality*. 2018, pp. 1–78. DOI: 10.2777/1524 (cit. on pp. 3, 12, 137).

[37] J. Downing, P. Murray-Rust, A. Tonge, P. Morgan, H. Rzepa, F. Cotterill, N. Day, and M. Harvey. «SPECTRa: The deposition and validation of primary chemistry research data in digital repositories». In: *Journal of Chemical Information and Modeling* 48.8 (2008). DOI: 10.1021/ci7004737 (cit. on p. 35).

[38] E. Duval and W. Hodgins. «Metadata principles and practicalities». In: *D-Lib Magazine* 8 (2002), p. 2002 (cit. on p. 49).

[39] O. for Economic Co-operation and Development. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. 2007. URL: https://www.oecd.org/sti/inno/38500813.pdf (cit. on pp. 9, 13).

[40] P. N. Edwards. *A Vast Machine: Computer Models, Climate Data and the Politics of Global Warming*. Ed. by Cambridge, MA, MIT Press. 2010 (cit. on p. 37).

[41] European Commission; Directorate-General for Research & Innovation. H2020 Programme. *Guidelines on FAIR Data Management in Horizon 2020*. 2016 (cit. on pp. 1, 8, 12, 13, 137).

[42] European Commission. *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*. December. 2013, pp. 1–14 (cit. on p. 1).

[43] I. Faniel, A. Barrera-Gomez J.and Kriesberg, and E. Yakel. «A Comparative Study of Data Reuse Among Quantitative Social Scientists and Archaeologists». In: *iConference* (2013). DOI: 10.9776/13391 (cit. on p. 3).

[44] I. Faniel, E. Kansa, S. W. Kansa, J. Barrera-gomez, and E. Yakel. «The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse». In: *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. 2013. DOI: https://doi.org/10.1145/2467696.2467712 (cit. on p. 35).

[45] A. Ferreira. «Application of the LabTablet app in a laboratory environement: Case study I3S». Faculdade de Engenharia da Universidade do Porto, 2019 (cit. on p. 5).

[46] M. Ferreira, L. Gales, V. Fernandes, C. Rangel, and A. Pinto. «Alkali free hydrolysis of sodium borohydride for hydrogen generation under pressure». In: *International Journal of Hydrogen Energy* 35.18 (2010), pp. 9869 –9878. DOI: https://doi.org/10.1016/j.ijhydene.2010.02.121 (cit. on p. 57).

[47] J. R. Finkel, T. Grenager, and C. Manning. «Incorporating non-local information into information extraction systems by Gibbs sampling». In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05* 1995 (2005), pp. 363–370. DOI: 10.3115/1219840.1219885 (cit. on p. 62).

[48] R. Gattelli. «Gestão de dados de investigação no domínio da Oceanografia Biológica: criação e avaliação de um perfil de aplicação baseado em ontologia». Faculdade de Engenharia da Universidade do Porto, 2016 (cit. on p. 5).

[49] A. González-Beltrán, E. Maguire, S. Sansone, and P. Rocca-Serra. «linkedISA: semantic representation of ISA-Tab experimental metadata». In: *BMC bioinformatics* 15.14 (2014), S4 (cit. on p. 74).

[50] S. A. Gore. «e-Science and data management resources on the Web». In: *Medical Reference Services Quarterly* 30.2 (2011). DOI: 10.1080/02763869.2011.562778 (cit. on p. 46).

[51] R. Harris Pierce and Y. Q. Liu. «Is Data Curation Education at Library and Information Science Schools in North America Adequate?» In: *New Library World* 113 (2012). DOI: 10.1108/03074801211282957 (cit. on p. 11).

[52] R. Heery and M. Patel. «Application profiles: mixing and matching metadata schemas». In: *Ariadne* 25 (2000) (cit. on pp. 26, 49).

[53] P. B. Heidorn. ≪Shedding Light on the Dark Data in the Long Tail of Science≫. In: *Library Trends* 57.2 (2008), pp. 280–299. DOI: 10.1353/lib.0.0036 (cit. on pp. 1, 46).

[54] A. J. Hey, S. Tansley, K. M. Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research, 2009 (cit. on p. 1).

[55] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos. ≪The role of ontologies in biological and biomedical research: a functional perspective≫. In: *Briefings in Bioinformatics* 16.6 (2015), pp. 1069–1080. DOI: 10.1093/bib/bbv011 (cit. on p. 75).

[56] R. Johnson, A. Chiarelli, and T. Parsons. *Data asset framework (DAF) survey results 2016*. Oct. 2016. DOI: 10.6084/m9.figshare.3796305.v4 (cit. on p. 11).

[57] B. Jörg. ≪CERIF: The Common European Research Information Format Model≫. In: *Data Science Journal* 9.July (2010), pp. 24–31 (cit. on pp. 28, 49).

[58] S. W. Kansa, E. C. Kansa, and B. Arbuckle. ≪Publishing and Pushing: Mixing Models for Communicating Research Data in Archaeology≫. In: *International Journal for Digital Curation* 9.October 2013 (2014). DOI: 10.2218/ijdc.v9i1.301 (cit. on p. 35).

[59] Y. Karimova. ≪Vocabulários controlados na descrição de dados de investigação no Dendro≫. Faculdade de Engenharia da Universidade do Porto, 2016 (cit. on pp. 5, 144).

[60] Y. Kim and J. Stanton. ≪Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis≫. In: *Journal of the Association for Information Science and Technology* 67.4 (2016). DOI: 10.1002/asi.23424 (cit. on p. 15).

[61] C. Landeira. ≪Gestão de dados de investigação do tipo experimental: casos de uso e contribuições para a melhoria da qualidade dos metadados≫. Faculdade de Engenharia da Universidade do Porto, 2018 (cit. on pp. 5, 142).

[62] M. Lassi, M. Johnsson, and K. Golub. ≪Research data services: An exploration of requirements at two Swedish universities≫. In: *IFLA Journal* 42.4 (2016). DOI: 10.1177/0340035216671963 (cit. on p. 35).

[63] O. Lassila and D. McGuinness. *The role of frame-based representation on the Semantic Web*. 2001 (cit. on p. 49).

[64] T. P. Lauriault, B. L. Craig, D. R. F. Taylor, and P. L. Pulsifer. ≪Today's Data are Part of Tomorrow's Research : Archival Issues in the Sciences≫. In: *Archiveria* 64.Fall (2007), pp. 123–179 (cit. on p. 9).

[65] Y.-f. Li, G. Kennedy, F. Ngoran, P. Wu, and J. Hunter. ≪An ontology-centric architecture for extensible scientific data management systems≫. In: *Future Generation Computer Systems* 29.2 (2013), pp. 641–653. DOI: 10.1016/j.future.2011.06.007 (cit. on pp. 28, 49).

[66] Likert, R. ≪A technique for the measurement of attitudes≫. In: *Archives of Psychology* 140.22 (55 1932) (cit. on p. 86).

[67] L. Lyon. *Dealing with data: roles, rights, responsibilities and relationships*. June. UKOLN, University of Bath, 2007, pp. 1–65 (cit. on p. 8).

[68] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. ≪An ontology for describing and synthesizing ecological observation data≫. In: *Ecological Informatics* 2.3 (2007), pp. 279–296. DOI: 10.1016/j.ecoinf.2007.05.004 (cit. on pp. 28, 49).

[69] L. Martinez-Uribe and S. Macdonald. ≪User engagement in research data curation≫. In: *ECDL*. Vol. 5714. Lecture Notes in Computer Science. 2009 (cit. on pp. 11, 47).

[70] M. E. Martone. ≪Brain and Behavior: we want you to share your data≫. In: *Brain and Behavior* 4.1 (2013). DOI: 10.1002/brb3.192 (cit. on p. 2).

[71] G. Mayer, A. R. Jones, P.-a. Binz, E. W. Deutsch, S. Orchard, L. Montecchi-Palazzi, J. Antonio, H. Hermjakob, D. Oveillero, R. Julian, C. Stephan, H. E. Meyer, and M. Eisenacher. ≪Controlled vocabularies and ontologies in proteomics: Overview, principles and practice≫. In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1844.1 (2014). DOI: 10.1016/j.bbapap.2013.02.017 (cit. on p. 75).

[72] M. Mayernik. ≪Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators≫. In: (2011). DOI: 10.2139/ssrn.2042653 (cit. on p. 82).

[73] M. Mayernik, A. Batcheller, and C. Borgman. ≪How institutional factors influence the creation of scientific metadata≫. In: 2011. DOI: 10.1145/1940761.1940818 (cit. on p. 35).

[74] M. Mayernik. ≪Research data and metadata curation as institutional issues≫. In: *Journal of the Association for Information Science and Technology* 67.4 (2016). DOI: 10.1002/asi.23425 (cit. on p. 35).

[75] L. McEwen and Y. Li. ≪Academic librarians at play in the field of cheminformatics: Building the case for chemistry research data management≫. In: *Journal of Computer-Aided Molecular Design* 28.10 (2014). DOI: 10.1007/s10822-014-9777-4 (cit. on p. 63).

[76] C. Monteiro, C. T. Lopes, and J. R. Silva. ≪Supporting Description of Research Data: Evaluation and Comparison of Term and Concept Extraction Approaches≫. In: *Digital Libraries for Open Knowledge. TPDL 2018.* Ed. by E. Méndez, F. Crestani, C. Ribeiro, G. David, and J. C. Lopes. Springer International Publishing, 2018. ISBN: 978-3-030-00066-0 (cit. on pp. 62, 69).

[77] National Information Standards Organization. ≪Understanding Metadata≫. In: *National Information Standards* MD:NISO Press (2004), p. 20. DOI: 10.1017/S0003055403000534 (cit. on p. 22).

[78] National Science Foundation. *Grants.Gov Application Guide A Guide for Preparation and Submission of NSF Applications via Grants.gov*. National Science Foundation, 2011. URL: http://www.nsf.gov/pubs/policydocs/grantsgovguide0111.pdf (cit. on p. 8).

[79] N. F. Noy and D. L. Mcguinness. ≪Ontology Development 101: A Guide to Creating Your First Ontology≫. In: (2000) (cit. on p. 28).

[80] A. Ogletree. ≪Metadata workflows across research domains: Challenges and opportunities for supporting the DFC cyberinfrastructure≫. In: 2014 (cit. on p. 35).

[81] Osgood, C.E., Suci, G.J., Tannenbaum, P.H. ≪The measurement of meaning≫. In: *Urbana:University of Illionois Press* (1957) (cit. on p. 87).

[82] C. Palmer, A. Thomer, K. Baker, K. Wickett, C. Hendrix, A. Rodman, S. Sigler, and B. Fouke. ≪Site-based data curation based on hot spring geobiology≫. In: *PLoS ONE* 12.3 (2017). DOI: 10.1371/journal.pone.0172090 (cit. on p. 35).

[83] I. V. Pasquetto, B. M. Randles, and C. L. Borgman. «On the Reuse of Scientific Data». In: *Data Science Journal* 16 (2017). DOI: 10.5334/dsj-2017-008 (cit. on p. 137).

[84] L. Perrier, E. Blondal, A. P. Ayala, D. Dearborn, T. Kenny, D. Lightfoot, R. Reka, M. Thuna, L. Trimble, and H. Macdonald. «Research data management in academic institutions: A scoping review». In: *Plos One* 12.5 (2017). DOI: 10.5281/zenodo.557043.Funding (cit. on pp. 32, 34).

[85] D. Perrotta, J. L. Macedo, R. J. Rossetti, J. F. D. Sousa, Z. Kokkinogenis, B. Ribeiro, and J. L. Afonso. «Route Planning for Electric Buses: A Case Study in Oporto». In: *Procedia - Social and Behavioral Sciences* 111 (2014), pp. 1004–1014. DOI: 10.1016/j.sbspro.2014.01.135 (cit. on p. 54).

[86] H. A. Piwowar, R. B. Day, and D. S. Fridsma. «Sharing detailed research data is associated with increased citation rate». In: *PLoS ONE* 2.3 (2007). DOI: 10.1371/journal.pone.0000308 (cit. on p. 22).

[87] L Pouchard. «The Earth System Grid Discovery and Semantic Web Technologies». In: *Workshop for Semantic Web Technologies for Searching and Retrieving Scientific Data. 2nd International Semantic Web Conference*. 2003, pp. 1–6 (cit. on pp. 28, 49).

[88] J. Qin, A. Ball, and J. Greenberg. «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* (2012), pp. 62–71 (cit. on pp. 26, 28, 47, 63, 73).

[89] J. Qin and K. LI. «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*. 2013, pp. 25–34 (cit. on pp. 2, 22, 26–28, 30, 85, 105, 137).

[90] J. Qin, X. Liu, and M. Chen. «Ontology-Enabled Metadata Schema Generator : The Design Approach». In: *Proc. of the International Conference on Dublin Core and Metadata Applications*. 2013, pp. 1–2 (cit. on p. 27).

[91] J. Ray. «The rise of digital curation and cyberinfrastructure». In: *Library Hi Tech* 30.4 (2011). DOI: 10.1108/07378831211285086 (cit. on p. 10).

[92] K. Read. *Common Metadata Elements for Cataloging Biomedical Datasets*. DOI: 10.6084/m9.figshare.1496573.v1 (cit. on p. 76).

[93] «Red flags in data: Learning from failed data reuse experiences». In: *Proceedings of the Association for Information Science and Technology* 53.1 (2016), pp. 1–6. DOI: 10.1002/pra2.2016.14505301126 (cit. on p. 15).

[94] C. Ribeiro and M. E. M. Fernandes. «Data Curation at U.Porto: Identifying current practices across disciplinary domains». In: *IASSIST Quarterly* 35.4 (2012). DOI: 10.29173/iq893 (cit. on p. 3).

[95] C. Ribeiro, J. Rocha da Silva, J. Aguiar Castro, R. Carvalho Amorim, J. Correia Lopes, and G. David. «Research Data Management Tools and Workflows: Experimental Work at the University of Porto». In: *IASSIST Quarterly* 42.2 (2018), pp. 1–16. DOI: 10.29173/iq925 (cit. on p. 3).

[96] R. Rice and J Haywood. ≪Research data management initiatives at University of Edinburgh≫. In: *International Journal of Digital Curation* 6.2 (2011) (cit. on p. 46).

[97] J. Riley. *Understanding Metadata. What is Metadata, and What is it for*. 2017 (cit. on p. 22).

[98] J. Rocha da Silva. ≪Usage-driven application profile generation using ontologies≫. PhD thesis. Faculdade de Engenharia da Universidade do Porto, 2016 (cit. on pp. 4, 19, 60).

[99] J. Rocha da Silva, J. Barbosa, M. Gouveia, J. Correia Lopes, and C. Ribeiro. ≪UPBox and DataNotes: a collaborative data management environment for the long tail of research data≫. In: *iPres Conference Proceedings* (2013) (cit. on p. 49).

[100] J. Rodrigues, J. A. Castro, J. Rocha da Silva, and C. Ribeiro. ≪Hands-On Data Publishing with Researchers: Five Experiments with Metadata in Multiple Domains≫. In: Jan. 2019, pp. 274–288 (cit. on p. 145).

[101] M. Sampaio. ≪Metadados para o uso de ferramentas de gestão com investigadores do I3S≫. Faculdade de Engenharia da Universidade do Porto, 2019 (cit. on p. 5).

[102] M. Sampaio, A. L. Ferreira, J. A. Castro, and C. Ribeiro. ≪Training Biomedical Researchers in Metadata with a MIBBI-Based Ontology≫. In: *Metadata and Semantic Research*. Cham: Springer International Publishing, 2019, pp. 28–39 (cit. on pp. 5, 74, 142).

[103] S. Sansone, P. McQuilton, P. Rocca-Serra, A. Gonzalez-Beltran, M. Izzo, A. L. Lister, and M. Thurston. ≪FAIRsharing as a community approach to standards, repositories and policies≫. In: *Nature biotechnology* 37.4 (2019), p. 358 (cit. on pp. 74, 75).

[104] C. J. Savage and A. J. Vickers. ≪Empirical study of data sharing by authors publishing in PLoS journals≫. In: *PLoS ONE* 4.9 (2009) (cit. on pp. 14, 15).

[105] J. Scaramozzino, M. Ramírez, and K. McGaughey. ≪A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University≫. In: *College Research Libraries* 73 (July 2011), pp. 349–365. DOI: 10.5860/crl-255 (cit. on p. 11).

[106] A. Schmidt-Kloiber, S. Moe, B. Dudley, J. Strackbein, and R. Vogl. ≪The WISER metadatabase: The key to more than 100 ecological datasets from European rivers, lakes and coastal waters≫. In: *Hydrobiologia* 704.1 (2013), pp. 29–38. DOI: 10.1007/s10750-012-1295-6 (cit. on p. 35).

[107] F. Silva, R. C. Amorim, J. A. Castro, J. R. da Silva, and C. Ribeiro. ≪End-to-End Research Data Management Workflows - A Case Study with Dendro and EUDAT≫. In: *MTSR*. 2016 (cit. on p. 21).

[108] L. Soldatova and R. D. King. ≪An ontology of Scientific Experiments≫. In: *Journal of the Royal Society Interface* 3.11 (2006). DOI: 10.1098/rsif.2006.0134 (cit. on pp. 28, 49).

[109] S. Stall, L. Yarmey, J. Cutcher-Gershenfeld, B. Hanson, K. Lehnert, B. Nosek, M. Parsons, E. Robinson, and L. Wyborn. ≪Make scientific data FAIR≫. In: *Nature* 570 (2019). DOI: doi:10.1038/d41586-019-01720-7 (cit. on p. 12).

[110] S. Stemler. ≪An Overview of Content Analysis≫. In: *Practical assessment, research & evaluation* 7.17 (2001) (cit. on p. 61).

[111] C. F. Taylor, D. Field, S. Sansone, J. Aerts, M. Ashburner, C. A. Ball, P.-a. Binz, M. Bogue, A. Brazma, R. R. Brinkman, A. M. Clark, E. W. Deutsch, O. Fiehn, J. Fostel, P. Ghazal, F. Gibson, T. Gray, J. M. Hancock, N. W. Hardy, H. Hermjakob, R. K. Julian, M. Kane, C. Kettner, C. Kinsinger, and E. Kolker. In: *Nature Biotechnology* (). DOI: 10.1038/nbt.1411.Promoting (cit. on pp. 73–75).

[112] C. F. Taylor, N. W. Paton, K. S. Lilley, P.-a. Binz, R. K. J. Jr, R Andrew, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M. J. Dunn, A. J. R. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S. D. Patterson, P. Ping, S. L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, and T. M. Vondriska. «The minimum information about a proteomics experiment (MIAPE)». In: *Nature Biotechnology* 25.8 (2007). DOI: https://doi.org/10.1038/nbt1329 (cit. on p. 75).

[113] C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. «Data Sharing by Scientists: Practices and Perceptions». In: *PLoS ONE* 6.6 (2011). DOI: 10.1371/journal.pone.0021101 (cit. on pp. 14, 46).

[114] C. Tenopir, E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, and K. Dorsett. «Changes in data sharing and data reuse practices and perceptions among scientists worldwide». In: *PLoS ONE* 10.8 (2015), pp. 1–24. DOI: 10.1371/journal.pone.0134826 (cit. on pp. 14, 137).

[115] C. Thanos. «Scientific Data Reusability: Concepts, Impediments and Enabling Technologies». In: *Publications* 5 (1 2016). DOI: 10.3390/publications5010002 (cit. on p. 137).

[116] M. Toups and M. Hughes. «When Data Curation Isn't: A Redefinition for Liberal Arts Universities». In: *Journal of Library Administration* 53 (May 2013). DOI: 10.1080/01930826.2013.865386 (cit. on p. 11).

[117] J. Townsend, S. E. Adams, C. Waudby, V. K. de Souza, J. M. Goodman, and P. Murray-Rust. «Chemical documents: machine understanding and automated information extraction.» In: *Organic & biomolecular chemistry* 2.22 (2004). DOI: 10.1039/b411033a (cit. on p. 63).

[118] A. Treloar and R. Wilkinson. «Rethinking Metadata Creation and Management in a Data-Driven Research World». In: *2008 IEEE Fourth International Conference on eScience* (2008). DOI: 10.1109/eScience.2008.41 (cit. on pp. 2, 29).

[119] P. Vandenbussche, G. A Atemezing, and B. Vatant. «Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web». In: *Semantic Web Journal* 1 (2014), pp. 1–5. DOI: 10.3233/SW-160213 (cit. on p. 62).

[120] T. H. Vines, A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J. S. Moore, S. Renaut, and D. J. Rennison. «The availability of research data declines rapidly with article age». In: *Current Biology* 24.1 (2014), pp. 94–97. DOI: 10.1016/j.cub.2013.11.014 (cit. on p. 1).

[121] H. C. White. «Considering personal organization: Metadata practices of scientists». In: *Journal of Library Metadata* 10.2-3 (2010), pp. 156–172. DOI: 10.1080/19386389.2010.506396 (cit. on pp. 35, 82).

[122] H. C. White. ≪Descriptive Metadata for Scientific Data Repositories: A Comparison of Information Scientist and Scientist Organizing Behaviors≫. In: *Journal of Library Metadata* 14.1 (2014). DOI: 10.1080/19386389.2014.891896 (cit. on pp. 35, 137).

[123] J. M. Wicherts, M. Bakker, and D. Molenaar. ≪Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results≫. In: *PLoS ONE* 6.11 (2011). DOI: https://doi.org/10.1371/journal.pone.0026828 (cit. on p. 15).

[124] A. Wiggins, R. Bonney, E. Graham, S. Henderson, S. Kelling, R. Littauer, G. LeBuhn, K. Lotts, W. Michener, N. Greg, E. Russell, R. Stevenson, and J. Weltzin. ≪Data managment guide for public participation in scientific research.≫ In: *DataONE* (Jan. 2013) (cit. on p. 10).

[125] C. Wiley and E. Kerby. ≪Managing Research Data: Graduate Student and Postdoctoral Researcher Perspectives≫. In: *Issues in Science and Technology Librarianship* Spring (2018). DOI: 10.5062/F4FN14FJ (cit. on p. 15).

[126] C. Wiljes and P. Cimiano. ≪Linked Data for the Natural Sciences: Two Use Cases in Chemistry and Biology≫. In: *Proceedings of the Workshop on the Semantic Publishing (SePublica 2012)* (2012) (cit. on pp. 35, 62).

[127] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. ≪The FAIR Guiding Principles for scientific data management and stewardship≫. In: *Scientific data* 3 (2016). DOI: 10.1038/sdata.2016.18 (cit. on p. 12).

[128] C. Willis, J. Greenberg, and H. White. ≪Analysis and Synthesis of Metadata Goals for Scientific Data≫. In: *Journal of the Association for Information Science and Technology* 63.8 (2012), pp. 1505–1520. DOI: 10.1002/asi.22683 (cit. on pp. 1, 9, 16, 22, 27, 61, 69, 76).

[129] C. Willoughby, C. L. Bird, S. J. Coles, and J. G. Frey. ≪Creating Context for the Experiment Record. User-Defined Metadata: Investigations into Metadata Usage in the LabTrove ELN≫. In: *Journal of Chemical Information and Modeling* 54 (2014), pp. 3268–3283. DOI: 10.1021/ci500469f (cit. on pp. 35, 77, 141).

[130] A. J. Wilson. ≪Toward Releasing the Metadata Bottleneck A Baseline Evaluation of Contributor-supllied Metadata≫. In: *Library Resources & Technical Services* 51.1 (2007), pp. 16–28 (cit. on pp. 2, 46).