

Describing data in image format: Proposal of a metadata model and controlled vocabularies

Joana Rodrigues and Carla Teixeira Lopes

Faculty of Engineering of University of Porto and INESC TEC,
Porto, Portugal.

Contributing authors: joanasousarodrigues.14@gmail.com;
ctl@fe.up.pt;

Abstract

Research data management (RDM) includes people with different needs, specific scientific contexts, and diverse requirements. The description is a big challenge in the domain of RDM. Metadata plays an essential role, allowing the inclusion of essential information for the interpretation of data, enhances the reuse of data and its preservation. The establishment of metadata models can facilitate the process of description and contribute to an improvement in the quality of metadata. When we talk about image data, the task is even more difficult, as there are no explicit recommendations to guide image management. In this work, we present a proposal for a metadata model for image description. To validate the model, we followed an experiment of data description, where eleven participants described images from their research projects, using a metadata model proposed. The experiment shows that participants do not have formal practices for describing their imagery data. Yet, they provided valuable contributions and recommendations to the final definition of a metadata model for image description, to date nonexistent. We also developed controlled vocabularies for some descriptors. These vocabularies aim to improve the image description process, facilitate metadata model interpretation, and reduce the time and effort devoted to data description.

Keywords: Research Data Management, Image Management, Image as research data, Metadata model, Controlled vocabulary

1 Introduction

The exponential production of research data increasingly drives science. Research Data Management (RDM) is an essential requirement for research projects. The absence of good RDM practices can lead to data never reaching its potential and, consequently, losing value over time [1].

The description of the data assumes a determining role in the research data management process. Description allows the data to be properly annotated, and it will enable other researchers to locate and reuse the data after being published [1, 2].

Studies in the field of research data management recognize that involving researchers in data description is problematic, as the lack of knowledge and the absence of RDM practices in research groups are common [3]. For these reasons, it is crucial to establish practices for the management and documentation of research data to foster the access, reuse, and preservation of data in the long term. In parallel, for data to be useful, it is essential to preserve all documentation related to the content, structure, context, and source of the data collection. Following these recommendations, researchers are in line with the European Commission's Guidelines on FAIR Data Management for Horizon 2020, which advocate a set of principles to make data Findable, Accessible, Interoperable and Reusable [4].

Increasingly, researchers are looking for tools and instruments to assist them in managing research data. The description task can be demanding and time-consuming, and using a metadata model can facilitate it. These models propose descriptors that researchers can use to provide a complete description that, in turn, improves the organization of information and enables a correct interpretation of the data [5].

Metadata models can be a valuable tool for improving data, allowing data to be not only material underlying articles but duly documented, preserved, and made available materials, capable of acquiring potential for future research [6]. On the other hand, controlled vocabularies can be auxiliary tools for metadata models, facilitating and improving the description process. The availability of a specific set of values for a descriptor lets the researcher realize the possible options for a given field and avoid confusing, ambiguous, or even wrong descriptions. Allied to this, it streamlines the description, as the researcher can perform this task more quickly and effectively.

Although the potential of metadata models is known, their applicability to data in image format has not been studied in depth. This work proposes a metadata model for the description of images in the context of research. The model can be applied in any research domain and allows the inclusion of metadata specific to particular domains, giving this model a flexible character within its exclusivity to images. That is, descriptors that are unique to a certain domain can be included in order to satisfy the needs of description (example: *chemical compound* descriptor for the *chemistry domain*). In addition to the model, we also propose controlled vocabularies for the metadata model descriptors.

The article begins by presenting a literature review, then moving on to methodology. Afterward, we describe an experiment that allowed its evaluation. After detailing the evaluation of the metadata model, we present the final version of the model and describe the construction of the controlled vocabularies. Finally, we conclude and offer future perspectives of work.

2 Metadata and Research Data Management

Defined as “data about data” [7], metadata is highlighted as essential for the representation and description of the data. In addition, metadata is vital for current scientific communication [8]. When datasets are associated with metadata, data retrieval by a third party is made easier. A more straightforward data interpretation is another benefit obtained from the contextualization gained through the metadata. Another advantage is the greater possibility of reusing data in the future [9].

Metadata is currently considered a key factor for effective functioning and interoperability between systems and can be defined as structured and standardized data that describe an informational resource, with the objective of facilitating its identification for location, search, and retrieval in an information system [10]. The data description process requires skills, effort, time, and adequate tools for high-quality metadata [11]. The level of description detail research data is often scarce, thus it is argued that quality metadata can contribute to ensuring access, interpretation, and consequent reuse of the same.

There are several metadata models for data description. Some are generic, others specific to research domains. These models are essential in the data management tasks of many researchers, as they guide the description process. Some examples are the Dublin Core¹, for generic description, the Data Documentation Initiative², for social science data, the Genome Metadata³, for biology data, and the Astronomy Visualization Metadata⁴, for astronomy data. However, there were no metadata models for the description of images or sets of images. Data in image format may have particularities that require specific metadata for its description, hence the importance of this study.

The development of controlled vocabularies can facilitate and simplify the steps involved in the data description process and, at the same time, improve the quality of the metadata.

Controlled vocabularies are organizations of words or phrases aimed at indexing or retrieving informational content. Generally, they include descriptors or preferred terms. According to Harpring [12], controlled vocabularies consist of an information instrument with standardized terminology to express ideas, physical characteristics, people, places, events, and subjects. Hedden [11] states that a controlled vocabulary consists of a restricted list of words or concepts that, by default, is used for descriptive cataloging or indexing. Controlled

¹<https://dublincore.org/>

²<https://ddialliance.org/>

³<https://www.dcc.ac.uk/resources/metadata-standards/genome-metadata>

⁴<https://www.dcc.ac.uk/resources/metadata-standards/avm-astronomy-visualization-metadata>

4 Describing data in image format

vocabularies support the organization of information, provide terminology for cataloging informational resources and support research and description of the data. One of its most important functions is to bring together the richness of terms variants and synonyms of concepts while promoting consistency by using preferred terms and assigning the same terms to similar content [12]. In this way, controlled vocabularies allow indicating activities or functions and, similarly, confidence in the system [13].

3 Methodology

The methodological approach was organized into five stages. Figure 1 systematizes these phases.

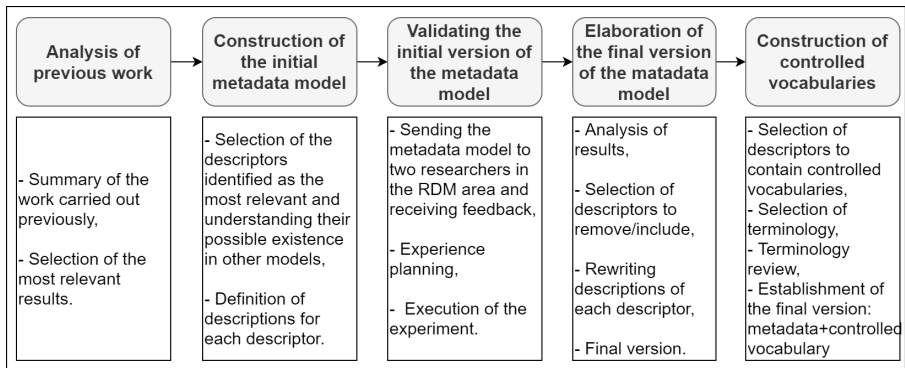


Fig. 1 Stages of the elaboration of the metadata model and controlled vocabularies

3.1 Analysis of previous work

After conducting an exhaustive search, we found no specific metadata model for the general description of images. There are some options to meet the needs of the various research domains, but no indication on how to proceed in the case of data in image format. This work follows two previous studies that explored the practices of researchers regarding the management of research data in image format. The first analyzes the results of a questionnaire answered by researchers from different research domains [14]. The second explores the insights given by researchers in the context of a set of interviews. The two preliminary studies were essential for elaborating the initial metadata model that is presented and evaluated in this study. These studies allowed raising the requirements for the elaboration of the initial metadata model.

3.2 Construction of the initial metadata model

Based on previous work, a set of requirements were established to elaborate the metadata model. These requirements included: descriptors to be included;

existence of descriptors in other models; adaptability of the model to different research domains; adaptability of descriptors to repositories and directories and model flexibility.

3.3 Validating the initial version of the metadata model

To validate this metadata model, an experiment with researchers from different research domains - Life and Health Sciences (LHS), Exact Sciences and Engineering (ESE), Natural and Environmental Sciences (NES), and Social Sciences and Humanities (SSH) - was designed. Metadata model descriptors were asked to describe two images from two of their research projects. This task was supported by a guiding document. From now on, referred to as description sessions.

For the description sessions, we recruited 11 participants in total, 2 (18,18%) from ESE, 3 (27,27%) from LHS, 2 (18,18%) from NES and 4 (36,36%) from SSH. Since the domain of SSH was the one with the lowest incidence in the interviews, it was good to get a larger number of participants from this field in this evaluation experiment.

For the development of this work, all the participants read and filled out a document called “Informed, Free and Clarifies Consent to Participate in a Research Project according to the Declaration of Helsinki and the Convention of Oviedo”, where they attested that they agreed to participate in the study and that they were guaranteed confidentiality, the exclusive use of the data collected and anonymity, promising never to public the identification of participants.

To recruit participants, an email was sent to participants requesting their collaboration and explaining the study in question. If they answered affirmatively, a second email was sent with the date and time of the experiment and asking them to send the two images to be included in the experiment. That said, a personalized form was prepared in Google Forms, with the images of each of the participants (the metadata model was the same in all forms, the images varied). On the day before the session, a new email was sent to recall the experience, with the access link and the informed consent. During the experiment, the participant performed the task and the person in charge of the session collected notes. In the discussion phase, the notes obtained in the previous phase were discussed and new questions and answers were asked. These experiences allowed the realization of the metadata model.

COVID-19 did not allow this experience to happen in person, so we devised a remote experiment. We used a Google Forms document with two pages (one for each image to be described). At the top of the page, we described the goal of the task. Then, we listed the descriptors and, for each, we included a caption that clarified its purpose. Mandatory descriptors were flagged. Figure 2 presents a Screenshot of the form used for the experiment.

Before being used, the forms were evaluated by two researchers knowledgeable in RDM. These researchers provided feedback on the model and the experiment.

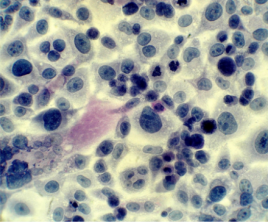
6 Describing data in image format

Descrição de dados em formato de imagem: caso x

Este formulário possui 2 páginas.

***Obrigatório**

Descreva a seguinte imagem através do modelo e metadados disponibilizado a seguir:



Título *
Nome dado à imagem
A sua resposta

Autor *
Entidade responsável pela autoria do recurso
A sua resposta

Descrição *
Uma descrição sobre a imagem
A sua resposta

Data de aquisição *
Ponto ou período de tempo em que a imagem foi adquirida
A sua resposta

Data de produção
Ponto ou período de tempo em que a imagem foi produzida. Caso tenha sido produzida por terceiros e só a intenção de realizar, este descritor não deve ser utilizado
A sua resposta

Período temporal
Ponto ou período de tempo a que o conteúdo da imagem do respeito
A sua resposta

Tipo *
Núcleo ou gênero da imagem. Exemplo: fotografias, imagens médicas, imagens microscópicas, desenhos (ilustrações, vetores, rascunhos, imagens feitas em computadores, mapas, gráficos, imagens digitais, imagens arquitetônicas)
A sua resposta

Fonte
Um recurso de onde a imagem deriva. Exemplo: artigos, livros, páginas web ou outros de onde derivam as imagens. Se a imagem tiver sido obtida pelo próprio ou outro investigador do projeto mencionar a fonte original
A sua resposta

Direitos de utilização *
Uma declaração sobre os direitos de propriedade associados à imagem. Exemplo: CC BY, CC BY-NC-ND, AP, etc. Exemplos: CC BY-NC-ND, CC BY-NC-SA, AAPA, Nature, ACS
A sua resposta

Recursos relacionados
Uma lista de recursos para um recurso intimamente relacionado com a imagem. Exemplos: ligações para a página web do projeto, artigos, dissertações, documentos relativos ao projeto, ligações para repositórios de dados
A sua resposta

Metodologia
Metodologia e o processamento envolvidos na produção da imagem
A sua resposta

Edição (Edition Statement)
Uma declaração sobre as edições/manipulações realizadas na imagem. Exemplo: adição de escala, compressão, recorte, adição de brilho, reaquecimento. Se a imagem for produzida por terceiros não preencher este descritor e apenas que tenha sido concorrentemente que edições foram feitas à imagem
A sua resposta

Formato *
Formato do ficheiro ou meio físico em que se encontra a imagem
A sua resposta

Qualidade
Informação sobre a qualidade dos dados ou quaisquer procedimentos de garantia de qualidade seguidos na produção da imagem. É verificado se a imagem corresponde aos padrões de qualidade exigidos para a investigação? Por exemplo: verificação do dispositivo focal e verificação de objetos saturados. Se a imagem for produzida por terceiros não preencher este descritor
A sua resposta

Instrumento
O nome do instrumento usado para produzir a imagem. Exemplo: máquina fotográfica, microscópio, computador, instrumento médico
A sua resposta

Cobertura espacial
Características espaciais relevantes para a imagem. Pode ser o nome de um local ou as suas coordenadas geográficas
A sua resposta

Escala
Especificar os valores numéricos ou textuais definidos na escala de medida
A sua resposta

Amostra
Descreva a amostra utilizada na experiência que faz uso desta imagem
A sua resposta

Seguinte Limpar formulário

Fig. 2 Screenshot of the form used for the experiment

All sessions were recorded. There was no interaction between the participant and the person conducting the session so that there was no influence on the responses. The person responsible for the session took notes about hesitations, spaces left empty, and time spent on some descriptors. We used these notes and seven other questions to guide a discussion at the end of the session. The questions will be detailed in the Section 5.

It is important to note that participants identification and personal data have not been collected. Before the experiment, participants gave their informed consent that presupposes an explanation and respective understanding of the goal, the procedures, and the expected result of the consented intervention. All recordings for the sessions are safely stored and will be deleted after the study is finished. Details about the realization of vocabularies can be read in the Section 5.

3.4 Elaboration of the final version of the metadata model

After analyzing the various results of the description sessions and listening to the opinions of the participants, the final version of the metadata model for the description of data in image format was elaborated. For each decision, the RDM guidelines were always taken into account. Details about the realization of vocabularies can be read in the Section 6.

3.5 Construction of controlled vocabularies

In a final model, to make the model more robust, reliable, and capable of helping the description, we developed controlled vocabularies for some of the metadata model descriptors. Firstly, we identified which descriptors are susceptible to having a controlled vocabulary. Then, we did an extensive search about the terminology related to each of these descriptors. In the end, we built a dataset containing the metadata model and the controlled vocabularies. With this, we aim to provide the scientific community with all the information regarding this metadata model. Details about the realization of vocabularies can be read in the Section 7.

4 Initial metadata model

The initial metadata model was elaborated based on the results obtained in previous studies, always taking into account the opinion of researchers from different domains. The initial model includes the descriptors most pointed out by the researchers. In addition, it was necessary to understand which descriptors are most used in the general research panorama and which are usually included in repositories, directories, and similar.

We aim to propose a generic metadata model that can be used in every research domain. However, these descriptors must be able to include all relevant information. To obtain a more flexible model, we defined a set of mandatory descriptors - important for interpretation - and optional ones - they are not fundamental for interpretation, but give depth to the data. The model also has to be flexible to the point of allowing the inclusion of new descriptors, that is, having a field of “others.”, where it is possible to include relevant descriptors, in addition to those already stipulated. Finally, the model should not be too long. Is clear that one of the reasons for not describing data is the time it takes to spend. If that time is improved with a pre-established and low number of descriptors, participants tend to describe it more often. In the Table A1, present in the Appendix A, we present the descriptors of the initial metadata model.

5 Evaluation of the metadata model

We analyzed the description sessions according to the notes collected during the experience and the questions asked to the participants at the end of the session. Whenever it was necessary to classify some detail, we resorted to recordings made during the sessions.

5.1 Observation of the description

There were some differences between the participants in the description times. Table 1 presents the time each participant took to describe each image and the number of descriptors used per image. We identify participantes with an

8 Describing data in image format

R followed by the acronym of his research domain and a number incrementally assigned. As can be noted, almost all the participants took more time to describe the first image than the second one. As for the number of descriptors used, there is usually no great difference between those used in the first and second images. In this experiment, 54.5% (6) of the participants used more descriptors in the first image than in the second, 9.1% (1) of the participants used more descriptors in the second image than in the first and 36.4% (4) used the same number of descriptors in both images.

It is important to mention that the experience took place with two images per participant. A large set of images was not used to simplify the experience process, but the metadata model can be applied to the description of a set of images. In this case, the description would be made about the characteristics of the dataset and not each image in particular.

Table 1 Time each participant took to describe images and number of descriptors used per image

	Image 1		Image 2		Image 1+2
	Time	Num	Time	Num	Time
R-H1	10:11	17	04:21	13	14:32
R-ESE1	10:22	15	09:24	15	19:46
R-LSH1	05:12	15	04:48	14	10:00
R-NES1	07:11	16	07:30	16	14:41
R-H2	08:29	17	05:15	16	13:44
R-H3	10:09	15	08:42	15	18:51
R-ESE2	06:59	18	05:52	18	12:51
R-H4	06:24	16	05:12	14	11:36
R-LSH2	08:37	16	08:10	17	16:47
R-LSH3	10:52	15	07:55	14	18:47
R-NES2	09:02	15	08:05	13	17:07
Average	08:37	14.6	07:30	15	14:41

5.1.1 Empty descriptors

During the description, we noticed that some descriptors were not used. These descriptors were: *scale*, *edition statement*, *quality spatial coverage*, and *sample*. Only one participant, from the ESE domain, used all the descriptors available in the model. As a rule, the descriptors that were not filled in the first image were also not filled in the second image.

5.1.2 Hesitations

To identify less clear descriptors, we also analyzed the hesitations that occurred during the sessions. Most of the participants (9 out of 11) struggled distinguishing the *date of production* and *date of acquisition*. Participants often “jumped” between descriptors to understand the differences between the

time descriptors. Two participants (one from SSH and one from LHS) hesitated filling the methodology descriptor. One did not understand what to include because he was reusing images. Another mentioned that, although the *methodology* is important, images are often small details in the projects and the associated procedures are not documented. Also, the descriptors *quality*, *edition* and *sample* were also associated with hesitations and were not always filled.

5.1.3 Results summary

To obtain detailed insights on each descriptor, we have analyzed the time spent by participants on each of them, the length of the description, the number of times they were used, the average time spent by participants on each descriptor, and the average time between the descriptors. In Table 2 we present the average values of these measures. In the table, the *start* indicates when the participant clicks on the box to start writing. The *end* indicates when he finishes writing the last word. The *time* indicates the average time spent by participants on each descriptor. The *time laps* refers to the average time between the end of filling out one descriptor and the beginning of filling out the next. The *char* is the number of characters used in each descriptor. The *used* refers to the percentage of times each descriptor was used.

Table 2 Descriptors usage during the sessions: average start and end times, average time spend on each descriptor, average time between descriptor the, average number of used characters and percentage of times used

descriptor	start	end	time	time lapse	char	used
Title	00:14	00:27	00:13	-	51	100%
Author	00:30	00:40	00:10	00:03	35	100%
Description	00:44	02:01	00:17	00:04	252	100%
Date acquisition	02:08	02:21	00:13	00:07	15	100%
Production date	02:32	02:37	00:05	00:11	15	81.82%
Temporal coverage	02:43	02:48	00:05	00:06	32	72.72%
Type	02:53	03:01	00:08	00:05	17	100%
Source	03:06	03:19	00:13	00:05	58	81.82%
Use rights	03:24	03:38	00:14	00:05	37	100%
Related resources	03:45	04:55	01:10	00:07	221	90.91%
Methodology	05:08	05:47	00:39	00:13	131	100%
Edition Statement	06:09	06:17	00:08	00:22	68	54.55%
Format	06:23	06:29	00:06	00:06	4	100%
Quality	06:39	06:54	00:15	00:10	118	72.72%
Instrument	07:00	07:15	00:15	00:06	11	100%
Spatial coverage	07:26	07:35	00:09	00:11	44	63.63%
Scale	07:45	07:53	00:08	00:10	6	27.27%
Sample	08:04	08:22	00:18	00:11	90	18.18%

The descriptors that, on average, take the longest to be described are *Related resources* and *Methodology* (highlighted in red). With less time, but

with an average of ten seconds or more, are the *Title*, *Description*, *Date acquisition*, *Source*, *Use rights*, *Quality* and *Sample* (highlighted in blue).

The *Related resources* descriptor has a longer filling time. During one session it was noticeable that participants used documents they had saved on your computers to verify these resources and copy their information to the descriptor. This made filling out take longer in most cases. The *Methodology* was seen as one of the descriptors that caused the most hesitation. Although participants consider it very relevant, they sometimes have difficulty selecting exactly what to include in this field, which is why it was one of the most time-consuming descriptors. The *Description*, given its possible length, could be a descriptor that would take longer to be filled. However, it was found that this is one of the descriptors that participants find easier to fill out, so it is faster. *Temporal coverage*, *Production date*, and *Format* are the ones that stand out for the shortest description time. These are descriptors that, as a rule, have a more direct answer, without the need for much detail. Furthermore, they are descriptors that do not use many characters in the writing.

It was noticed that participants usually describe in sequential order. There was no complete reading of the model before filling it out. On average, the participants took fourteen seconds to start filling in the fields, which indicates that they spent some time reading the exercise, even though they already knew what it was about.

As for the time spent between the end of filling out one descriptor and the beginning of filling out the next one, the longer period of time are between the *Date acquisition* - *Production date*; *Related resources* - *Methodology*; *Methodology* - *Edition Statement*; *Instrument* - *Spatial coverage* and *Scale* - *Sample*. The longer time between descriptors can be justified because they were the ones that caused the most hesitation to the participants, as they took some time to interpret and to realize what content would be included in that field.

Description, *Related resources*, *Methodology* and *Quality* are the descriptors with longer descriptions. *Description* and *Related resources* stand out unsurprisingly, as detailed information about the data is expected in both. The first allows to give more context to the data and the second allows to identify a series of resources that can be important for a better understanding of the data and the study where they fit (articles, theses, presentations, posters). Although the *Description* is one of the descriptors with more characters, it is placed within seventeen seconds of completion, which suggests that although it is a long filling descriptor, the participants have information regarding the completion.

We noticed that some descriptors were used more often than others. The *Sample* and *Scale* are the descriptors that were least used, with 18.18% and 27.27%, respectively. Also other descriptors were not always filled. *Related resources* was filled in 90.91% of the times. *Production date* and *Souce* were filled 81.82% of the times. *Quality* and *Temporal coverage* were filled 72.72% of the times. *Edition statement* was filled in 54.55% of the times. Reasons pointed

out by participants include: goal of the descriptor was not clear, descriptor was not relevant, information to fill the description is not available/recorded.

5.2 Feedback from the participants

Some simple questions guided the discussion with the participants at the end of the sessions.

5.2.1 What difficulties were there?

Some topics were highlighted by participants as the greatest difficulties, namely: 1) distinction of the date descriptors, 2) not knowing many quality improvement techniques for including in Quality descriptor, 3) understand which image to describe when it is a digitization (description about the original image or about the digitization?), 4) how to place the source when it is a team hired by the project to obtain the images (is the designation “own authorship.” correct?), 5) the most common is that there is a *Use rights* that covers the entire project and not just the images, so it will be necessary to see if there are specific *Use rights* for the images?, 6) difficulty in understanding exactly what a sample is (may be confused with “universe.”), 7) understand if the acquisition date is important when dealing with historical images and 8) doubt in the inclusion of the Related Resources descriptor, since, although it is very pertinent, it is not a quick descriptor to fill out and that often implies resorting to external search.

5.2.2 Did you miss any descriptors?

Most participants pointed out descriptors that could be included in the metadata model. An SSH participant said he was not familiar with the description of the data, although he considered it important. For that reason, he said he was unable to identify new descriptors, but said that this issue could be discussed in group meetings. An NES participant said that, at the generic level, no descriptor is missing, he just said that an optional free-writing field for specific elements could be useful, ie the model should allow an optional empty field where it is possible to include information that does not match any of the other descriptors. The most requested descriptors were: *Keywords*: most mentioned element (8 times). *Research center*, *Supervision*, *Publisher*, *Research domain*: some of the most mentioned elements and seen as essential for other participants to quickly find a direct relationship with the data (5,4,4,4 times, respectively). *Research group*, *Contacts*, *Others*: often mentioned (all 3 times). *Partnerships/financiers*: considered bureaucratic but essential, because many projects are required to show their financiers/partners (2 times).

5.2.3 What descriptors did you not use and why?

The participant that used all the descriptors said the metadata model seemed relatively simple to fill out and the descriptors useful. The remaining participants mentioned mostly the *Sample*, the *Scale*, the *Edition Statement* and the

Quality as less useful descriptors. For the *Sample*, they say that this can be more useful in a dataset. For the *Scale*, the participants say it may be relevant, but only for certain types of images. In terms of editing and quality, those who did not respond were not sure what to say, however they consider it important. Although most participants have left some descriptors empty, everyone thinks they are relevant, they simply did not apply the images in question, but they may be important to others.

5.2.4 Do you agree with the obligation defined for descriptors?

Seven of the participants agree with the necessity to define mandatory fields and with the fields that were made mandatory. One of these reinforced the idea of including keywords and the research domain as mandatory as well. Another said that contacts and the publisher should also be mandatory. Of the remaining four participants, one said that the mandatory distribution is reasonable, but the date descriptors should be rethought. Another said he would not make the acquisition date mandatory and would only put the descriptor “data” (which would include the images produced and acquired). Another participant questioned whether any descriptor would be mandatory. For this participant this condition can discourage the description of data. He also says that what is necessary for him may not be for others. However, he considers that this matter should be discussed. Finally, a participant sees the existence of mandatory fields as controversial. On the one hand, he thinks that everything should be optional because we never know what it is important for a researcher. On the other hand, having mandatory fields helps to create greater focus and attention at the time of description. The participant thinks that combining the two is the ideal, but never with too much mandatory descriptors.

5.2.5 Is the model Is the model applicable to your research domain?

All participants saw the metadata model as relevant and easily applicable to their projects. In general, the participants mentioned that, if there were descriptors specific to the research domain, the model would be even more applicable. They also stated that the description of the data, regardless of the data format, is necessary and should be a constant practice. Some participants said they did not describe their images, as they considered that the metadata generated by the capture instruments would be sufficient. However, these participants see advantages in this more formal description. Most participants do not describe their images (only two admit to doing so), but are considering starting to do so.

5.2.6 Would some descriptors be more easily filled by other participants in the research project?

Three of the participants replied that they did not. They said that they know all the procedures of the project. They also stated that these particular images would hardly be better described by others, but if they were different images, perhaps this would happen. The participants say that only a comparison could give a correct answer to this question and that the tasks of description have to be assigned according to the knowledge of the images. The rest of the participants say that perhaps some descriptors would be more easily filled in by other participants. One of the participant mentioned that the description could be faster, but that all members of the group are able to describe the images, as everything goes through everyone, mainly by the supervisor. Some participants say that the producer of the original image will always have more advantage in the description than the rest of the group.

5.2.7 Would you like the idea of a collaborative fill?

When asked about making descriptions collaboratively, with other elements of the research group, all participants answered it would speed up the process. They see this as an ideal scenario, as it would decrease time and work and improve sharing. They also claim that this method would help junior participants in the learning process. It is also mentioned that a collaborative description would decrease errors, as the description is reviewed by several people. In addition, the result would be more complete. One participant say that a form in a cloud environment would be a solution, as it allowed each member of the group to describe the data anywhere. Only one participant said that this collaborative scenario should depend on the context. He said that there are cases where only one person knows the details of the images, so adding more people to the description can be confusing.

6 Final metadata model

After analyzing the results of the description sessions, we elaborated the final version of the metadata model. This new version has 26 descriptors, 9 mandatory, and 17 optional.

In this new version, the temporal coverage descriptor was eliminated. It was found that the date descriptors were sufficient for the description, and it is not necessary to include more than the production date and the acquisition date. In the description of the date of production and the date of acquisition, the format in which the date should be written was included. In the previous version of the metadata model, the descriptor date of acquisition was mandatory, however, it was realized that this information does not always exist, so this descriptor became optional.

Four new mandatory descriptors have been included: keywords, research domain, contacts, and use rights. The first two arise, above all, from the needs

presented by the participants. Keywords are seen as one of the most fundamental descriptors, as they allow a quick view of the data content, understanding which sphere it belongs to, and identifying as main characteristics of the data. The researchers also consider it very relevant to demonstrate to which scientific area the data belong, as it is also one of the first pieces of information they look for when they need to do research. The research domain can focus the search, restrict it to the searched results and forward it to the intended area. In turn, contacts and use rights were not often mentioned by researchers, but the research carried out during this work, experience with researchers and projects, and awareness of international guidelines for data management, made us believe that these two descriptors are essential. Without a contact descriptor, it is difficult to contact the data authors, ask for more details, and talk about the data production process, for example. So, reuse and citation can be compromised. Without use rights, it is not possible to guarantee the integrity and intellectual property of the data, allowing them to be lost and not properly associated with the authors, and the citation process may be compromised.

Seven optional descriptors were also included: research group, project, contacts, partnerships, financiers, and material. The first five were also mentioned by the participants as relevant, however, as they may not be necessary in all contexts, they will be optional. The idea of a flexible metadata model was also be contemplated. In addition to having been mentioned by some participants, it was understood through a search on the topic that it is an important feature.. That is, if participants deem it necessary, they can include more descriptors specific to their domains to this metadata model. The material descriptor was added at the initiative of the authors of this work as they consider that this descriptor can be very relevant, especially when it comes to analog images, as the material used for the elaboration of the image can be an important piece of information to be mentioned. This conclusion only emerged after the completion of the first version of the metadata model, especially after better understanding the possible scenarios for the use or production of images.

In the new metadata model, we also improved the descriptions associated with the descriptors to clarify their goal. We detailed the contextualization, gave more examples, and indicated if more than one value was accepted (e.g., *several Edition statement can be indicated*). In the date descriptors, we specified the format expected for the date. Table A2, present in Appendix A, shows the final metadata model. Therefore, was generated a dataset that was described and will be deposited in INESC TEC's Data Repository⁵, in open access for understanding the data in detail. While the article is under review the dataset can be viewed through a shared folder⁶.

This metadata model is not intended specifically for the description of an image, but for datasets of images. Joint description is common in data management practices. Most of the time, not only one image is produced or used in a project, but several. It was considered that the images can be described in

⁵<https://rdm.inesctec.pt/>

⁶<http://shorturl.at/uCMS5>

sets, reducing the time spent and grouping the images by categories. However, it is not ruled out that the images can be individually described, when necessary. The metadata model presented is prepared for both situations, although its main function is aimed at sets of images.

7 Construction of controlled vocabularies

The desire to make the model more interoperable, easier to understand and easier to use, boosted the development of controlled vocabularies, associated with some of the metadata model's descriptors. We expect these vocabularies will facilitate the image description process.

The first step in the development of controlled vocabularies was to select the descriptors that should have an associated vocabulary, that is, the ones with a limited set of possible values. For example, on the one hand, the *Author* descriptor has an infinite set of possibilities, and it is not possible to pre-register all possible names. On the other hand, the *Type* descriptor makes it possible to select all types of images and list them in a controlled vocabulary.

To achieve this goal, it was necessary to select information sources that would clarify in detail the terminology associated with each descriptor. Some descriptors were immediately excluded, as the countless possibilities of terms did not enable the construction of vocabularies.

We selected ten descriptors for the construction of controlled vocabularies: *Research domain*, *Type*, *Format*, *Instrument*, *Material*, *Use rights*, *Methodology*, *Edition Statement*, *Quality* and *Scale*.

In the next step, we defined the specific terminology that would compose each of the controlled vocabularies for describing images. This phase is essential, as a poor choice of terminology can affect the credibility of the model, so the choice of good sources of information (for example, studies, scientific articles, dissertations and theses) is essential. As explained next, we carried out an individual study for each descriptor.

For the *Research Domain*, we followed the guidelines of the Foundation for Science and Technology (FCT)⁷, which is an agency of the Ministry of Education and Science of Portugal that evaluates and finances scientific research activities in the country in all scientific areas. This vocabulary contains seven terms.

For the *Type*, a search was made for all types of analog and digital images possible to capture or elaborate, whether in a recreational or professional nature. This vocabulary contains twenty terms.

For the *Format*, all the possibilities were identified, and the digital images are the ones with the most options. For this descriptor, the elaboration of vocabularies made it clear that formats and supports can be included in this descriptor because when it comes to analog images we speak of support, that is, support is what sustains something, the basis of physical support. This vocabulary contains one hundred and twelve terms.

⁷<https://www.fct.pt/>

For the *Instrument*, we identified, for each type of image, which instruments could be used for its capture or elaboration. This vocabulary contains fifteen terms.

The descriptor *Material* was challenging because, before the vocabularies began to be elaborated, the material and the instrument were being confused. In fact, material and instrument are different things, and in the image domain, they have different roles. Therefore, for this descriptor, we identified all the materials where it was possible to see or stamp the images. As it is a physical material, this descriptor and its controlled vocabulary only apply to analog images. This vocabulary contains twenty one terms.

For the *Use rights*, a study was made of the existing licenses to support the research. Were analyzed the web resources of the various accreditation and licensing institutions available for scientific production, such as Creative Commons and Elsevier. This vocabulary contains thirty five terms.

As participants had difficulties filling out the descriptor *Methodology*, we defined six main methodological scenarios. In case the vocabulary information is not enough, the participant still has the possibility to give more details about the methodology in the description or select the option “other.” and specify other methodological approaches. This vocabulary contains six terms.

We studied the various programs and methods for editing and analyzing images to create the controlled vocabularies for the descriptors of *Edition Statement* and *Quality*. This vocabulary contains forty one terms.

For the *Scale* all possibilities of applying scales were selected. This analysis allowed selecting a set of possible scales for the types of image: maps, photographs, and screens. This vocabulary contains one hundred and five terms.

Due to the high number of terms of the controlled vocabularies, it is difficult to present them in this article. Therefore, was generated a dataset that was described and will be deposited in INESC TEC’s Data Repository⁸, in open access for understanding the data in detail. While the article is under review the dataset can be viewed through a shared folder⁹.

8 Discussion and Conclusion

Research Data Management is associated with several challenges and potentials. The description is an essential phase in data management, however, it is often seen as a difficult, complex, and time-consuming process. Although most participants recognize that description practices are very relevant for their projects, in most cases this is not done.

This work confirms the idea, already demonstrated, that metadata models can be strong allies of participants in the data management process. The benefits are based on the possibility of using a valid, robust, and capable instrument of meeting their needs. Furthermore, the models reduce the time spent,

⁸<https://rdm.inesctec.pt/>

⁹<http://shorturl.at/uCMS5>

as they present participants with a set of terms that are relevant to describe their data, without having to search for the descriptors in external sources. We realized, therefore, that having a metadata model implies less work and this is a valid argument that participants are looking for, since the speed of the days and deadlines to meet make it necessary to select the most important tasks to perform.

The description of data in image format still causes strangeness. It was possible to realize that many participants associate the description only with numerical and textual data. This is supported by the inexistence of a metadata model dedicated exclusively to the description of images/sets of images from a generic point of view, regardless of the research domain. In this sense, this proposed metadata model is a necessary step in the evolution of image description.

It became evident that participants are looking for direct and easily interpretable descriptors. The inclusion of captions in the descriptors was seen as essential for a faster description, it was easier to understand what is necessary to include in the field. Furthermore, the support of examples in the caption is seen as very useful and even essential in certain descriptors, such as Quality, Edition statement, and Sample.

Another characteristic that the participants pointed out is the need for the model to be sufficiently generic, as most do not see the need for an overly exhaustive description. That is, they are looking for a complete description, but without too much depth.

The metadata model presented in this work aims to meet this need. It is generic and does not depend on a research domain, as it includes descriptors that cover all the characteristics of the image. However, it allows for specificity through the possibility of including new descriptors, when necessary. Furthermore, through 26 descriptors, are given several possibilities for describing images, precisely with the purpose of contextualizing this data typology.

It was also clear that the model could not only have mandatory descriptors, as not all images have the same characteristics. Sometimes, what is useful in the description of one image may not be relevant in another. The nature of the image, the research context, the purpose of use or production of the image, are some of the many characteristics that can dictate which descriptors to use in the description moment. However, it is very important to include optional descriptors, to remember and identify description possibilities that otherwise might be forgotten. The choice to determine the mandatory descriptors involved reflecting on those that are essential to contextualize the data, albeit superficially.

Finally, it was clear that participants take longer than desired to fill in the description fields. This delay is often associated with realizing exactly what must be placed in the field. Captions with examples help with this task. However, the inclusion of controlled vocabularies can be important in reducing this time, as they offer a set of terms for description. Therefore, the participant will not have to think about what to write, just select the correct option.

Controlled vocabularies are a very useful tool in description and a significant complement not only in the description but also in quality, as they avoid doubtful and deviant descriptions.

Acknowledgements

Joana Rodrigues is supported by research grant from FCT - Fundação para a Ciência e Tecnologia: PD/BD/150288/2019. Special thanks to João Castro e José Devezas for the availability and help given in the validation of the form. Thanks to Joana Almeida for her contribution to the construction of controlled vocabularies.

Data availability statement

Availability of data: After the article review phase, the data will be openly available in a public repository that issues datasets with DOIs.

Template for data availability statement: Was generated a dataset that will be deposited and described in INESC TEC's Data Repository (<https://rdm.inesctec.pt/>), in open access for understanding the data in detail. The deposit will have an associated DOI. While the article is under review, the dataset can be viewed through a shared folder: <http://shorturl.at/uCMS5>.

Policy: All.

References

- [1] Christine Borgman. Advances in Information Science: The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, 6(63):1059–1078, 2012.
- [2] Veerle Van den Eynden, Louise Corti, Matthew Woollard, Libby Bishop, and Hortonv Laurence. Managing and Sharing Data: A Guide to Good Practice. *UK Data Archive*, 2011.
- [3] João Aguiar Castro, Ricardo Carvalho Amorim Amorim, Rúbia Gattelli, Yulia Karimova, João Rocha da Silva, and Cristina Ribeiro. Involving Data Creators in an Ontology-Based Design Process for Metadata Models. *Developing Metadata Application Profiles*, page 181–214, 2017.
- [4] Directorate-General for Research and Innovation. Guidelines on fair data management in horizon 2020. *European Commission*, 2016.
- [5] João Aguiar Castro, Cristiana Landeira, João Rocha da Silva, and Cristina Ribeiro. Role of Content Analysis in Improving the Curation of Experimental Data. *International Journal of Data Curation*, 15(1), 2017.

- [6] K.G. Akers and J. Doty. Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Data Curation*, 8(2):5–26, 2013.
- [7] Jenn Riley. Understanding Metadata: What Is Metadata, and What is it for? *NISO Primer. National Information Standards Organization (NISO)*, 2017.
- [8] Craig Willis, Jane Greenberg, and Hollie C. White. Analysis and Synthesis of Metadata Goals for Scientific Data. *Journal of the American Society for Information Science and Technology*, 8(63):1505–1520, 2012.
- [9] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6):1–21, 2011.
- [10] J. C. Castro, D. Perrotta, R. Amorim, J. R. da Silva, and C. Ribeiro. Ontologies for research data description: a desing process applied to Vehicle Simulation. *Metadata and Semantics Research Conference*, 2015.
- [11] Heather Hedden. DTaxonomies and Controlled Vocabularies Best Practices for Metadata. *Journal of Digital Asset Management*, 2010.
- [12] Patricia Harping. Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works. *Getty Publications*, 2010.
- [13] Kenneth P. Smith, Leonard J. Seligman, and Vipin Swarup. Everybody Share: The challenge of data-sharing systems. *Computer*, 41(9):54–61, 2008.
- [14] M. Fernandes, J. Rodrigues, and C. T. Lopes. Management of research data in image format: An exploratory study on current practices. *International Conference on Theory and Practice of Digital Libraries - Digital Libraries for Open Knowledge*, pages 212–226, 2020.

Appendix A

Table A1 Image description metadata model descriptors - before description sessions

Descriptor	Description
DC:Title ¹	Name given to the image.
DC:Author ¹	Entity responsible for authoring the resource.
DC:Description ¹	A description about the image.
DC:Date acquisition ¹	Point or time period at which the image was acquired.
DC:Production date ²	Point or time period at which the image was produced. If it has been produced by a third party and is only being reused, this descriptor should not be used.
DC:Temporal coverage ²	Point or time period to which the image content relates.
Type ¹	Nature or gender of the image. Example: photographs, medical images, microscopic images, drawings, illustrations, portraits, videos, prints, computer-made images, maps, graphics, animations, paintings, architectural plans.
DC:Source ²	A resource from which the image derives. Example: articles, books, web pages or others from which the images are derived. If the image was obtained by the project's own or other researcher, mention "own authorship."
DC:Use rights ²	A statement of ownership rights associated with the image. Example: CC BY, CC BY-NC-ND, APA, Elsevier, CC BY-NC-ND, CC BY-NC-SA, AAAS, Nature, ACS.
DC:Related resources ²	Link or note to a resource strictly related to the image. Examples: links to the project's web page, articles, dissertations, documents related to the project, links to data repositories.
DDI:Methodology ²	Methodology and processing involved in image production.
Edition Statement ²	A statement about the edits/manipulations made to the image. Example: adding scale, compositing, cutting, adding brightness, reframing. If the image is produced by a third party, do not fill in this descriptor, unless you know specifically which edits were made to the image
DC:Format ¹	File format or physical medium in which the image is located.
DIFF:Quality ²	Information about the quality of the data or any quality assurance procedures followed in the image production. Is it checked whether the image meets the quality standards required for the investigation? For example: checking for focal blur and checking for saturated objects. If the image is produced by a third party, do not fill in this descriptor.
DIFF:Instrument ²	The name of the instrument used to produce the image. Example: camera, microscope, computer, medical instrument.
DC:Spatial coverage ²	Spatial features relevant to the image. It can be the name of a place or its geographical coordinates.
QUDV:Scale ²	Specifies the numerical or textual values defined on the measurement scale.
DDI:Sample ²	Describes the sample used in the experiment that makes use of this image.

¹ Mandatory² Optional

DC: Dublin Core; DDI: Data Documentation Initiative; DIF: Directory Interchange Format; QUDV:Quantities, Units, Dimensions and Values

Table A2 Image description metadata model descriptors - after description sessions

Descriptor	Description	Controlled vocabulary terms
Research domain ¹	Work field in which the images fits.	7
DC:Title ¹	Name given to the set of images.	
DC:Author ¹	Entity responsible for authoring the resources. Several authors can be indicated.	
DC:Description ¹	A description of the set of images.	
DC:Keywords ¹	Describe the theme of the images (indicates informational content). Several keywords can be indicated.	
Type ¹	Nature or gender of the images. Example: photographs, medical images, microscopic images, drawings, illustrations, portraits, videos, prints, computer-made images, maps, graphics, animations, paintings, architectural plans. Several types can be indicated.	20
DC:Format ¹	File format or physical medium in which the images is located. Several formats can be indicated.	112
Contacts ¹	Contact of the person in charge of the images.	
DC:Use rights ¹	A statement of ownership rights associated with the images. Example: CC BY, CC BY-NC-ND, APA, Elsevier, CC BY-NC-ND, CC BY-NC-SA, AAAS, Nature, ACS.	35
DIFF:Instrument ²	The name of the instrument used to produce the images. Example: camera, microscope, computer, medical instrument, pencil, . Several instruments can be indicated.	15
Material ²	The name of the materials used to make the images. Example: charcoal, ink, graffiti. Several materials can be indicated.	21
DC:Date acquisition ²	Point or time period at which the images were acquired. If the images were produced by the working group, do not fill them out. Follow the format YYYY-MM-DD.	
DC:Production date ²	Point or time period at which the images were produced. If they were produced by a third party and are only being reused, this descriptor should not be used. Follow the format YYYY-MM-DD.	
DC:Temporal coverage ²	Point or time period to which the images content relates.	
DC:Spatial coverage ²	Spatial features relevant to the images. It can be the name of a place or its geographical coordinates. Several spatial coverage can be indicated.	
DC:Source ²	A resource from which the images derives. Example: articles, books, web pages or others from which the images are derived. If the images were obtained by the project's own or other researcher, mention "own authorship.". Several source can be indicated.	
DC:Related resources ²	Link or note to a resource strictly related to the images. Examples: links to the project's web page, articles, dissertations, documents related to the project, links to data repositories. Several related resources can be indicated.	
DDI:Methodology ²	Methodology and processing involved in images production or acquisition.	6
Edition Statement ²	A statement about the edits/manipulations made to the images. Example: adding scale, compositing, cutting, adding brightness, reframing.If the images are produced by a third party, do not fill in this descriptor, unless you know specifically which edits were made to the images. Several edition statement can be indicated.	41
DIFF:Quality ²	Information about the quality of the data or any quality assurance procedures followed in the production of the images. Is it checked whether the images meet the quality standards required for the investigation? For example: checking for focal blur and checking for saturated objects. If the image is produced by a third party, do not fill in this descriptor. Several quality procedurescan be indicated.	24
QUDV:Scale ²	Specifies the numerical or textual values defined on the measurement scale.	105
DDI:Sample ²	Describes the sample used in the experiment that uses these images.	
Research group ²	Name of the research center bearing the set of images.	
Project ²	Name or designation of the project involved in the use/production of the images.	
Partnerships ²	Person or entity that was a partner/collaborator in obtaining or producing the set of images.	
Financiers ²	Entity or institution that financed resources for obtaining or producing the set of images.	

¹Mandatory²Optional

DC: Dublin Core; DDI: Data Documentation Initiative; DIF: Directory Interchange Format; QUDV:Quantities, Units, Dimensions and Values