

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

DOCTORAL PROGRAM IN DIGITAL MEDIA

SPECIALIZATION IN JOURNALISM, DATA AND INFORMATION MANAGEMENT

Images as data and metadata: management practices to promote Findability, Accessibility, Interoperability and Reusability of research data

Joana Patrícia de Sousa Rodrigues

D

2024



Images as data and metadata: management practices to promote Findability, Accessibility, Interoperability and Reusability of research data

Joana Patrícia de Sousa Rodrigues

Faculdade de Engenharia da Universidade do Porto

Doctoral Program in Digital Media

Specialization in Journalism, Data and Information Management

Supervisor: Carla Teixeira Lopes

Joana Rodrigues was supported by research grant from FCT - Fundação para a Ciência e Tecnologia: PD/BD/150288/2019.

Abstract

The evolution of the use and production of images is evident. Technological development favors the emergence of images. This scenario is also applicable in the field of scientific research, where more and more instruments are used to facilitate image capture. With this context as a motto, and knowing that images have particular characteristics, this work explores the dynamics of images in research according to two scenarios: images as data and images as metadata. From the perspective of images as data, the aim is to understand how they can be important agents in the research process. From the perspective of images as metadata, the aim is to understand how they can be used as a tool capable of faithfully describing a dataset and promoting its adequate contextualization and interpretation.

Through exploratory research, several experiments with images as data and as metadata were conducted, based on the dynamics and principles of Research Data Management (RDM). The work begins with an overview of the literature on the subject, in which the management of research data and its relationship with the use and production of images is explored, namely a study of the evolution of the image in the context of scientific, going through its challenges and opportunities, to the way it is treated by researchers, always with RDM as a guiding line.

Subsequently, we move on to a view of the informational behavior of those who use and produce images in the context of research, namely in the search to identify behavior patterns in four studies. The first study starts from the perspective of a specific research project and the remaining three studies take a perspective that includes more stakeholders from different research domains. Thus, the scenarios of images as data and as metadata are explored, through the presentation of experiences that mostly involved researchers.

Para o caso das imagens como dados, são apresentados quatro estudos, com contribuições específicas, começando pela proposta de um novo modelo de metadados para a descrição de dados imagéticos e respectivos vocabulários controlados. The remaining three studies focus on automatic image processing, the publishing and sharing of data, and, finally, the definition of guidelines for using and producing images in the context of research.

For the case of images as metadata, we conducted two studies, the first exposing a new perspective on data description, through the use of images as metadata. It is this new perspective that challenges us to take a new approach, which culminates in the second study, which is based on an experience developed with researchers and aims to alert and raise awareness of the use of other types of documents in the scientific process.

This work has made it clear that images do not always have a formal, standardized role in research data management. However, researchers are open to including images more formally in the scientific process. Furthermore, it is clear that researchers can identify advantages in the use and production of images. The general trend of exponential growth in the production and use of images (motivated by technological developments and new sharing and dissemination tools) is also seen in the context of research, as the various participants in the studies carried out identified

it as such, which points to an imminent need to make the role of images representative in the context of research data collection.

The contributions obtained from this work are reflected in the response to the need for including images in the RDM. Starting from a more conceptual approach, a consistent overview was obtained about the habits of those who use and produce images through the analysis carried out in the literature, namely through the systematic review carried out and which includes several analysis variables, up to studies carried out through questionnaires and surveys and which allowed direct contact with researchers, thus allowing us to relate what theory tells us with what practice does. In the context of images as data, the metadata model and controlled vocabularies are a significant contribution, as they explore new possibilities of description, much more focused on the type of data in question (images). Furthermore, studies of sharing platforms, as well as the possibilities of automatic description of images, made it possible to generate a starting point and determine the path to follow. Still, in the context of images, it is important to highlight the guidelines for the use and production of images that appear as a contribution to the scientific community, in the sense that they guide specific image practices and needs. In the context of images as metadata, the great contribution lies, precisely, in the presentation of an innovative approach to data description, where the particular characteristics that images have are highlighted and that position them as allies of researchers in the data description process and present practical examples of how this description through images could occur.

Keywords: Image, Image as Data, Image as Metadata, Information Behavior, Research Data Management.

Resumo

A evolução da utilização e produção de imagens é evidente. O desenvolvimento tecnológico favorece o surgimento de imagens. Este cenário é também aplicável no domínio da investigação científica, onde cada vez mais são utilizados instrumentos que facilitam a captura de imagens. Tendo este contexto como mote, e sabendo que as imagens têm características particulares, este trabalho explora a dinâmica das imagens na investigação segundo dois cenários: imagens como dados e imagens como metadados. Numa perspectiva de imagens como dados, o objetivo é perceber como podem ser agentes importantes no processo de investigação.. Na perspectiva das imagens como metadados, o objetivo é perceber como podem ser utilizadas como uma ferramenta capaz de descrever fielmente um conjunto de dados e promover a sua adequada contextualização e interpretação.

Através de uma pesquisa exploratória, foram realizadas várias experiências com imagens como dados e como metadados, tendo por base as dinâmicas e os princípios da Gestão de Dados de Investigação (GDI). O trabalho inicia-se com uma visão geral da literatura sobre o tema, na qual é explorada a gestão de dados de investigação e a sua relação com a utilização e produção de imagens, nomeadamente um estudo da evolução da imagem no contexto científico, passando pelos seus desafios e oportunidades, até à forma como é tratada pelos investigadores, sempre com a GDI como linha orientadora.

Posteriormente, passamos a uma visão do comportamento informacional de quem utiliza e produz imagens no contexto da investigação, nomeadamente na procura de identificação de padrões de comportamento em quatro estudos. O primeiro estudo parte da perspectiva de um projeto de investigação específico e os restantes três estudos adoptam uma perspectiva que inclui mais intervenientes de diferentes domínios de investigação. Assim, são explorados os cenários das imagens como dado e como metadado, através da apresentação de experiências que envolveram maioritariamente investigadores.

For the case of images as data, four studies are presented, with specific contributions, starting with the proposal for a new metadata model for the description of image data and respective controlled vocabularies. Os restantes três estudos incidem sobre o processamento automático de imagens, a publicação e partilha de dados e, finalmente, a definição de orientações para a utilização e produção de imagens no contexto da investigação.

Para o caso das imagens como metadados, realizámos dois estudos, o primeiro expondo uma nova perspectiva sobre a descrição de dados, através da utilização de imagens como metadado. É esta nova perspectiva que nos desafia a uma nova abordagem, que culmina no segundo estudo, que se baseia numa experiência desenvolvida com investigadores e que pretende alertar e sensibilizar para a utilização de outro tipo de documentos no processo científico.

Este trabalho tornou claro que as imagens nem sempre têm um papel formal e normalizado na gestão de dados de investigação. No entanto, os investigadores estão abertos a incluir as imagens de forma mais formal no processo científico. Além disso, é evidente que os investigadores podem identificar vantagens na utilização e produção de imagens. A tendência geral de crescimento

exponencial da produção e utilização de imagens (motivada pelos desenvolvimentos tecnológicos e pelas novas ferramentas de partilha e disseminação) verifica-se também no contexto da investigação, como os vários participantes nos estudos realizados a identificaram isso como tal, o que aponta para uma necessidade iminente de tornar representativo o papel das imagens no contexto da recolha de dados de investigação.

As contribuições obtidas com este trabalho reflectem-se na resposta à necessidade de inclusão das imagens na GDI. Partindo de uma abordagem mais concetual, obteve-se um panorama consistente sobre os hábitos de quem usa e produz imagens através da análise efetuada na literatura, nomeadamente através da revisão sistemática realizada e que inclui diversas variáveis de análise, até aos estudos realizados através de questionários e inquéritos e que permitiram o contacto direto com os investigadores, permitindo assim relacionar o que a teoria nos diz com o que a prática faz. No contexto das imagens como dados, o modelo de metadados e os vocabulários controlados são contribuições significativas, na medida em que exploram novas possibilidades de descrição, muito mais direccionadas para o tipo de dados em questão (imagens). Além disso, o estudo das plataformas de partilha, bem como as possibilidades de descrição automática de imagens, permitiram gerar um ponto de partida e determinar o caminho a seguir. Ainda no contexto das imagens, é importante destacar as orientações para o uso e produção de imagens que surgem como um contributo para a comunidade científica, no sentido em que orientam práticas e necessidades específicas de imagem. No contexto das imagens como metadados, a grande contribuição reside, precisamente, na apresentação de uma abordagem inovadora à descrição de dados, onde são realçadas as características particulares que as imagens possuem e que as posicionam como aliadas dos investigadores no processo de descrição de dados e apresentam exemplos práticos de como esta descrição através de imagens poderia ocorrer.

Palavras-chave: Comportamento Informacional, Gestão de Dados de Investigação, Imagem, Imagem como Dado, Imagem como Metadado.

Acknowledgements

The completion of this doctoral work involved a long and laborious process. Along the way, many people have accompanied and guided me to achieve the best results.

First of all, I would like to say a special thank you to Professor Carla Teixeira Lopes, the supervisor of this work, who made a point of being there for me throughout the whole process, with the best advice, tips, and suggestions, with a critical spirit and always expecting the best from me. At no time, not even when I was most vulnerable, did she show any kind of concern or fear that I wouldn't succeed. That trust was essential and that's why I want to thank you so much.

I would like to thank the professors who have accompanied me along the way, always with a word of motivation and courage. To Professor Cristina Ribeiro, who opened the doors to research and helped me grow a lot. To Professor Gabriel David, who gave me new and incredible opportunities, some of them of great responsibility. To Professor António Coelho, director of the doctoral program, for his guidance and for always being available for any clarification. To Professor Teresa Silveira, who always has a word of encouragement to give, confident in my abilities and proud of my path. To Professor Olívia Pestana, for whom I have and will always have great affection, for whom my relationship goes beyond academia and to whom I owe a great deal. From the very first day, she placed all her trust in me and I have the greatest admiration for her.

Thanks to all my colleagues at InfoLab who have made work a place of good humor and sharing, especially Inês Koch, a colleague and friend on the journey. I would also like to thank Marisa Silva, who is responsible for the secretariat of the Doctoral Program in Digital Media, and Ana Paula Silva, who is responsible for the secretariat of INESC TEC's HumanISE research center, who have always been willing to help and have always had a friendly word for me.

I would like to mention the students who crossed paths with my work. I would like to thank Maria Nunes, Miguel Fernandes, Joana Almeida and Pedro Oliveira, whom I had the pleasure of supervising in their academic work, for accepting the challenge I set them with dedication, enthusiasm and professionalism, and for collaborating in some stages of this work. I would also like to thank everyone who, voluntarily and kindly, participated and collaborated in the studies that led to this work, namely all researchers from different domains and research institutions.

The next thank you goes to my friends, those who enthusiastically share my achievements and who always allow me to feel supported. My thanks go to those who have been with me forever and those who have been with me since 2012.

Finally, but always first in my life, thank you to my family, who have always been proud to see me grow. With special affection, I would like to thank my father, my mother, my sister and my boyfriend. They are the pillars that hold me up, support me, accompany me, and give me stability. Without them, it would be harder and less exciting. There hasn't been a moment when I haven't seen their unwavering pride and confidence in my path. No words will ever be enough to thank them for everything they are and everything they do for me. This achievement, which I celebrate in this document, is for me, but everything I have achieved is also for them and that is why I dedicate this doctorate to them.

Contents

1	Introduction	1
1.1	Contextualization	1
1.2	Motivation	2
1.3	Objectives	3
1.4	Methodological approach	5
1.5	Publications and contributions	6
1.6	Structure of the dissertation	9
2	Literature review	11
2.1	Research data management	11
2.1.1	Research data definition	11
2.1.2	FAIR principles and open data	12
2.1.3	Benefits and challenges in sharing research data	13
2.1.4	Research data management cycle	14
2.2	Use of images in research	19
2.2.1	Imagery resources	20
2.2.2	The image as scientific evidence	21
2.3	Description of research data	22
2.3.1	Metadata	23
2.3.2	Metadata models	24
2.3.3	Controlled vocabularies	27
2.4	Image description	27
2.4.1	Challenges and trends	27
2.4.2	Automatic description of images	30
3	Informational behavior of researchers using and producing images	31
3.1	The case of the Generation.Mobi project	31
3.1.1	Contextualization	31
3.1.2	The Generation.Mobi project	32
3.1.3	Generation.Mobi workflow	32
3.1.4	Methodological approach	33
3.1.5	Results	34
3.1.6	Final remarks	38
3.2	Exploring image management practices through questionnaires	39
3.2.1	Contextualization	39
3.2.2	Methodology	39
3.2.3	Results	41
3.2.4	Discussion	48

3.2.5	Final remarks	49
3.3	Exploring image management practices through interviews	49
3.3.1	Contextualization	50
3.3.2	Methodology	50
3.3.3	Results	52
3.3.4	Discussion	59
3.3.5	Final remarks	61
3.4	What does literature tell us about research image management practices?	62
3.4.1	Contextualization	62
3.4.2	Methodology	63
3.4.3	Results	65
3.4.4	Discussion of results	78
3.4.5	Final remarks	79
4	Image as data	81
4.1	Image data description: proposal of a metadata model and controlled vocabularies	81
4.1.1	Contextualization	82
4.1.2	Methodology	82
4.1.3	Initial metadata model	85
4.1.4	Evaluation of the metadata model	86
4.1.5	Feedback from the participants	89
4.1.6	Final metadata model	91
4.1.7	Construction of controlled vocabularies	93
4.1.8	Final remarks	95
4.2	Automatic description of research images: utopia or reality?	96
4.2.1	Contextualization	96
4.2.2	Description of images	97
4.2.3	Methodology	98
4.2.4	Results	99
4.2.5	Final remarks	100
4.3	Solutions for data sharing and storage: a comparative analysis of data repositories	101
4.3.1	Contextualization	101
4.3.2	Repository analysis	102
4.3.3	Final remarks	107
4.4	Guidelines for a standardized practice in the use and production of images in research	107
4.4.1	Contextualization	108
4.4.2	General guidelines	108
4.4.3	Guidelines for research image management	109
5	Image as metadata	119
5.1	A new perspective for describing research data	119
5.1.1	Contextualization	119
5.1.2	Methodology	120
5.1.3	Data description in RDM: Standard Approach	121
5.1.4	Images as Metadata: A New Perspective	122
5.1.5	Challenges and Opportunities for RDM	126
5.1.6	Final remarks	127
5.2	Images to describe research data: a case study on the use of imagery metadata . .	128
5.2.1	Contextualization	128

5.2.2	Methodology	128
5.2.3	Results	130
5.2.4	Inclusion of images in the description process	133
5.2.5	Discussion	136
5.2.6	Final remarks	137
6	Conclusion	139
6.1	Informational behavior of researchers using and producing images	140
6.1.1	RQ1. How does the production and use of images occur in science? . . .	141
6.2	Images as data	142
6.2.1	RQ2. How can RDM practices be adapted to the case of images as re- search data?	143
6.3	Images as metadata	144
6.3.1	RQ3. Can images improve the description and interpretation of research data through their use as metadata?	144
6.4	Perspectives and future work	145
	References	147
A	Semi-structured interview guide: Exploring image management practices through interviews	159
B	Semi-structured interview guide: Images to describe research data: a case study on the use of imagery metadata	161

List of Figures

1.1	Methodological approach to doctoral work	5
1.2	Research typology	6
1.3	Research outline	6
2.1	Research data management cycle, 2018	15
2.2	Research data management cycle, 2004	16
2.3	Research data management cycle by UK Data Archive	17
2.4	Research data management cycle by DataONE	18
2.5	Research data management model by DCC	19
2.6	Example of illustration in a research context	21
2.7	Example of photograph in a research context	22
3.1	Generation.Mobi workflow	33
3.2	Questionnaire - Answers to the question 2	34
3.3	Questionnaire - Answers to the question 3	35
3.4	Questionnaire - Answers to the question 4	36
3.5	Questionnaire - Answers to the question 5	36
3.6	Questionnaire - Answers to the question 7	38
3.7	Methodology workflow	40
3.8	Image use by research domains.	41
3.9	Guide by area and frequency.	41
3.10	Instrument of capture by scientific area.	43
3.11	Instrument of capture by frequency of use.	43
3.12	Image edition by research domains.	44
3.13	Image edition by frequency of use.	44
3.14	Elements used by research domains.	46
3.15	Elements relevance by research domains.	46
3.16	Storage format by research domains.	47
3.17	Storage format by frequency of use.	47
3.18	Our methodological approach	52
3.19	Mind-map of the topics examined in the results	52
3.20	Methodological Approach	63
3.21	Flow diagram of study selection	63
4.1	Stages of the elaboration of the metadata model and controlled vocabularies	82
4.2	Screenshot of the form used for the experiment	84
4.3	Example of an image description	99
4.4	Correspondence of labels between the three types of description	100
4.5	Research life cycle	110

5.1	Methodological approach to the new perspective	120
5.2	Simulated work task example of a data deposit	125
5.3	Mockup of the data repository	129

List of Tables

3.1	Significant comparisons between scientific areas	42
3.2	Significant comparisons between rates of image use	43
3.3	Researchers profile	51
3.4	Image management planning cycles identified by researchers	53
3.5	Dimensions of analysis	65
3.6	Images type used per domain and significant comparisons	67
3.7	Instrument type used per domain and significant comparisons	68
3.8	Use and production of images in research domains and significant comparisons .	69
3.9	Collection scenarios per domain and significant comparisons	69
3.10	Image format per domain and significant comparisons	70
3.11	Image editing per domain and significant comparisons	71
3.12	Analysis type per domain and significant comparisons	72
3.13	Collection technique vs analysis type	72
3.14	Programs used per domain and significant comparisons	73
3.15	Image description per domain and significant comparisons	74
3.16	Image type vs Image description	75
3.17	Image deposit per domain and significant comparisons	76
3.18	Licenses description	76
3.19	Licenses per domain and significant comparisons	77
3.20	Image acquisition vs licenses	77
4.1	Image description metadata model elements - before description sessions	86
4.2	Time each participant took to describe images and number of elements used per image	87
4.3	Elements usage during the sessions: average start and end times, average time spend on each element, average time between element the, average number of used characters and percentage of times used	88
4.4	Image description metadata model elements - after description sessions	93
4.5	Comparison of platforms for the infrastructure component	103
4.6	Comparison of platforms for the ingestion component	103
4.7	Comparison for the content organization and control component - part 1	104
4.8	Comparison for the content organization and control component - part 2	105
4.9	Comparison for the metadata component	106
4.10	Comparison for the user interface of component	106
4.11	Comparison for the articulation with other services component	107
4.12	Guidelines for the planning phase	111
4.13	Guidelines for the creation, collection and compilation phase	112
4.14	Guidelines for the processing/analysis phase	113

4.15	Guidelines for the ensure the quality phase	113
4.16	Guidelines for the store phase	114
4.17	Guidelines for the description phase	115
4.18	Guidelines for the ensure the discover/share phase	116
5.1	Profile of interviewees	130
5.2	Existing elements where it would be possible to include images	134
5.3	New elements where it would be possible to include images	135

Acronyms and abbreviations

AAT	Art and Architecture Thesaurus
ABCD	Access to Biological Collection Data
CASRAI	Consortium for Advancing Standards in Research Administration Information
CERIF	Common European Research Information Format
CESSDA	Consortium of European Social Sciences Data Archives
DDI	Data Documentation Initiative
DCC	Digital Curation Centre
DIF	Directory Interchange Format
DMP	Data Management Plan
DOI	Digital Object Identifier
EML	Ecological Metadata Language
EU	European Union
EXIF	Exchangeable Image File Format
FAIR	Findability, Accessibility, Interoperability, and Reusability
FEUP	Faculty of Engineering of University of Porto
GDI	Gestão de Dados de Investigação
GDPR	General Data Protection Regulation
FITS	Flexible Image Transport System
JEITA	Japan Electronics and Information Technology Industries Association
LCTGM	Library of Congress Thesaurus of Graphic Materials
LTER	Long Term Ecological Research Network
NCEAS	National Center for Ecological Analysis and Synthesis
OAIS	Open Archival Information System
OECD	Organisation for Economic Co-operation and Development
PIA	Privacy Impact Assessment
QUDV	Quantities, Units, Dimensions and Values
RDA	Research Data Alliance
RDM	Research Data Management
RQ	Research Question
ROAR	Registry of Open Access Repositories
SQ	Starting Questions
WoS	Web of Science

Chapter 1

Introduction

1.1 Contextualization

Modern science is characterized by the creation, dissemination, and analysis of large amounts of data and the variety of resources for data creation and collection. This new phase of science has given rise to the term data-intensive science, where data collection, curation, and analysis are essential activities for data to have value in the future [14, 19]. The challenges encompass the issues of security, privacy, exchange, and copyright of research data, and increasing the ability to extract knowledge from research data by sharing and reusing it [53]. The creation and development of software and hardware, of institutions, infrastructures, and specialized people, as well as of guidelines and policies directed at the management of research data are aimed at mitigating the challenges to their management.

The technological and scientific revolution made it possible to highlight new methods, instruments, and tools. These technological and scientific revolution have led to an increase in the volume and complexity of research data. Together with the increase in computational and digital storage capacity, data collection is increasingly intensive, causing changes and challenges to the way data is stored, preserved, accessed, and shared within the context of scientific activity [100]. The widespread use of information technology in all areas of research has enhanced the openness and free exchange of ideas, information, and knowledge, as well as the emergence of new ways to collaborate. Therefore, access to research data in a responsible and efficient way should be seen as one of the opportunities and benefits brought by information technology [89].

Research data management (RDM) involves a set of practices directed toward the information collected and produced during the research process. It is a structured process aligned with the research context and scientific area [65], it is something that takes place before, during, and after the research process. It encompasses planning data management, documenting and organizing data, improving analysis, as well as ensuring the security and proper storage of the data produced in the research. Data management combines different practices with the goal of making data easily locatable, less likely to be lost, and more usable, both during research and in the future [20].

Research data management means that various resources must be properly planned, captured,

processed, organized, described, preserved, stored, and shared. The typology of these resources can be very different. Among them are images.

Capable of safeguarding the most diverse moments and manifestations, image occupied a prominent place at the time of the appearance of television and other artistic demonstrations, such as cinema. This type of documentary is now seen as an integral part of society since it is an excellent mechanism for establishing communication. It can hold several interpretative possibilities, depending on the function attributed to it. These functions can be business, emotional, news, stimulus, expression, memory, heritage, or scientific.

Images have several functionalities and their capacity to register actions and information is one of the most important. It is impossible to dissociate the image from its ability to carry materiality and distinct communication and expression resources, for these are the characteristics that bring it closer to other documentary types, even with its particularities. However, the role of photography in science is still not entirely clear, although the scientific character given to the photographic document dates back to the 19th century when it began to be used in Astronomy studies.

1.2 Motivation

Digital technologies have taken on a prominent role in the information society, favoring the emergence of new trends and making informational resources highly valuable. The image and its production have also changed been improved, with capture, reach, and dissemination. Allied to this is the notorious evolution in storage capacity and the growing incentive to produce and store more images, whether in a professional, academic or personal context.

The motivation for this doctoral work lies precisely in these trends and challenges associated with images and their involvement with research data management. In the field of Information Science, particularly in the management of research data, images tend to be an ally for researchers to capture evidence and record facts associated with research. In many cases, they are included in their datasets and can be organized and described in information systems so that, when necessary, it can be reused.

In line with what was said previously, it is true that new technologies, from image capture devices to image analysis and processing programs, have promoted the production of images. It's also true that the overwhelming majority of images produced today are digital, as opposed to the analog images that were so prevalent 20 years ago. This is also evident in the context of research: digital images make up the majority of the imagery universe. However, the role of analog images and their use and production cannot be dismissed. This work has made it possible to understand that analog images, although fewer in number, continue to exist, standing out in some specific areas. For this reason, this work has considered both digital and analog images, always safeguarding their differences and recommending specific guidelines for their characteristics. Digital is more challenging because of all the editing possibilities available today. However, analog images also have characteristics that make them stand out, such as, for example, the challenges in disseminating and sharing these images.

This work is motivated by images in research and seeks to understand how this can support in scientific research, both as data, due to its ability to be easily represented and obtained, and as metadata, through its use as a tool capable of faithfully describing a set of data and promoting its proper contextualization and interpretation. Furthermore, the study of data curation was an additional motivation, through the work of professionals with a specialized profile in data management. With this study it is possible to understand how data curation can promote better management of scientific images, raise awareness and train researchers who use and produce images, and create standards, policies and guidelines for the implementation of good practices. This is where the so-called social and human layer comes in.

The lack of knowledge about image management in research, on the part of the scientific community in general, also motivates this study. Images are fundamental pieces in the research process, facilitating interpretation and analysis. For this reason, it is essential to know how researchers deal with images and to make them aware of the value of images, motivating their adequate management.

The work represented here, although extending to various components of research data management, has always taken into account those guiding principles for the management and administration of scientific data. The four components highlighted by the FAIR principles¹ (findability, accessibility, interoperability and reusability) underpin what are the fundamental aspects of research data management and therefore guide all the work carried out to ensure compliance.

Published in 2016, the FAIR Principles are intended to provide a guide for research data management practices. The FAIR Principles are divided into four: Findable, Accessible, Interoperable, and Reusable, and are framed within the Open Science ecosystem.

In a nutshell, the Findable principle aims to make data properly locatable through various practices, such as assigning persistent unique identifiers or describing data with personalized metadata. The Accessible principle indicates the need to make data accessible to others, for example through communication protocols in which the metadata remains accessible or even the data is completely open access and free of charge. The principle of interoperability addresses the importance of data being effectively interoperable, whether through vocabularies based on the FAIR principles, the use of formal, accessible, and widely applicable language in the field of knowledge representation, or through data or metadata that includes qualified references. Finally, the principle of reusability reinforces the need to ensure that data, regardless of its type, is reused, through practices such as associating data or metadata with its origin or even providing an accessible license for data use. The FAIR Principles will be described in greater detail in the literature review chapter (Chapter 2), specifically in the Section 2.1.2.

1.3 Objectives

It is predictable that when analyzing this work one will wonder about the importance of highlighting images in the context of data management. Shouldn't all typologies be treated in a standard

¹<https://www.go-fair.org/fair-principles/>

way? The answer is clearly obtained with this work. Each type of data carries unique characteristics that should not be underestimated. As is known, most of the data produced presents a textual or numerical typology. However, the exponential growth in the use and production of images, previously mentioned, means that images are increasingly noticeable in the scientific context. This evidence is often not widely discussed, as it is part of everyday investigation, however, it does happen.

It is based on these ideas that the thesis underlying this work emerges: Images are research data with their own specificities that can be explored in information data management strategies and also studied as metadata for other research data.

With the results of the in-depth research on the subject of images in the context of research data management and following the theoretical approaches, we intend to verify the idea that underpins the aforementioned thesis. To do this, it was essential to work with arguments and facts, with relevant, reliable data based on trustworthy sources of information, to reinforce and justify the development of ideas.

This work seeks to understand if and how researchers use images in the research process and how we can help them take advantage of this type of data in the research process. As a second objective, it will investigate whether and how automatic image processing tools, combined with data curation, can promote better description and interpretation of data.

In order to achieve the specific objectives mentioned above, a set of three general objectives was defined:

- Raise awareness about the use of images in research, showing their relevance, suggesting consistent and standardized practices;
- Show how RDM practices can be adapted to the case of images as research data;
- Show that the use of images as metadata can improve the description and interpretation of data.

The research questions developed to guide the entire research activity are as follows:

- RQ1. How does the production and use of images occur in science?
- RQ2. How can RDM practices be adapted to the case of images as research data?
- RQ3. Can images improve the description and interpretation of research data through their use as metadata?

In the course of the work, it became clear that all the questions were clearly related to each other, making it possible to study the initially defined thesis. Each of the research questions is directly related to one of the main chapters of this work (RQ1 - Chapter 3; RQ2 - Chapter 4; RQ3 - Chapter 5). At the appropriate time, this relation will be mentioned again.

1.4 Methodological approach

For the development of this work, we draw a general scheme of objectives to understand the stages of the work. We defined three stages to fulfill the specified objectives. Each of them has a set of specific tasks, as can be seen in Figure 1.1.

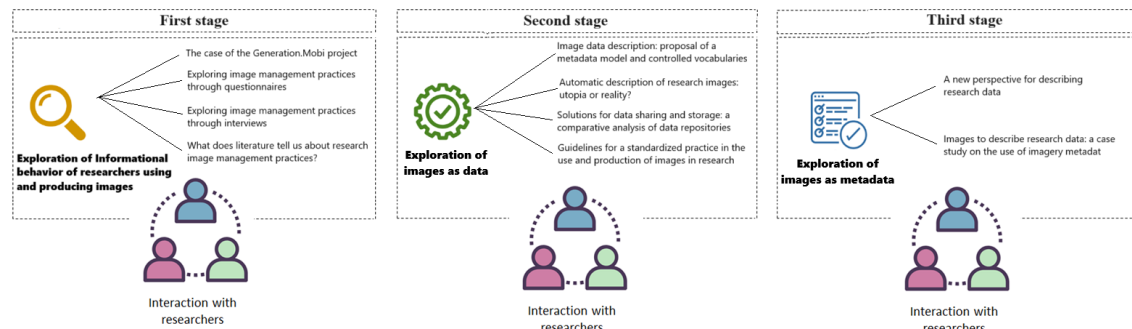


Figure 1.1: Methodological approach to doctoral work

Source: Own authorship

In the first stage, the use and production of images in science was explored through several exploratory studies. We evaluate the information behavior of image-producing researchers and understand the life cycle and importance of images in their projects. The objective was to understand the role that researchers attribute to images in different scientific areas and to understand what their habits are. To achieve this objective, four studies were carried out, using various methodological approaches.

In the second stage, within the scope of RDM, the aim is to analyze research data management practices in a specific scenario: image as data. The focus is on defining a metadata model for describing images in image format, as well as determining a set of controlled vocabularies for some of the model's elements. Still in this scenario, issues related to the automatic description of images are studied, as well as technological solutions for publishing data, such as repositories and guidelines for standardizing practices for the use and production of images in research.

In the third stage, a new perspective on using images to describe data is presented through a conceptual study and a case study. The latter served as an evaluation of the new perspective presented in the first study. This assessment is decisive for the reliability of the results and for giving strength to the proposed solutions.

To achieve our objectives, we initially decided that this study would involve mostly a qualitative approach (although some quantitative approaches are included), as the objective is to understand and document reality. Secondly, it was determined that this work would take place through non-documentary empirical studies since the objective is to identify evidence through experience, that is, the study has a documentary and empirical character. Third, it was essential to define the nature of the experiences. This work fits into a quasi-experimental typology, since there is a manipulation of the experience, but not of the sample. For this stage, from a methodological point of view, specific instruments were used during the experiments. With this approach, it will be

possible to evaluate and compare results, as well as confirm or refute the initial thesis. In addition, with this research typology, it will be possible to develop critical thinking, define a preliminary diagnosis, and gain knowledge. Figure 1.2 shows the design of the research typology that sustains the work.

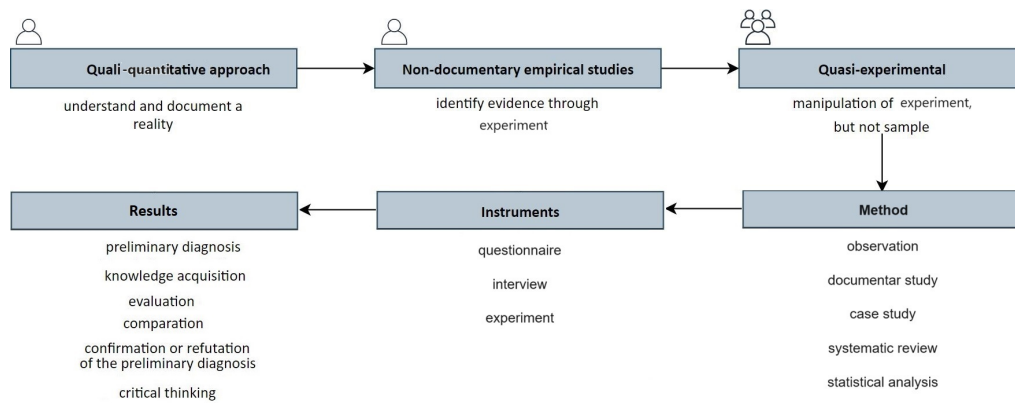


Figure 1.2: Research typology

Source: Own authorship

To implement the methodological strategy and put the action plan into practice, the analysis tools that make up each of the selected methods were defined. This methodology, using different but complementary methods, makes it possible to select the most precise and targeted approach to the problem identified, to obtain the best results. Figure 1.3 shows the methods and analysis tools used in each of the stages defined in Figure 1.1. Although each stage has its own methodological approach, they are all related, as if they were a cycle.

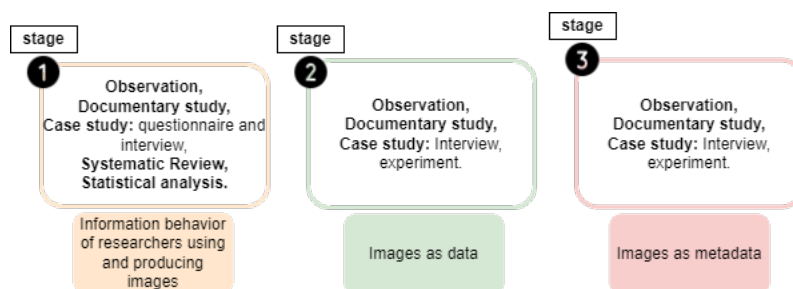


Figure 1.3: Research outline

Source: Own authorship

This is the general methodological approach of this work. However, in each chapter of this document, the methodology that underpinned the work will be reported in detail.

1.5 Publications and contributions

This work allowed us to better understand the unique characteristics of the images. By unique we do not mean more or less important, just unique and, therefore, important. Furthermore, images

constitute, in this work, a new and innovative role, in the sense that it is proposed that they assume the role of metadata. This scenario has not been explored before and we present here as a new perspective. For all studies that were carried out in the context of this work, it was a concern that they were presented to the scientific community. This new perspective followed the same premise, as it was considered that it represents not only a challenge but also an opportunity and a responsibility to advance and diversify theories in the field of research.

When exploring this topic of analysis, it was noticed that the literature does not adequately explore the issues of images combined with research data management, much less the possibility of them assuming the role of metadata. This fact triggered a greater desire to explore the topic, not only to open new paths and ideas but also to provide the scientific community with consistent literature in this field.

During this work, 12 papers were published, including 8 in scientific conference proceedings (66,67%) and 4 in scientific journals (33,33%). In addition to these articles, 6 datasets were also published in INESC TEC's Data Repository. Finally, 3 posters were presented and communicated at scientific conferences. Of all these 21 contributions, scientific articles account for 57,14%, datasets for 28,58%, and posters for 14,28%.

Papers

Journal papers

- Joana Rodrigues and Carla Teixeira Lopes. **Images as Metadata: A New Perspective for Describing Research Data**. 2023. Journal of Library Metadata [114]
- Joana Rodrigues and Carla Teixeira Lopes. **Research Image Management Practices Reported by Scientific Literature: An Analysis by Research Domain**. 2023. Open Information Science [115]
- João Aguiar Castro, Joana Rodrigues, Paula Mena Matos, Célia Sales and Cristina Ribeiro. **Getting in touch with metadata: a DDI subset for FAIR metadata production in clinical psychology**. 2023. IASSIST Quarterly [26]
- Joana Rodrigues and Carla Teixeira Lopes. **Describing Data in Image Format: Proposal of a Metadata Model and Controlled Vocabularies**. 2022. Journal of Library Metadata [110]

Conference papers

- Joana Rodrigues and Carla Teixeira Lopes. **Images to describe research data: a case study on the use of imagery metadata**. 2023. 17th International Conference on Metadata and Semantics Research. in publication

- Joana Rodrigues and Carla Teixeira Lopes. **Automatic description of research images: utopia or reality?** 2023. 17th International Conference on Metadata and Semantics Research. in publication
- Joana Rodrigues and Carla Teixeira Lopes. **Solutions for Data Sharing and Storage: A Comparative Analysis of Data Repositories.** 2022. 26th International Conference on Theory and Practice of Digital Libraries [109]
- Joana Rodrigues and Carla Teixeira Lopes. **Research Data Management in the Image Lifecycle: A Study of Current Behaviors.** 2022. International Conference on Research Challenges in Information Science [108]
- Miguel Fernandes, Joana Rodrigues and Carla Teixeira Lopes. **Management of Research Data in Image Format: An Exploratory Study on Current Practices.** 2020. 24th International Conference on Theory and Practice of Digital Libraries [48]
- Joana Rodrigues and Carla Teixeira Lopes. **A utilização e tratamento de documentos fotográficos em projetos de investigação o caso do Generation.Mobi.** 2019. IV Congreso ISKO España-Portugal [106]
- Joana Rodrigues, João Castro and Cristina Ribeiro. **Hands-On Data Publishing with Researchers: Five Experiments with Metadata in Multiple Domains.** 2019. 15th Italian Research Conference On Digital Libraries [105]
- Joana Rodrigues. **Photography as data and metadata: description requirements for information retrieval and semantic interoperability.** 2019. Bulletin of IEEE Technical Committee on Digital Libraries - Joint Conference on Digital Libraries [107]

Datasets

- Joana Rodrigues, Pedro Oliveira and Carla Teixeira Lopes. **Images annotated according to their content: a study on the description of data in image format in multiple domains.** 2023. INESC TEC Data Repository [117]
- Joana Rodrigues and Carla Teixeira Lopes. **Images to describe data: Summary of interview transcripts to assess the role of imagery metadata in data description.** 2023. INESC TEC Data Repository [113]
- Joana Rodrigues and Carla Teixeira Lopes. **Research image management practices reported by scientific literature: Studies that explore the use and production of images in research.** 2023. INESC TEC Data Repository [116]
- Joana Rodrigues and Carla Teixeira Lopes. **Research Data Management behaviors in the image lifecycle.** 2022. INESC TEC Data Repository [111]

- Joana Rodrigues and Carla Teixeira Lopes. **Describing data in image format: Proposal of a metadata model and controlled vocabularies**. 2022. INESC TEC Data Repository [112]
- Joana Rodrigues and Carla Teixeira Lopes. **Research data management in image format - Survey**. 2020. INESC TEC Data Repository [49]

Posters

- Joana Rodrigues and Carla Teixeira Lopes. **Solutions for Data Sharing and Storage: A Comparative Analysis of Data Repositories**. 2022. International Conference on Theory and Practice of Digital Libraries
- Maria Nunes, Joana Rodrigues and Carla Teixeira Lopes. **Research Data Management: an approach according to GDPR**. 2019 IV Congresso ISKO Espanha-Portugal
- Joana Rodrigues and Carla Teixeira Lopes. **The use of photographs in research data management: the case of the generation.mobi project**. 2019. 13th Research Data Alliance Plenary Meeting

1.6 Structure of the dissertation

This work is divided into six chapters, this introduction being the first of them. Next, the literature review aims to provide all the theoretical and conceptual material that supports the various studies carried out.

The third, fourth, and fifth chapters present the three most practical components of the work, based on several properly structured studies, strategically defined and conducted accordingly. The third chapter addresses the issue of the informational behavior of researchers using and producing images. It was essential to draw the current panorama regarding this topic, to understand current practices and challenges, to better guide the next steps. The fourth chapter addresses the issue of images as data, going through all its facets and phases in the context of research, namely, in the final phase of the chapter, the presentation of guidelines for the use and production of images in a research context. The fifth chapter, in turn, explores the new perspective of images as metadata, to promote the capabilities of images and improve the procedures inherent to research data management.

The sixth chapter aims to conclude the work carried out, namely discussing the fulfillment of previously established objectives and research questions, and looking at future work. References and appendices are presented at the end of the document.

Chapter 2

Literature review

This chapter aims to offer a conceptual basis that supports all the work developed. The theoretical foundations, corroborated by valid sources of information, are a fundamental contribution to the elaboration of the steps to follow and the definition of assumptions and hypotheses. This literature review will address the topics of research data management, the use of images in research, the description of research data, and the challenges of image description. Each topic has several subtopics.

2.1 Research data management

Research Data Management (RDM) is gaining the attention of researchers, who often ask for support in the process [23]. Tasks such as adequate data organization, rigorous description, storage, and sharing allow data to keep their meaning and facilitate their interpretation [30]. Data management is also fundamental in the communication between researchers and the scientific community, where images can play a central role in the "teaching-learning" process [66].

2.1.1 Research data definition

According to the Open Archival Information System (OAIS) reference model cited by the Digital Curation Centre (DCC)¹, data are reinterpretable representations of information, formalized in a way that can be understood, processed, or communicated, such as sequences of bits, a table of numbers, the characters on a page, the sound record, or a moon rock specimen.

The Organisation for Economic Co-operation and Development (OECD) document "Principles and guidelines for access to research data from public funding" [89] defines research data as the factual records used as primary sources in scientific research, accepted by the scientific community and indispensable for validating research results. These records can be textual, numerical, image, or sound records. The report "Making open science a reality" [90] adds that research data are collected, and produced in a wide variety of formats from digital spreadsheets and databases

¹<https://www.dcc.ac.uk/about/digital-curation/glossary>

to compilations of questionnaires, images and objects. It also states that their consultation and use may require, in many cases, the use of specific computer programs, software, among others.

In the same line of thought, the Consortium for Advancing Standards in Research Administration Information (2015) [21] states that research data refer to data, observations, or numerical, textual, image, and video records, in a wide variety of digital formats, that serve as evidence of the research process, used as primary sources to support the research and/or accepted by the scientific community as necessary to validate the research results. With respect to their nature, research data can be experimental, observational, operational, third-party, public, processed, monitoring, or reused data.

Research data can be distinguished into primary and secondary data, depending on the degree of processing they undergo. Primary data are data that have not undergone any kind of transformation or processing, e.g., data collected through an instrument without any processing. Secondary data result from the interpretation, processing, or transformation of primary data, such as the application of statistical models to data collected by an instrument [52]. According to their origin, research data can be classified as follows [90, 16]:

- **Observational data** - data resulting from recording historical information and phenomena that only happen once, through telescopes, satellites, sensors, questionnaires, and other data collection instruments. In many cases, the data cannot be replicated, so its preservation is advisable. Some examples of this type of data are photographs of a supernova and direct observation of ocean temperature on a specific date;
- **Experimental data** - data collected through controlled experiments, such as clinical and biomedical trials, pharmaceutical testing, and others. Particularly important to preserve experimental data whose replication conditions are not easy to reproduce, costly monetarily, or for ethical reasons. Identification of patterns in gene expression is an example of such data;
- **Computational data** - data created from large-scale computer simulations. It is important to preserve these data because although they are recreated by repeating the simulation, the time available to use the computer may be limited and the data can be used as a basis for analysis, visualization, or data mining. Sometimes it is more relevant to preserve the model that gave rise to the data when resources exist that allow the simulation to be easily reproduced. The results of running a computational model are an example of computational data;
- **Reference data** - highly curated data desired by several scientific areas, these data are created for various purposes, ranging from human genome mapping and protein documentation to longitudinal data on economic and social aspects.

2.1.2 FAIR principles and open data

Closely linked to research data management are the Findable, Accessible, Interoperable, Reusable (FAIR), and Open Data principles. Although they are different concepts, when applied together

they strengthen, emphasize, and enrich the goals of each concept. According to Higman, Bangert, and Jones [65], data management is the foundation that enables FAIR and open science principles, if data is not created and managed correctly from the beginning of the research, its sharing will not take place efficiently, hindering the application of FAIR and open data concepts.

The concept of FAIR data emerged in 2014 at a Lorentz workshop, driven by the challenge of knowledge discovery in data-intensive science. FAIR Principles come to assist humans and machines to discover, access, integrate and analyze data appropriate for research tasks as well as for associated workflows and algorithms [51]. The FAIR Principles define the properties that contemporary data resources, tools, vocabularies, and infrastructures should have to support data discovery and access. They are not a specification that suggests the technology, standards, and mode of implementation. The high-level FAIR Principles precede implementation, acting as a guide that assists data curators in choosing the solution that best suits their research data [136].

The concept of open data translates into data that can be accessed and reused without any legal or technical restrictions, although the use of the data is subject to the legal restrictions that apply to them [90]. The benefits of open data are reflected in the development of new products and technologies. At the government level, encourages greater transparency in decision-making, as well as in public policies, in the creation of new jobs, and in the attraction of investment [124]. Applied to scientific research, the goal is to make research data publicly available for access and reuse with as few restrictions as possible [65].

Data sharing allows the reuse of data by third parties, this can be done through private or public channels, such as depositing the data in public data collections, and its target audience can range from researchers in a specific scientific branch to the general public. The degree of usefulness, credibility, and value of research data can vary greatly, data can be highly structured and described, as well as unstructured and accompanied by little descriptive documentation [18].

In an international study on data management practices [128], it was found that researchers are positive about sharing all or part of their data within certain conditions and restrictions regarding access and use. Some ways to encourage sharing among researchers include formally citing the data, obtaining copies of the articles (duly following the license and copyright conditions), and learning from the products and publications that are reused from third. Researchers also pointed out that their ability to answer certain scientific questions, such as global warming and climate change, was limited by a lack of access to data created by other researchers.

2.1.3 Benefits and challenges in sharing research data

Sharing and access to research data avoids the repetition of investigations and the costs inherent to their realization. In this sense, several scientific funding entities have produced guidelines and regulations for the dissemination of research data. Science policy promotes data sharing and open data. The arguments for sharing data are vast, such as reproducing research, making public assets available to the public, leveraging investments in research and promoting research and innovation [94].

The reuse of research data allows the verification, validation, and detection of inaccuracies, as well as the testing of new hypotheses not tested in the original research, serving as the basis for new projects. On the other hand, the evaluation and review of applications for research grants become simpler processes, due to the easier access to the data produced by researchers and their impact on scientific development, i.e., the quality and efficiency of research. The dissemination of data produced in the course of research can be beneficial in increasing or enhancing the recognition and credibility of researchers, institutions involved, and the repositories that help preserve and disseminate the data [100, 46].

The ability to access the data that serve as research evidence increases transparency, identification of fraud, and deterrence of plagiarism [100]. Transparency is achieved through the combination of research behaviors, practices, and processes associated with their reproducibility, serves to demonstrate the rigor of research methodologies and protocols, and consequently enhances public opinion about the quality, trustworthiness, and integrity of the results, conclusions, and theories drawn from research [77].

However, sharing research data and its reuse, as well as interpretation by others, is a difficult task due to several layers of complexity related to the nature of data and research, innovation, research scholarship, privacy, data protection, motivation, and incentives for researchers, public policy, intellectual property, and economic return [18]. Another facet negative side of data sharing is that data is misused or misunderstood, which affects the reproduction and pace of research, and on the other hand, leaves available data that can lead to scientific error [19, 53].

The sharing and reuse of large amounts of data depend on the existence of data sharing infrastructures such as databases, repositories, and data centers that enable storage, curation, and continued access. In turn, data to be retrieved correctly requires standardized formats and meta-data [37].

2.1.4 Research data management cycle

Research data management cycle models represent explicit abstractions and simplifications of the elements and processes that make up data management, consists of a simple, understandable, and organic way to visualize the different phases of research data management through a descriptive model. Models can range from abstractions of the research process to detailed models of data creation, reuse, or curation applied in a specific context, such as a university or institution, or generalized to any research [30]. The cycle seeks to define the concepts and key moments in the planning and implementation of research data management, these concepts and key phases may change based on the scientific area, the type of data, among other factors [52].

This approach becomes necessary as digital resources are fragile and susceptible to change due to technological development during their life cycle. Thus, the existence or lack of activities during their cycle directly influences the ability to manage and preserve the resources, and the reuse of digital materials is only possible if there are curation steps that safeguard their integrity and authenticity [95].

Research data management involves organizing data from the moment it enters the research cycle to the archiving and dissemination of the results. It aims to ensure the quality of the results obtained as well as to contribute to a new and innovative investigative process from existing information. Curation and preservation of research data are two concepts closely linked to research data management. Curation is about ensuring that project results are eligible to be archived and that they constitute important research assets that can be reused. While preservation focuses on the long-term accessibility and reuse of resources [134].

The cycle of research data management should be synchronized with the research process, so it is important to know how it works. The research process usually begins with the emergence of an idea or hypothesis that will be developed and worked on to give rise to a product, in most cases, a scientific article. Throughout the process, equipment, methodologies, and steps are used to advance the research process, each of these steps is supported by data that help develop the research and support decision-making. Pryor [97] presents a data-centric research model, where data collection begins with formulating hypotheses to analyze what has already been done and select data for the investigation. Next, the research will need to collect published data to re-analyze the data in new contexts or to compare with the data produced in the course of the research, this phase is the main source of new research data. All phases of the cycle use or generate data at some point, publication and reuse after the data are interpreted, analyzed, and synthesized are essential to the development of knowledge and new research opportunities. Image 2.1 presents the life cycle of research data management proposed by Pryor [97].

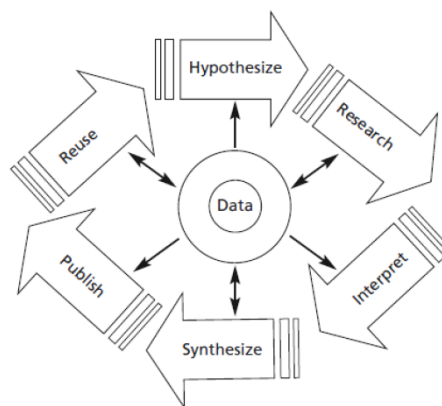


Figure 2.1: Research data management cycle, 2018

Source: Pryor, 2018 [97]

2.1.4.1 Research data lifecycles

In the context of the research process, data go through a life cycle that reflects the different stages that research data go through. Several cycles attempt to represent these dynamics.

One such cycle comes from the Data Documentation Initiative (DDI)² (2.2), which presents a life cycle designed to support the programming and archiving activities typical of the application of the metadata model it has developed. The model focuses on data from its development, encoding, preservation, dissemination, and analysis. The model aims to capture information about the creation of the data in a given context and to have information about the structures and relationships that allow the discovery, extraction, and manipulation of the data [125]. The cycle begins with research planning, it involves planning what data is needed, the methods of collection, and how it will be processed, analyzed, and used. Data collection and processing consists of capturing and/or creating data, followed by processing and analyzing it to answer the research questions. The archiving step adds value to the data over time through metadata assignment, preservation, and long-term access. The data are then distributed either directly or through a library or data archive and may be published in books, journals, web pages, or other services. The analysis phase consists of using the data for purposes other than that originally conceptualized. In the same vein, reuse involves using the data in a different context, such as restructuring the data, integrating it with data from similar data sets, and producing pedagogical materials [9].

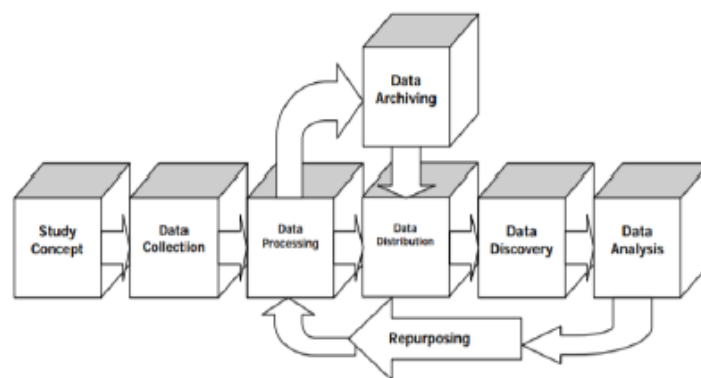


Figure 2.2: Research data management cycle, 2004

Source: Structural Reform Group by DDI Alliance

The life cycle proposed by the UK Data Archive³ (2.3) consists of planning, collection, processing and analysis, publication and sharing, preservation and reuse of research data.

The planning phase reflects on data management, collection methods, protocols and standards, consent to sharing, and the data sources to be used. The collection involves the gathering of data and their metadata if possible. Processing and analysis encompass the actions that are taken to process and analyze the data during the research, this phase incorporates the following actions: data entry, digitization, transcription, and translation; verification, validation; cleaning and anonymization; data derivation; data description and documentation; data management and storage; analysis and interpretation; production of results and citation of data sources. In data publishing and sharing, data copyrights are established, user documentation and discovery metadata are produced,

²<https://ddialliance.org/>

³<https://www.data-archive.ac.uk/>

and access control and data publication, sharing, and promotion are defined. Preservation actions encompass migrating data to the most appropriate format, storing and performing backups, creating preservation documentation, and preserving and curating data. The reuse of data comprises conducting secondary analysis, scrutinizing and reviewing research, and using the data in learning resources.

The UK data Archive cycle was adapted by Vejvoda, Burpee, and Lackie [133] to the life cycle of images in research. The cycle aims to provide a conceptualization of the management of image collections throughout the research life cycle and recommend some practices to consider in their management.



Figure 2.3: Research data management cycle by UK Data Archive

Source: UK Data Archive, 2019

The life cycle proposed by DataONE⁴ (2.4) addresses the life cycle of data in the research context, in this model data goes through the phases of planning, assurance, description, preservation, discovery, integration, and analysis. Planning involves the creation of a data management plan, which defines how the data is going to be handled. Data collection involves gathering data through various equipment and sources and organizing it into documents or databases. The data are then reviewed under standards, regulations, and quality control processes, such as checking for duplicate data. The description phase involves adding metadata to the data to make it easier to understand and reuse. Preservation refers to depositing the data in data centers or repositories and protecting them against format degradation and obsolescence. Through the previously aggregated metadata and deposit in repositories, data can be discovered and accessed. The data can then be

⁴<https://www.dataone.org/>

integrated with other datasets, taking into account their semantic and methodological differences without losing their origin. Finally, the data are analyzed to meet research needs [9].

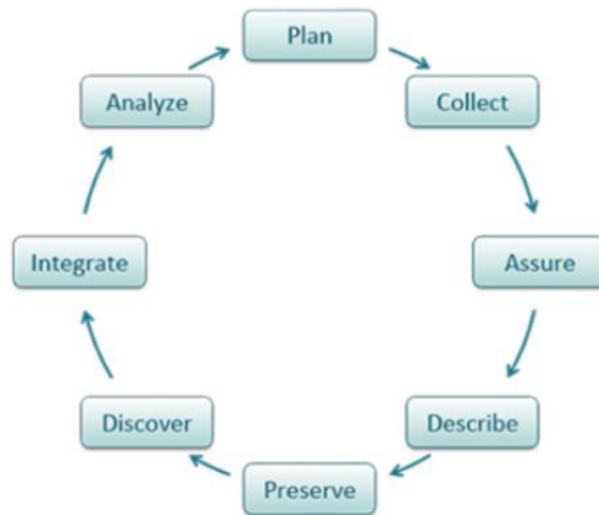


Figure 2.4: Research data management cycle by DataONE

Source: DataONE, 2019

2.1.4.2 Research data management model

The management model proposed by the Digital Curation Centre (DCC) (Figure 2.5) aims to provide an approach for managing digital materials, for their curation and preservation from conceptualization to reuse and long-term preservation. The model identifies the curation actions, applicable at different levels of granularity, that occur throughout the cycle, those that must occur sequentially, and those that are only performed occasionally [64]. The actions performed throughout the cycle encompass information description and representation, preservation planning, community participation and observation, and curation and preservation. The description and representation of information include the use of standards to ensure adequate description and control over time and includes the representation of information needed to understand and present digital resources and associated metadata. Preservation planning includes the management and administration plans for all actions in the cycle. Community participation and observation concerns the monitoring of community activities and participation in the creation of standards, tools, and programs. Curation and preservation involve raising awareness and carrying out planned curation and preservation actions [64].

The activities performed sequentially encompass conceptualization, creation and reception, arrangement and selection, ingestion, preservation actions, storage, access, use and reuse, and transformation. Conceptualization addresses the design and planning of data creation, including methods of capture and forms of storage. Creation and reception involve the creation of data

according to defined policies as well as the capture of administrative, descriptive, structural, and technical metadata. The evaluation and selection of data for curation and long-term preservation is part of the disposition and selection phase. They are then transferred and securely stored in an archive, repository, data center, or other custodial type, according to existing policies and legal requirements. Preservation actions include data cleansing, validation, allocation of preservation metadata, representation of information, and verification of the appropriateness of data structure and file type. Ensuring that data is accessible to users is part of the access, use, and reuse phase, where access control and authentication procedures are defined. The transformation phase involves creating new data from the originals [64].

Occasional activities incorporate the destruction of data or transfer to another type of storage of data that has not been selected for long-term preservation and curation. The recovery, re-evaluation, and selection activities of the data that previously failed the evaluation procedures. And lastly, the migration of data to different formats as storage needs or obsolescence of the technology used [64].

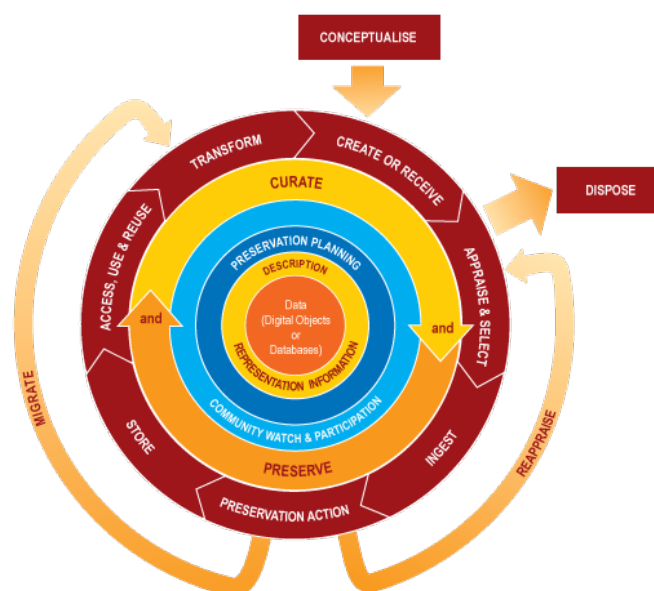


Figure 2.5: Research data management model by DCC

Source: Digital Curation Centre, 2019

2.2 Use of images in research

The development of information and communication technologies, triggered a new era in visual communication. The rapid development of processing capacity, image management, particularly digital images, and the ease of replication and dissemination, have increased the access and value of image collections [44]. This section aims to frame the use of images as data in the context of research, it begins by addressing the nature of imagery resources, then moves on to the use of

imagery as scientific evidence and its role in science. Due to the complexity of the photographic image, its increased value, and its use relative to other types of images, photography is briefly discussed in the next sections.

2.2.1 Imagery resources

Image is something that portrays, it is a form of representation of which photography, video, drawing, and painting are part, among many others. The major difference that distinguishes photographic images from non-photographic images (drawing and painting) is their method of production, photography is obtained through a mechanical process while drawing and painting traditionally do not [2].

The image contains information or meaning differently than text, while text means through conventionalized and arbitrary symbols, the image means through the representation and likeness of objects as they are. The fact that the image is used in conjunction with text suggests that the image carries within it information different from text, managing to convey things that text cannot [74].

Roland Barthes [13] states that photography is always linked to its referent, that is, to what it represents, it can be the object of three practices, doing, experiencing, and looking, and have various functions, inform, represent, surprise, give meaning and provoke desire. Technically, it is at a crossroads between processes of chemical origin, the action of light on certain substances, and processes of physical origin, image formation in an optical device. Boccato and Fujita [17] say that photography is a document that transmits information in analog or digital support through a technique, with the ability to record a past or present moment. The image is representative, referential, aesthetic, objective, subjective, synthetic, emotive, and real, in the case of the latter, because it documents. In the same vein, Meskin and Cohen [82] state that photography is an important source of information, typically providing information about the visual characteristics of what it depicts.

The photographic document as an informational element can be used to meet information needs in research and/or other contexts, thus it follows the information cycle that goes through editing, selection, acquisition, technical processing, storage, organization, dissemination, retrieval, and use [17].

Regarding images used in the research context, Dondero [38] revisits the semiotics of Charles Sanders Pierce to state that scientific images always contain three dimensions, index dimension, iconic dimension, and symbolic dimension. The indexing dimension perpetuates its context and traces of production. The iconic dimension refers to the stabilization of the image forms of scientific objects, it consists of the transformation of the traces of phenomena and data into concrete scientific objects through visualization. Finally, the symbolic dimension allows the image to be considered as the negotiated product of a negotiated communication practice, it concerns the fact that we cannot study and understand the image only on its basis. One has to take into account its production goals, the disciplinary practices of the scientific field, and the relations between the institution, policies, and discipline.

2.2.2 The image as scientific evidence

When we talk about research data, a dictionary in the area of RDM tells us that they are "Facts, measurements, recordings, records, or observations about the world collected by scientists and others, with a minimum of contextual interpretation. Data may be in any format or medium taking the form of writings, notes, numbers, symbols, text, images, films, video, sound recordings (...)" [21]. In its turn, an image is the visual representation of something. It can be captured through various instruments or obtained through manual techniques (such as drawings, paintings, engravings, and illustrations). Images can be in an analog or digital format. Examples of images are photographs, microscopic images, medical images (e.g., x-ray images), paintings, illustrations, videos (moving images), computer-made images, engravings, graphics, drawings, and maps.

The image provides evidence about the existence of a certain object and its characteristics only when it is produced by a reliable process, that is, the process that connects the image and the object must be able to create true perceptions about the object. The process must correctly reproduce the information about the object and do so representatively. In addition to the need for the image production process to be reliable, the images must also be informationally rich. An image only provides information about the characteristics of an object when there is information necessary to perceive those characteristics. Photography is the richest type of image because they are currently produced through processes that ensure a high degree of detail [2]. In the following figure (2.6), we can see an example of a photograph obtained in the context of research.

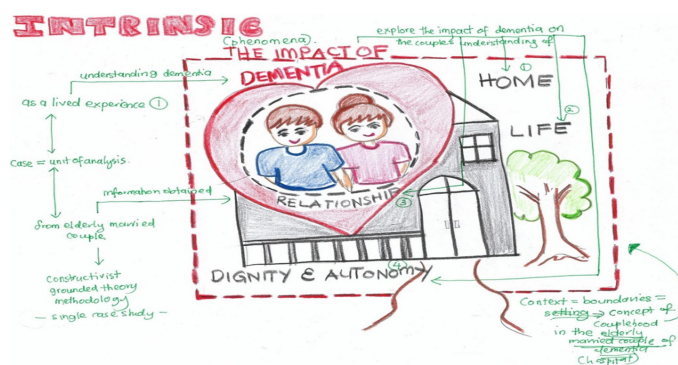


Figure 2.6: Example of illustration in a research context

Source: Janis, 2020 [67]

In the 19th century, photography, driven by positivism, served as scientific proof, as it was perceived as a replica of reality, an imitation of its object, whose neutrality was achieved through the machinic and technical mode of the object. The photographer's role was to understand the context of the photographic act and to photograph the object to create images that were understandable, clear, and identical to the object. Because the laws of seriousness and honesty that existed in a research context were applied, photography as a document was not shaken by photographic simulations [121].

Photography revolutionized the observations made by scientists by increasing what could be observed, as well as by increasing the quality of recording the observations. Before photography, observations were recorded manually by the scientist, which took years of training to produce adequate records [135]. However, the mere detection does not immediately reveal its relevance to the investigation, the image produced by the camera has to be analyzed by the researcher in light of the theory and methods that guide the investigation, and it has to be demonstrated that the image and the process of obtaining it are logically independent of what is intended to be proven. The photograph represents a concrete object, that is, it naturally does not represent something that is not there, except when it is manipulated to obtain a certain objective. Thus, when the use of photography as a detection and recording tool is appropriate and leaving fraud aside, one can consider the photographic process to be reliable and liable to be used as evidence [87]. In the following figure (2.7), we can see an example of a photograph obtained in the context of research.

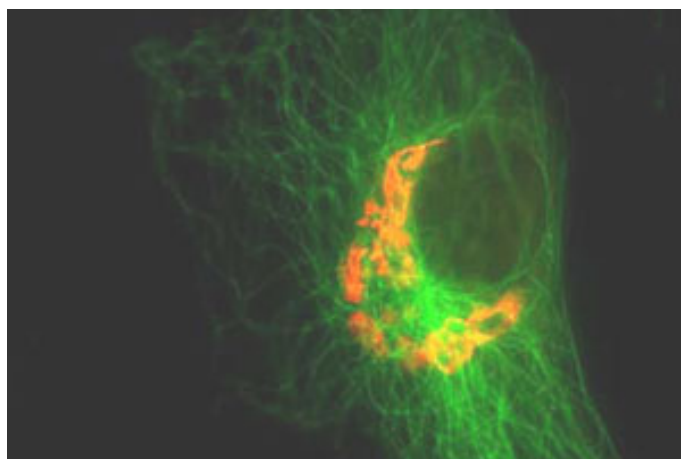


Figure 2.7: Example of photograph in a research context

*Source: NIH Cell Biology Interest Group*⁵

For the image to be considered valid and successful it must constitute a commensurable result of its production process, and take into account the assumptions, expectations, and necessity for research [38].

2.3 Description of research data

Describing research data through highly structured metadata is sometimes seen as difficult to achieve, yet it needs to be highly structured for data reuse and sharing. Metadata comprises a set of information that allows the description of both physical objects and objects [40]. This section aims to contextualize the importance of describing research data through metadata, addresses the issues related to metadata, metadata models in the scientific context, and the challenges facing image description.

2.3.1 Metadata

Defined as "data about data" [103], metadata is highlighted as essential for the representation and description of the data. In addition, metadata is vital for current scientific communication [137]. 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 [128].

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 [22]. The data description process requires skills, effort, time, and adequate tools for high-quality metadata [61]. 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.

A large amount of data available, particularly in digital form, coupled with the increasing dependence of researchers on the production and analysis of large amounts of data and policies encouraging open access to research data, have made metadata crucial in the digital curation and preservation process. The discoverability and comprehensibility of data by researchers depend on the existence of quality metadata so that it can be reused, re-analyzed, and integrated with other data sources [37]. The main benefits of describing research data for science are increased longevity of data, i.e. by attaching metadata to research data the time that data have value is extended, and increased reuse and sharing of data [83].

Data documentation should be considered good practice in research data management because it allows for the contextualization of data by explaining how the data were created or collected, what they mean, what their content and structure are, and what manipulation processes they were subjected to [46]. Metadata attribution is a highly structured way of documenting information about data, contextualizing it is essential for its understanding and future use, as research data without contextualization documentation is worthless as no meaning can be drawn from it [20]. Metadata has evolved from initially promoting the discovery, acquisition, understanding, and use of data by researchers to more recently enabling the automation of discovery, ingestion, processing, and analysis [83].

Metadata can be characterized according to the objects they characterize. Descriptive metadata identifies and describes the content of the object, for example, title, author, description, language, and keywords. Technical metadata refers to how the data is created or collected and captures the technical characteristics of the formats, some examples are file type, file size, and capture or creation tool. Structural metadata, encodes the internal structure of documents or the structural relations with other objects, for example, the number of pages, sections, and related documents. Preservation metadata, collect information that is needed for preservation and curation activities. Provenance metadata describes changes that have been made to the data and is important in au-

ding processes, an example of this type of metadata is file versioning. Finally, rights metadata describes the restrictions/licenses or lack thereof to which the use of the data is subject, this type includes the license to use, copyright, and rights holder [130]. Willis, Greenberg, and White [137] through their study of domain-specific metadata models, developed a list of fundamental requirements of metadata schemas for documenting scientific data:

- Abstraction - promote through abstraction the capture of new needs and updating of the metadata schema;
- Extensibility, flexibility, and modularity - safeguard the longevity of the model, facilitate its adoption, modification, and extension to needs that have not yet been identified;
- Comprehensibility and sufficiency - define the set of elements so that they are understandable and capture the essential elements to document the domain;
- Simplicity - take into account different levels of expertise and support them with tools and resources to assist in their use;
- Data exchange - having the ability to foster the exchange, sharing and communication of data among the scientific community;
- Retrieval - being able to discover and access data through metadata, taking into account the access standards of specific domains;
- Archive - to enable the long-term preservation of the information contained in the data;
- Publication - ensure the association of the data with scientific publications.

The description of the resources that support research is a complex process due to the diversity of the typology of research data, the purposes for which they are collected, and their use. The existence of diverse scientific domains and concepts specific to them also contributes to the increased complexity of description. Collaboration between researchers and research data managers is often necessary to ascertain the requirements needed to describe data collection of a given scientific domain [24, 137].

Another aspect that demonstrates the complexity of the description process is the cost of metadata production. The moment of capturing metadata implies an analysis of its value and benefits for its users. The greater the detail and structure of the metadata, the greater the cost, especially in relation to time consumption [83]. The idea of capturing as much metadata as possible and then deriving the benefits from the large amount of metadata collected has to be weighed. Metadata collection should start with recognizing its usefulness and doing a cost/benefit analysis [130].

2.3.2 Metadata models

Metadata models for research data collection contain essential information about the creators, context, temporal information, location, and quality of the data. These elements form an interrelated

network of nodes that support discovery, selection, and access for data users and organization, management, and preservation for research data managers. The models together with technological components and the policies, rules, and best practices for metadata use constitute the basic infrastructure for metadata application [98]. Metadata models can be of generic description, meaning that they can be applied to different domains, or models specific to a scientific area. The latter provides better documentation and facilitates access to data from particular scientific domains by incorporating them into thematic repositories and other tools [137]. The application of a single metadata model may not fully match the description needs in a specific context. In these cases, the collection and selection of elements from one or more models are performed to build a combined schema, without losing the interoperability of the base schemas. The final result is the application of a profile characterized by allowing to specify of the controlled vocabulary, format, and relations of the elements, refining definitions of the standards used, and combining different models to specify their use and thus support a particular application, function, community, and environment [24, 62, 39].

The following subsections will describe some metadata models in more detail. The first three examples below have been selected because they are generic, used in different fields of research and widely used by the scientific community. The third, in particular, is an information model specifically designed for the field of scientific research and is even used by the OpenAIRE Guidelines for CRIS Managers⁶. The fourth example was selected precisely because it falls within the domain of images and is therefore of great interest to the study in question.

2.3.2.1 DublinCore

The Dublin Core Metadata Initiative⁷ is a metadata model of the generic application, that emerged in 1995 at a meeting in Dublin, Ohio, initially designed to facilitate the discovery of information on the web, through the introduction of metadata on pages. Developments have taken it towards interoperability, a characteristic of the principles of the semantic web and linked data, using Uniform Resource Identifiers to identify the objects described by the metadata and the terms used in the vocabulary. The most recent version of the terms that make up the Dublin Core dates from July 2012, changing the 2010 specification used until then. The model consists of a set of fifteen generic elements ("Creator", "Contributor", "Publisher", "Title", "Date", "Language", "Format", "Subject", "Description", "Identifier", "Relation", "Source", "Type", "Coverage", and "Rights"), which can be combined with elements from other schemes through the Resource Description Framework and thus form diverse vocabularies.

2.3.2.2 Data Documentation Initiative

The Data Documentation Initiative⁸ is a data description standard based on the XML annotation language applied in the social sciences. The DDI arose from the existing need to better docu-

⁶<https://guidelines.openaire.eu/en/latest/cris/index.html>

⁷<https://www.dublincore.org/>

⁸<https://ddialliance.org/>

ment and communicate the characteristics of empirical data produced, collected, or compiled by researchers. Its specification makes it compatible and complementary to other metadata models such as Dublin Core, MARC, ISO 11179, SDMX, and ISO 19115. The standard is currently in version 3.2, as of March 2014, and consists of 17 modules with various terms: "Archive", "Comparative", "Conceptual Component", "Data Collection", "Dataset", "Group", "Instance", "Logical Product", "Physical Data Product", "Physical Data Product - NCube - Inline", "Physical Data Product - NCube - Normal", "Physical Data Product - Proprietary", "Physical Data Product - NCube - Tabular", "Physical Instance", "DDI Profile", "Reusable", and "Study Unit" [132]

2.3.2.3 Common European Research Information Format

The Common European Research Information Format (CERIF)⁹ is the metadata standard that the European Union recommends to its members for documenting research. It allows the description of research entities, activities, context, and results as well as great flexibility with formal semantic relationships, which fosters quality control, archiving, access, information, and knowledge exchange among decision-makers, researchers, reviewers, research managers, publishers and, ultimately, the general public. CERIF is represented through a relational database model supported by the entity-relationship model. Recent updates to the model have added a formal semantic layer and XML formatting. The model consists of the main entities ("Person", "OrganisationUnit" and "Project"), result entities ("ResultPublication", "ResultPatent" and "ResultProduct"), second-level entities ("Funding", "Facility", "Equipment", "Prize", "CV", "Expertise", "Qualification", "Citation", "Metrics", "Event", "PostalAddress" and "ElectronicAddress"), link entities, multilingual specifications, and semantic specifications [68, 50].

2.3.2.4 Exchangeable Image File Format

The Exchangeable Image File Format (EXIF)¹⁰ is a standard format for digital camera image files that uses a controlled vocabulary to represent administrative image metadata. It was developed by the Japan Electronics and Information Technology Industries Association (JEITA)¹¹ and is currently in version 2.2. It aims to capture and provide useful information about image attributes, processing, and location information [129]. It also allows the organization of images and enables the search of image attributes. The standard only supports the TIFF and JPEG formats and specifies the structure of the image data, the fields used by the standard as well as the definition and management of the format versions [41]. The imaging system assumed in the standard encompasses equipment, commercial applications, and specialized applications. Equipment includes image-capturing devices such as digital cameras. Commercial applications refer to generic programs that enable image processing. Specialized applications encompass programs that specialize

⁹<https://joinup.ec.europa.eu/collection/eu-semantic-interoperability-catalogue/solution/common-european-research-information-format/about>

¹⁰<https://docs.fileformat.com/image/exif/>

¹¹<https://www.jeita.or.jp/english/>

in image processing [41]. This standard allows you to collect information about the image file, lens used to capture, image data, personal data, and image rights.

2.3.3 Controlled vocabularies

Nowadays, the process of describing data is increasingly challenging. Metadata plays an important role in this context, since it adds informational content that enables data interpretation by others, thus promoting data reuse and preservation. However, data description can be a very demanding and long-lasting process, so the development of controlled vocabularies can facilitate and simplify the steps inherent in the data description process while improving the quality of the metadata and the conditions to which users are subjected.

Controlled vocabulary can be characterized as an organization of words or phrases intended for indexing or retrieving information content. Generally, they include elements or preferred terms. According to Harpring [60], controlled vocabularies consist of an information tool that has a standardized terminology to express ideas, physical characteristics, people, places, events, and subjects, among others. In a more simplified way, [61] states that a controlled vocabulary consists of a restricted list of words or concepts that are typically used for descriptive cataloging or indexing.

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 elements or preferred terms.

In short, controlled vocabularies support the organization of information, provide terminology for cataloging information resources, and support the search and description of information. One of their most important functions is to bring together the wealth of variant terms and synonyms of concepts, while promoting consistency through the use of preferred terms and the assignment of the same terms to similar content [60]. Thus, controlled vocabularies allow for indicating activities or functions and, likewise, confidence in the system [71], as they are valuable tools that assist in the process of indexing, categorizing, describing, and retrieving information.

2.4 Image description

Images are dominant in communication, and their abundance, diversity of origins, and variety of holders create a multiplicity of practices regarding description, interpretation, and use [118]. Many authors talk about the role of human behaviors in the life cycle of images [30]. The awareness and education of others for the practices of image record description are essential, commenting on the importance of eliminating description mistakes associated with automatic systems [11].

2.4.1 Challenges and trends

Rapid technological development, particularly in storage and telecommunications, as well as in the ease with which images are incorporated into applications that enable the display, publication,

and dissemination of images, has led to an increase in users' informational needs for images [28].

The capture and production of images are increasingly facilitated by the emergence of technologies that allow you to obtain an image in seconds. With a simple smartphone, we can capture or produce an image quickly and easily. In addition, technological devices also allow, in a fraction of a second, an image to be broadcast to any part of the world. This scenario of image empowerment leads to more images being used and produced in the context of research.

Image indexing has also evolved with technological development. Previously, indexing was dependent on human-made description through textual elements and classifiers, supported by controlled vocabularies. Indexing schemes were mostly created locally or to reflect the characteristics of specific image collections. Some attempts have been made to create generic image description schemes, for example, the Art and Architecture Thesaurus (AAT)¹² and the Library of Congress Thesaurus of Graphic Materials (LCTGM)¹³. Currently, there has been automation of the retrieval and indexing process due to the increasing increase of image content-based solutions [56, 139].

The goals of image indexing are to provide access to images based on their attributes and to provide access to groups of images, not just individual images [74]. Image description and retrieval bring with it the knowledge that interpreting visual signs and images is not the same as interpreting text and language, especially when images are not accompanied by text that points to the meaning of the image [99]. The main approaches to image indexing and retrieval are concept-based and content-based [28, 44].

Concept-based retrieval uses textual description for image retrieval. It can use textual queries, controlled vocabularies, natural language, and captions, and can be purely descriptive, subjective, or abstract [28]. Content-based retrieval uses computational image analysis to automatically extract features such as color, texture, and shape, locally or globally, in digital images [28, 56]. Concept-based indexing provides a high-level image of its content, but its implementation is very costly, especially in time, and suffers from a lack of consistency in indexing due to subjectivity in interpreting images [28]. Content-based indexing is less expensive to implement but has some technical challenges regarding image representation and organization as well as image similarity measurement [139].

Image description performed implies knowledge of some assumptions namely:

- Images are rich, they can contain useful information for researchers in different fields [15];
- Images are sometimes used for purposes other than those anticipated by the creator [15];
- Capturing in words the content or meaning of an image is a significant intellectual challenge [44];
- The same image can have distinct meanings to different people [28];
- The image can be generic or specific, what it represents can be useful at the generic or specific identity level [74];

¹²<https://www.getty.edu/research/tools/vocabularies/aat/>

¹³<https://www.loc.gov/pictures/collection/tgm/>

- It can be difficult to reach a consensus on the content and meaning of the image, as well as on the appropriate attributes to describe the image [28];
- The indexer has to select abstract aspects shared by him and the user to provide suitable elements. The indexer's or user's mental models can be descriptively rich or poor depending on the situation [63];
- Difficulties in matching the terms used by the user to describe their information need with those in the controlled vocabulary [57].

Likewise, the description of images performed automatically also implies knowledge of some key aspects, namely:

- The semantic gap, that is, the difficulty in translating low-level visual features into high-level semantic concepts [139];
- Difficulties in identifying elements that are not directly visible in the image namely, time, space, events and meaning, abstract and emotive concepts as well as unwanted features [45];
- The intent gap, the difficulty the user has in accurately expressing the visual content they want to retrieve [139];
- Humans segment images better than machines due to the ability to associate elements, acquired over the years [35];
- Image understanding techniques still face some challenges regarding execution speed, segmentation reliability, and robustness in classifying similar images [36];
- Techniques that enable automation can be computationally complex, which makes their application difficult in real-world situations [36].

To mitigate some challenges and improve image retrieval there has been the introduction of folksonomies and tags in image descriptions, where users freely assign keywords to images. This approach allows the addition of a particular vocabulary, sometimes very different from the one that the indexer assigns in the same context [86]. While social classification is not a solution to the existing challenges in image description, it offers a new opportunity to improve image indexing and user interaction. The traditional approach to indexing is complemented by this one, on one hand, the traditional one offers more consistency, a similar level of specification, interoperability, and uniform access through metadata standards and controlled vocabularies. Social classification brings the possibility for users to describe in their native language, new perspectives, specialization, interaction with users, and user-oriented indexing [79]. There has also been the emergence of hybrid image retrieval systems that involve features and functionalities of both concept-based and content-based retrieval [44].

In describing images it is important to strike a balance between which images should be indexed individually or collectively, what level and how different attributes of the images should be differentiated, and what depth of indexing is desirable [74].

2.4.2 Automatic description of images

In Research Data Management, data description is one of the core tasks to ensure that all data are properly interpreted and understood and that they can be cited and reused. However, in many cases, the process involved in this task is time-consuming and can sometimes be discarded by researchers [7].

Scientific images constitute a great challenge, as there are still no sufficiently enlightening guidelines for orienting the process of describing images, taking into account their content and the sense and meaning by which they were produced.

Science benefits from images since the possibility of accurate registration enhances scientific research. If, in its beginnings, the image records were related to tribal peoples or new species discovered, with the passage of time it allowed to evolve in studies in which successive photographs were shot (one of the most famous studies in this regard is from 1878, by researcher Eadweard Muybridge [85], in which he took photographs in a sequence of running horses, in order to study the gallop), even others in which the technology associated with this documentary typology has boosted innovative studies in the scope of brain mapping and in the human body assessment, for example [66].

The production of images is, more than ever, in a growing phase. Nowadays, through various technological devices, such as mobile phones, tablets, and drones, it is possible to capture and broadcast images with great ease.

Vision is known as one of the most refined human senses and, for this reason, images are considered essential in human perception. Unlike human vision, which is limited to the visible band of the electromagnetic spectrum, computers, and imaging devices cover almost all of the electromagnetic spectrum [4]. Thus, imaging devices can process images generated from ultrasound, electron microscopy, and computer-generated images. This factor motivated the use of digital image processing in a wide field of applications [47].

Automatic image processing is an increasingly common practice in several areas. There are several everyday applications where automatic image processing appears. Among the various applications are robotics, facial animation, image editing and manipulation, deblurring (the process of removing distortions in images), noise removal, contrast enhancement, visual surveillance, traffic monitoring, games, human-machine interaction, commercial applications, industrial automation, identification of individuals and improvement of visual information for superior human interpretation [47].

Image processing is divided into three different groups. The low group, in which the processing acts at the pixel level and the operations are basic, such as noise reduction, increasing contrast, and smoothing images [4]. The middle group involves more advanced tasks, including segmentation, the recognition of individual objects, and the partition of images into regions or objects [4]. The high group implies the interpretation of the content of the image, in which the objective is to perform cognitive functions that, as a rule, are inherent to human vision [4].

Chapter 3

Informational behavior of researchers using and producing images

To properly understand the habits and behaviors of those who use and produce images in research, it is necessary to collect data from a variety of sources in order to validate findings, ascertain standards, and discuss results. To this end, four sequential studies were developed. The first focused on a particular case, in which the behaviors and habits of a mobility research group were evaluated. This study was intended to be the starting point for subsequent studies, as it provided preliminary results that opened up directions for more in-depth studies. The second and third studies are based on exploratory studies in which questionnaires and interviews were used, respectively, to collect data on the current habits of the largest number of researchers, regardless of their research domain. The fourth study is based on a literature review, where, through a unique and specific methodological approach, it analyzed what the literature tells us about the behaviors of researchers who use and produce images and the management process involved.

This chapter is aligned with RQ1, aiming to give depth to the question, trying to raise hypotheses and record conclusions that lead to its answer.

3.1 The case of the Generation.Mobi project

3.1.1 Contextualization

To better understand the photographic document, from the phenomenon of transformation from analog to digital to its inclusion in research projects, this experience intends to present a case study of a project in the domain of dynamic mobility management, called Generation.Mobi. The work starts by analyzing the use of photography in Generation.Mobi, starting with the presentation of the project, up to the workflow approach. Finally, the analysis of the results corresponding to the questionnaires applied to Generation.Mobi members is presented.

3.1.2 The Generation.Mobi project

The result of a consortium between Iberia, Revolution Answer, the Centre of Engineering and Product Development (CEiiA), Cardio ID, and the Faculty of Engineering of the University of Porto (FEUP), Generation.Mobi is a research project that aims to develop and validate a dynamic mobility management system, designed for the new generation, whose concept is based on the notion of a social network of interoperable and interactive bicycles with the ecosystems of cities.

In addition to the companies and entities mentioned, Generation.Mobi has as a partner the RENER Living Lab, a Portuguese network of smart cities. This alliance is expected to enhance the level of radical innovation of services and operations related to smart urban mobility, both nationally and internationally.

The project explores issues ranging from communications architecture, models, and algorithms for predictability of demand and supply of mobility services, interactions with the electrical grid, modularity, flexibility, and portability of urban mobility micro-hub, flexible bicycle architectures, and cardiac biometrics technologies applied to bicycle components. Generation.Mobi aims to develop a mobility management system that incorporates a network composed of interactive devices, such as bicycles. This project was designed to respond to a series of societal challenges in the domain of the European Strategy of Horizon 2020.

3.1.3 Generation.Mobi workflow

The study of the Generation.Mobi project, through the interactions with its members, allowed us to understand their informational behavior.

It was noticeable that Generation.Mobi produces and uses photographs in several moments of the research. In the initial phase of the project, in which the requirements survey and project planning are performed, the photographic record is non-existent. However, in the following phases, it is always present.

The second phase of the project consists of the preparation of the material. This moment is crucial to the success of the experiments. For the project leader, poor preparation of the material that will be used in the experiments can compromise the results. This preparation includes marking the floor, positioning bicycles, placing antennas, and measuring angles, for example. Sometimes researchers are unable to perceive certain errors and flaws that occur in the preliminary stages, and with the use of photography they can, at a later stage, perceive these possible deviations in the results, confirm scenarios, and investigate anomalies in the material, without having to repeat processes.

In the third phase of the project, inherent to the tests, photography is also used, this time to document the various moments of the experiments and their participants. The advantage of using photography in this phase of the project is the same as that presented for the previous phase. That is, it allows that at later times the photographs can be used to analyze the moments at which they were produced.

In the fourth and last phase of the project, the analysis of results, photography is no longer produced. However, its use continues to exist, this time as an input to analyze the data collected. As explained in the previous paragraphs, the representation capacity that photography enhances, allows researchers to better understand the results, enabling them to go back to the moment of the experiments through the images, without having to repeat the experiments. This information can be seen in Figure 3.1.

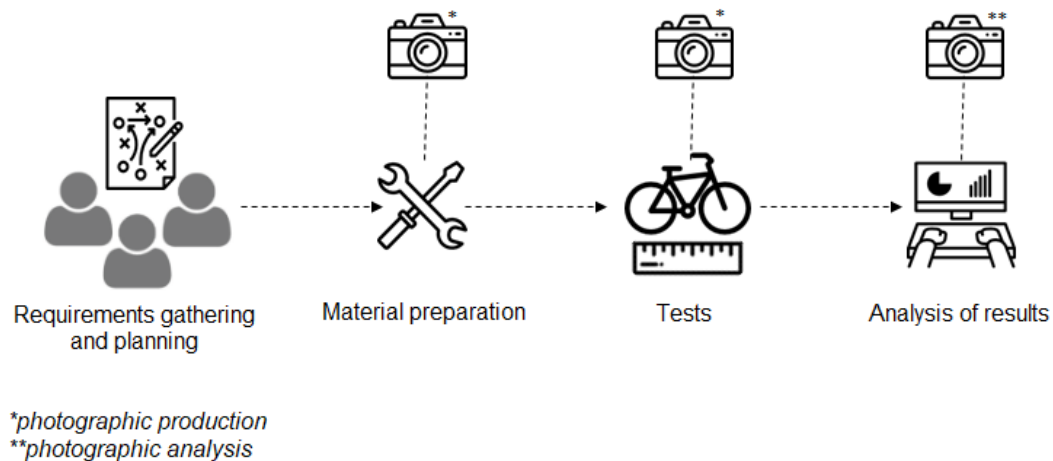


Figure 3.1: Generation.Mobi workflow

Source: Own authorship

3.1.4 Methodological approach

After understanding the Generation.Mobi workflow, through interaction with the project manager, it was necessary to conduct rigorous research to acquire meaningful data on the use of photography by the various project members. With hypotheses formulated, seven questions were asked in order to be able to obtain an explanation for what had been speculated.

Since the questionnaire survey is a non-participant observation technique supported by a sequence of written questions addressed to a set of individuals who make up the sample of respondents, it is possible to collect the most varied factual information about the theme under analysis.

The criteria for the elaboration of the questionnaire in question were related to a few factors. First, it had to be brief, as this would make it easier to be predisposed to its completion. Second, it had to be intuitive and with a simple lexicon, since most researchers are not yet familiar with the terminology inherent to data management, for this reason, in the seventh question we chose to put the explanation of the terms inserted in the question so that it was easy to understand the intent of the question and its options. Third, it was essential to explain to the respondents the purpose of the study, as well as the research domain inherent to it. Fourth, it was essential to ensure the anonymity of the respondents, not only because the identification of each individual is not determinant for the analysis of results, but mainly because of the privacy issues inherent in the New General Data Protection Regulation. For these reasons, the questionnaire contains a

brief header where respondents can see all these details explained and confirmed, thus ensuring full transparency between the applicant and the respondents.

Although the questionnaire survey should be, for the most part, a study applied to large groups of individuals, its use and relevance should not be ruled out when the groups are smaller in number. In the case of Generation.Mobi, six people were part of the project, so this was the universe defined for the study. However, only four of these members answered the questions, so the results presented summarize the answers of these four researchers.

3.1.5 Results

Below is an individual analysis of each of the answers given to the questionnaire. Throughout this analysis, the questions that make up the questionnaire can be seen and, where necessary, the respective graphical representation.

1. Were you one of the researchers who produced or used photographs in the Generation.Mobi research process? If you answer "No", skip to question 5.

Most (75%) of the researchers produced and used photographs during the research process. Only one of the researchers said he only produced photographs. The remaining researchers showed that they had used photography either as a working tool, allowing the capture of research moments, or as an essential complement for the analysis of the data produced.

With this question, it was possible to confirm what had been said by the researcher responsible for the project: the photographic image was present in much of the Generation.Mobi workflow, with utility for all its intervenients.

2. What conditions led you to produce photographs?

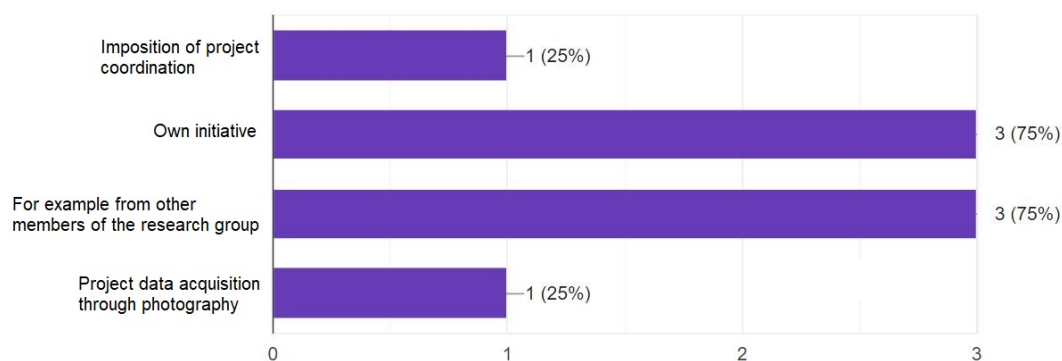


Figure 3.2: Questionnaire - Answers to the question 2

This question allowed respondents to give multiple answers (Figure 3.2), as the production of the photograph, in many cases, is not only about one factor. Thus, there are three answers for "Own initiative" and for "By the example of other research groups". These answers allow us to

verify an interesting factor: most of the answers are not conditioned to compulsory factors, that is, the picture was produced because they really saw a use for it.

The option "Imposed by the project coordination" has one answer. This allows us to verify that those responsible for the projects, particularly Generation.Mobi, are beginning to be sensitive to the use of photographs as research support, seeing more clearly their potential for the projects.

Finally, a response that came at the initiative of one of the respondents, since it was mentioned through the option "Other". In this case, the researcher states that the conditions that led him to produce photographs of Generation.Mobi was related to "Gaining data from the projects through photography". This finding is really interesting because it allows one to understand that scientific imagery can assist in data collection. In the case of Generation.Mobi, this had to be carried out very rigorously and cautiously, because deviations in the results could compromise the entire research, given the nature of the data, and with this response, it was realized that photography played a key role in data acquisition.

3. What conditions led you to use photographs?

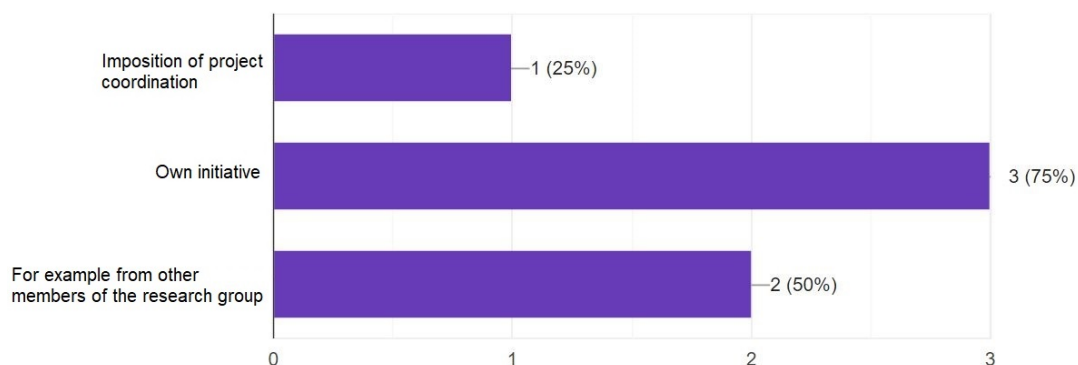


Figure 3.3: Questionnaire - Answers to the question 3

As for the conditions that led the researchers to use photographs, the answers were more uniform (Figure 3.3). Also in this case the respondents had the possibility of selecting multiple answers. Thus, the answer "Own initiative" was the one that obtained the highest number, with three. It was followed by "By the example of other research groups", with two. Finally, "Imposed by project coordination", with one response. Once again it was found that the use of photographs is not, for the most part, related to a question of compulsion, but by the finding of a significant advantage to the research.

4. What do you consider to be the advantages of including photography in the research process of Generation.Mobi? Select up to 3 of the following options.

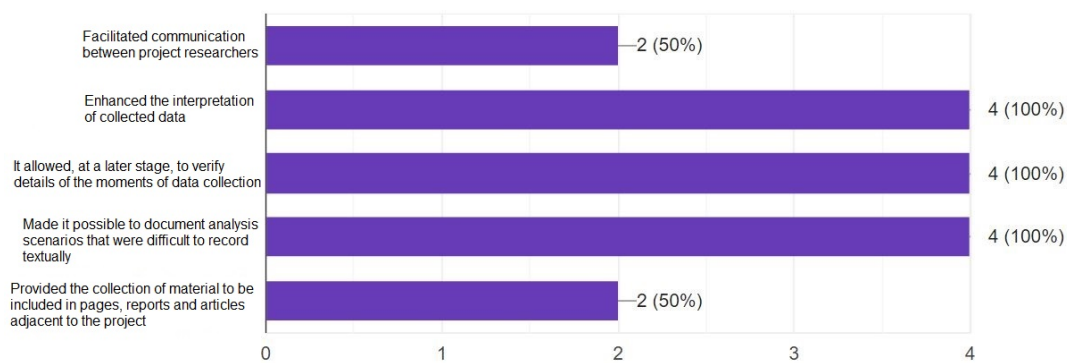


Figure 3.4: Questionnaire - Answers to the question 4

With question number 4 (Figure 3.4) it was verified that the researchers did indeed see advantages in the use of photographic records. As this was also a multiple-answer question, it was possible to understand more clearly the benefits of photography in Generation.Mobi.

One hundred percent of the respondents selected three of the five hypotheses presented, namely "Enabled the interpretation of the data collected", "Enabled, at a later stage, the verification of details at the time of data collection" and "Enabled the documentation of analysis scenarios difficult to record textually". This series of answers highlights the interpretative power of the image and the unique representativeness that the photographic image possesses, enhanced by its effect of duplicating a reality, a characteristic that no other documentary typology has.

Besides these answers, two others were recorded, each answered by two of the respondents. In these responses, they indicated that photography "Facilitated communication between project researchers" and that it "Provided for the collection of material to be included in pages, reports, and articles adjacent to the project".

5. Consider the following scenario: a project researcher sends you a photograph that reports a certain data collection moment in which you were not present. Do you think you will easily interpret the information that reports that moment or would you prefer the same content as the photograph, but textually documented?

The photo should be sent with a contextual text.
I think the photos can show how the experiments were performed but not the technologies used for data collection. That is, not in all experiments we use the same scripts and programs for data collection and I think that is hard to fit into a photograph.
I will easily interpret the context with photography.
The content of the photograph documented textually would be preferable.

Figure 3.5: Questionnaire - Answers to the question 5

Since this is an open-ended question (Figure 3.5), the main objective was to understand each

researcher's point of view when faced with a scenario in which they had to interpret the content of a photograph. The truth is that image content analysis still lacks further study and guidelines for its proper implementation. From the research point of view, this analysis can have many variants and distinct procedures, depending on the research domain.

In the case of Generation.Mobi, one researcher mentioned that "The photo should be sent with a contextual text", showing the importance he attaches to text to facilitate interpretation. Researcher 2 mentioned, "I think the photos can show how the experiments were performed but not the technologies used for data collection. That is, not in all experiments do we use the same scripts and programs for data collection and I think that is hard to fit into a photograph". This answer shows that, from this researcher's perspective, the photograph cannot always stand alone, as there are certain details that are not possible to document with the photograph, and a textual complement is needed in order to avoid failure in noting all the details of the research or experiment. In turn, researcher 3 stated that "I will easily interpret the context with photography", clearly stating that in the case in question he sees no difficulty in interpreting the photographs relating to the project he was involved in. Finally, researcher 4 says that "The content of the photograph documented textually would be preferable", giving evidence of his preference for documents of textual character for the recording and analysis of research processes.

6. Would you like to see the photograph accompanied by textual elements that give it context (such as description, keywords, spatial and temporal coverage...)?

The answers to question number 6 left no doubt as to the usefulness, in the view of Generation.Mobi researchers, of the textual component in attributing context to photography. As already mentioned, the content analysis of images still involves several difficulties and doubts, making it natural that the overwhelming majority of documents are textual. Although the advantages and potentialities of photography are noted by all, the textual elements are still seen as fundamental for a better interpretation of the photograph, since they allow obtaining an explanation of what is being reported in the photographic record. Therefore, this question obtained one hundred percent of the answers for the option "Yes".

7. Of the following, which do you consider to be the fundamental elements to describe a photograph in the field of research?

Taking into account the answers to question 7 (Figure 3.6), in which the researchers show an interest in seeing the photographs accompanied by textual elements, it is interesting to check, with the answers to question 7, which is the vocabulary they consider most relevant for the description of photographs in the field of research in question.

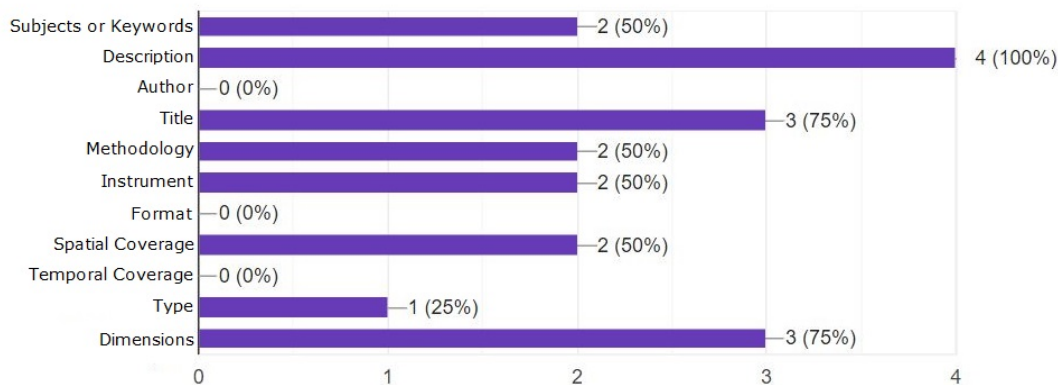


Figure 3.6: Questionnaire - Answers to the question 7

All researchers consider "Description" to be relevant, thus being the element with the highest number of responses. Right after that, with three responses, are "Title" and "Dimensions". Mentioned by half of the respondents are "Subjects or Keywords", "Methodology", "Instrument" and "Spatial Coverage". With only one answer is "Type". Missing any matches are "Author", "Format" and "Temporal Coverage".

This question did not imply a restricted number of answers, so the researchers chose those that, in reality, they considered necessary for a correct description. In addition, the possibility of adding other options to those already mentioned was given, which none of the researchers did, predicting satisfaction with the options presented.

3.1.6 Final remarks

This study allowed us to have a first approach to several aspects related to the scientific image, in particular photography inherent to research practices. In fact, the overwhelming dominance of new technologies and new digital practices enables photography and its production to grow vertiginously, since through the most diverse devices, photography is easily reachable.

On one hand, photography can be a valuable element in safeguarding information, since it captures, with precision, the various components photographed. In this way researchers are given the possibility to see various details and ensure that nothing is forgotten by human memory which, due to its limitations, does not guarantee the preservation of all particularities. Therefore, human recollection, with the help of photography, is made natural. On the other hand, photography can be a consequence of the research process and, therefore, one of the constituents of a data set. For this reason, it is susceptible to being analyzed and processed in order to obtain a series of results and conclusions. As such, the photograph will be susceptible to description in order to give context and achieve a level of depth that would not otherwise be possible. With description, the loss of information and the trivialization of photographs is avoided, improving the research process from the initial to the final stages, and making the researchers' time more profitable.

If there is standardization regarding the work to be done with scientific imaging, inconsistent practices will be discarded, giving way to solid techniques and decreasing the difficulty in giving meaning to the photographs. For this to happen, it is necessary to understand the contours of image production in science and to understand the informational behavior of researchers who use photography as research support. We can also raise awareness of its importance, raise awareness of management issues, and find out whether their informational needs are more readily satisfied through the use of research data management techniques. Indeed, educating researchers on this issue will possibly trigger more consistent and active behavior toward imaging in the research context.

3.2 Exploring image management practices through questionnaires

3.2.1 Contextualization

The rapid development of processing capacity, image management and the ease of replication and dissemination increased the access and value of image collections [44]. In the context of research, various image-capture devices have emerged. With the increased use and volume of image collections, new challenges and opportunities have arisen in image research data management. Vejvoda, Burpee and Lackie [133] present preliminary recommendations for image management in the context of research through the recommendations established for numerical data sets. In our search for related work, we found no other studies that addressed these issues.

The work described in this section is motivated by the lack of knowledge concerning the production processes, use, and management of images in a research context. As images constitute a valuable informational element for research. It becomes necessary to include them in research data management processes to make them accessible and reusable. It is therefore essential to know the processes and habits in the production and use of images in research to inform the production of appropriate recommendations.

If the study reported in section 3.1 allowed a first analysis of the habits and behaviors of those who use and produce images, this one allows a deeper analysis, covering a larger sample, in order to obtain more concrete results.

3.2.2 Methodology

A questionnaire was used to study practices and habits in the management of research data in image format. The structure of the questionnaire and the design of the questions were informed by the research data lifecycle of the Data Documentation Initiative [125], DataONE [9] and UK Data Archive [131]. Questions were grouped by stages of the research data lifecycle: planning, creation/compilation, quality assurance, processing/analysis, description, storage, and sharing. These stages were chosen after analyzing the lifecycle models referred to in the previous paragraph.

Figure 3.7 shows the workflow of methodological approach. The methodology consists of three major phases: elaboration, dissemination, and analysis of the questionnaire.

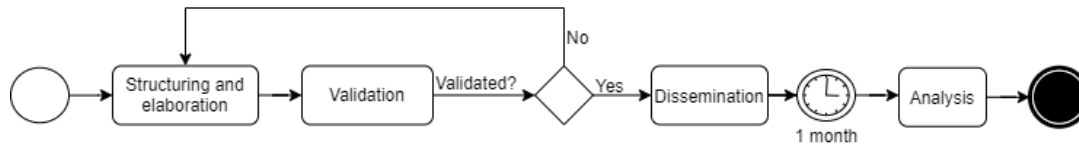


Figure 3.7: Methodology workflow

The questionnaire underwent several revisions by the authors, where they discussed and reflected on the suitability and wording of the questions. Before it was released, the questionnaire was tested by an external researcher to analyze the interpretation of each question. The objective was to ensure that the questions were well understood by researchers not specialized in the subject.

The questionnaire opened on 20 February 2020 and accepted answers until 26 March 2020. The questionnaire was distributed by email to the research community at the University of Porto, namely i3S¹ and INESC-TEC², and in research units outside the University of Porto funded in 2019 by the Foundation for Science and Technology³. Namely, Centre for Informatics and Systems of the University of Coimbra, Cardiovascular Centre of the University of Lisbon, Centre for Philosophical and Humanistic Studies of Universidade Católica Portuguesa and Center for Mathematics and Applications of Universidade Nova de Lisboa⁴. An email was sent to the coordinators of each group, asking them to share the questionnaire with their members.

Respondents were people with research experience, namely PhD students, fellows, and university professors. Answers were given anonymously, not allowing the identification of the participants.

Quantitative methods were used to analyze closed-ended questions and content analysis for open-ended questions. There were two types of closed-ended questions. A kind of question collected answers on a 5-point Likert scale (Never, Rarely, Occasionally, Often and Always), generating ordinal variables. The other type, involved the selection, or not, of provided answer options, creating nominal variables.

For each question, the general trend in the overall set of answers was analyzed and comparisons made between four scientific areas (Life and Health Sciences - LHS; Exact Sciences and Engineering - ESE; Natural and Environmental Sciences - NES; Social Sciences and Humanities - SSH) and between three rates of image use in research (low - less than 25% of the projects; moderate - about 50% of the projects; high - more than 75% of the projects). For the ordinal variables, the Kruskal-Wallis test was used to detect whether there were differences between the

¹<https://www.i3s.up.pt/>

²<https://www.inesctec.pt/en>

³<https://www.fct.pt/>

⁴<https://www.cisuc.uc.pt/>; <http://ccul.pt/>; <https://cefh.braga.ucp.pt/>; <https://www.cma.fct.unl.pt/>

groups. Where differences were found, the Pairwise Mann-Whitney test with Bonferroni correction was used to identify the location of the differences. For nominal variables, the Chi-square test for equality of proportions was used.

The scientific domains were identified according to the division made by the Foundation for Science and Technology. When reporting our results, we use * to indicate results significant at $\alpha = .05$ and ** to indicate results significant at $\alpha = .01$.

3.2.3 Results

Responses were collected from 107 researchers. The questionnaire, response data and detailed statistical results are available for consultation online in INESC TEC's Research Data Repository [49].

Of the respondents, 41 (38.3%) work in the Life and Health Sciences, 30 (28%) in the Exact Sciences and Engineering, 12 (11.2%) in the Natural and Environmental Sciences, and 24 (22.4%) in the Social Sciences and Humanities domain. Figure 3.8 relates the use of images as data with the research domain. We can observe a greater tendency to use images as research data in the domain of Life Sciences and Health. In turn, Figure 3.9 presents information on the frequency of use of images as data, in each of the research domain.

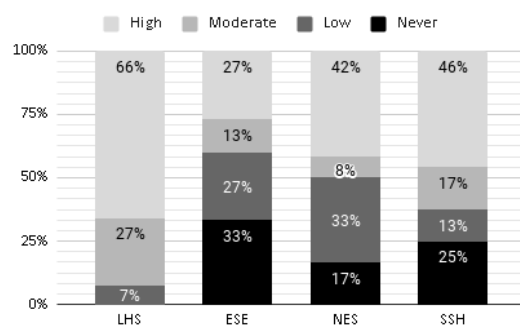


Figure 3.8: Image use by research domains.

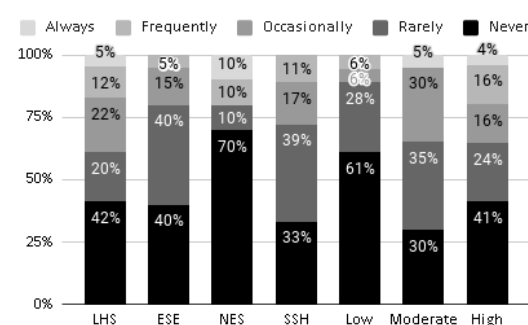


Figure 3.9: Guide by area and frequency.

Results are presented according to the stages of the research data lifecycle.

3.2.3.1 Planning

When asked about the existence of a document to guide the use and production of images during the research, 69.7% answered never or rarely, 16.9% occasionally and only 13.5% replied frequently or always. Even the researchers more accustomed to using images (High group) rarely make a document to guide the production and use of images, as can be seen in Figure 3.9.

We did not find significant differences between scientific areas or rates of image use.

Regarding the way and frequency in which researchers acquire and produce images, most of them produce images (73%) always or frequently, although about 59.5% of the researchers also

consider that they occasionally or frequently use images from past projects. It should be noted that 89.9% of the respondents state that they never buy images from others and 61.8% of them say that they rarely or never acquire images from third parties, even images with no associated cost.

It was found that the Life and Health Sciences domain is associated with significantly more images being produced in a research context when compared to the Natural and Environmental Sciences and Social and Human Sciences domains. This is visible in Table 3.1 which shows the significant differences between scientific areas.

Table 3.1: Significant comparisons between scientific areas

Phase	Var	Description	LHS>	ESE>	NES>	SSH>
P	O	Images produced in the research	NES*			
			SSH*			
CC	O	Image capture via traditional camera			LHS*	LHS**
						ESE**
CC	O	Image capture via microscope	NES**			
			SSH**			
			ESE**			
CC	O	Perform simple editions to the image	NES*			
QA	O	Check the quality of images	ESE*			
PA	O	Manual image analysis	SSH*			
PA	O	Automatic image analysis	SSH**	SSH**		
			NES**	NES*		
PA	O	Document the conducted analysis	NES*			
			SSH**			
			ESE*			
D	O	Describe image on paper	ESE**			ESE*
D	N	Use of Author as a element	ESE*		ESE*	
D	N	Use of Description as a element	ESE*			
D	N	Use of Capture Instrument as a element	ESE**			
			SSH**			
D	N	Use of Methodology as a element	SSH**			
D	N	Use of Sample as a element	ESE**			
			SSH*			
ST	N	Save images as TIFF	ESE**			
ST	N	Save images as PNG		LHS*		
SH	N	Sharing promoted by PI	ESE**			
SH	O	Include images location in scientific articles		LHS**		LHS*

Phases: Planning (P), Creation/Compilation (CC), Quality Assurance (QA), Processing/Analysis (PA), Description (D), Storage (ST), and Sharing (SH). Variables: Ordinal (O) and Nominal (N).

Likewise, those with low use of images, produce fewer images that researchers with moderate and high use of images in research. This is visible in Table 3.2 that shows the significant differences between rates of image use.

Table 3.2: Significant comparisons between rates of image use

Phase	Var	Description	H>	M>	L>
P	O	Images produced in the research	L**	L**	
P	O	Images come from past research	L**		
CC	O	Image capture via microscope	L**	L*	
CC	O	Perform simple editions to the image	L**	L**	
QA	O	Check the quality of images	L*		
PA	O	Manual image analysis	L*		
PA	O	Automatic image analysis	L*		
PA	O	Document the conducted analysis	L**	L**	
D	O	Annotate sets of images	L*		
ST	N	Save images as TIFF	L*		
SH	O	Include images location in scientific articles			H*

Phases: Planning (P), Creation/Compilation (CC), Quality Assurance (QA), Processing/Analysis (PA), Description (D), Storage (ST), and Sharing (SH). Variables: Ordinal (O) and Nominal (N). Use Rates: Low (L), Moderate (M), High (H)

3.2.3.2 Creation/Compilation

Images are always or frequently produced by the computer (51.7%), the microscope (42.7%), the camera of the mobile phone (31.5%), and the traditional photo camera (21.3%). In Figures 3.10 and 3.11, we can see the percentage of use of these instruments by scientific area and frequency of use.

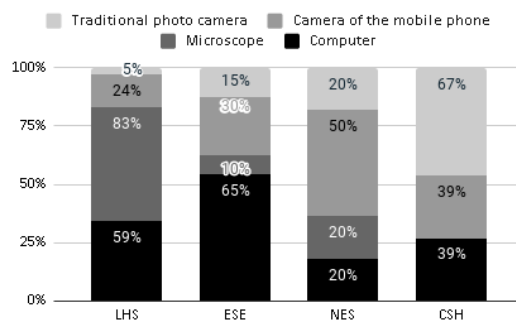


Figure 3.10: Instrument of capture by scientific area.

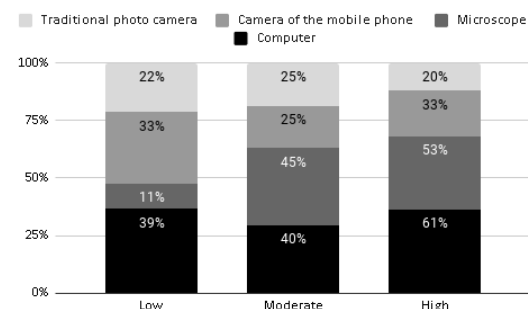


Figure 3.11: Instrument of capture by frequency of use.

Regarding significant differences between scientific areas (Table 3.1), Social Sciences and Humanities researchers use the traditional camera significantly more than researchers from Exact Science and Engineering and Life and Health Science. The latter researchers also use this instrument significantly less than Natural and Environment Sciences ones. Similarly, Life and Health Sciences researchers are the ones who mostly use the microscope in comparison with the other scientific areas. In Table 3.2, it can be seen that those who use images less often use microscope images less often.

Regarding image edition/manipulation, 65.2% of the researchers say they always or frequently use image clipping, 47.2% always or frequently use simple adjustments to properties such as contrast, brightness, and saturation, and 49.4% change the dimensions of the image, always or frequently. About 59.6% of researchers say that they rarely or never combine objects from different

images to create a new one and 53.9% say they rarely or never use filters to improve image quality.

Figure 3.12 relates the use of images as data with the research domain. In terms of significant differences, Natural and Environmental Sciences researchers perform significantly less simple adjustments to images comparatively to Life and Health Sciences researchers (Table 3.1).

Figure 3.13 relates the use of images as data with the frequency of use. Researchers that use images less often perform simple editions significantly less than the others (Table 3.2).

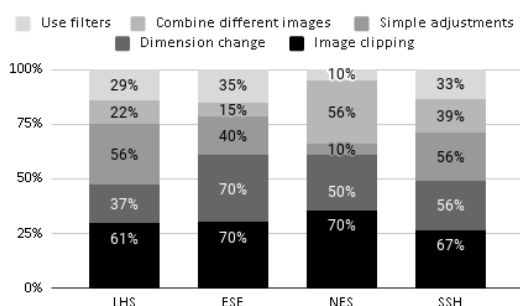


Figure 3.12: Image edition by research domains.

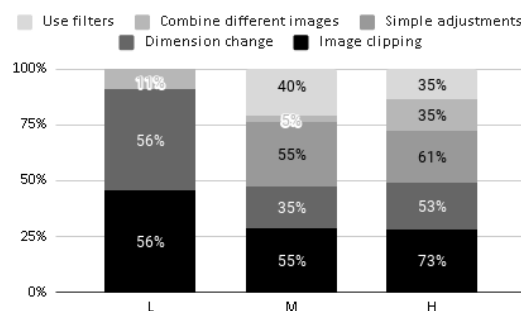


Figure 3.13: Image edition by frequency of use.

Of the respondents who edit the images, 57.6% indicated that they always preserve the original image, and 23.5% frequently preserves it. Only 2.4% of the researchers never keep the original image. No significant differences were found between scientific areas or rates of image use.

In turn, documentation of editions is always or frequently done by 34.1% of respondents, and 47.1% rarely or never document. No significant differences were found regarding research domains and rates of image use.

3.2.3.3 Quality assurance

Of the respondents, 53.9% answered that they always or frequently check the quality of their images. Only 7.9% of the researchers say they never do it, and 14.6% rarely check the quality.

Natural and Environmental Sciences researchers check the quality of images significantly less than Life and Health Sciences researchers (Table 3.1). The same happens in researchers that use images less often. These researchers check the quality of the images significantly less than researchers accustomed to using images (High use group) (Table 3.2).

As for the processes mentioned by the researchers to ensure the quality of the images, these are varied. However, the following stand out: a review of images and their properties (36%), calibration of instruments (23.3%), and disposal of inappropriate ones (9%). There were no significant differences in relation to scientific areas and rates of image use.

3.2.3.4 Processing/Analysis

When asked about the most used computer programs in image processing and analysis, the researchers mostly mentioned ImageJ (33.7%), Photoshop (19.7%), Paint (10.1%), Cell Profiler

(9%) and programming languages (7.9%). About a quarter (24.7%) of the researchers do not use any computer program.

It can be said that images are the object of different types of analysis. Those that stand out the most are content analysis (37.1%), mathematical calculations (18%), quantifications (11.2%), and measurements (5.6%). Of the respondents, 20.2% do not analyze images.

Finally, it was found that researchers tend to combine manual and automatic image analysis. There is a balance between manual mode (38.2% answered always or frequently) and automatic mode (42.7% answered always or frequently). In Table 3.1 it can be seen that automated analysis is significantly higher in the Life and Health Sciences and in the Exact Science and Engineering domains. On the other hand, Social Sciences and Humanities researchers are the ones who use manual analysis less in contrast with Life and Health Sciences. Researchers that rarely use images use significantly fewer automatic and manual modes to analyze images in comparison with those who highly use images (Table 3.2).

It was found that 46% of the researchers said that they never or rarely document their analysis of the images. Life and Health Sciences researchers are the ones that document more frequently the steps taken in the analysis of images in comparison with the other scientific areas (Table 3.1). The vast majority of researchers who do not use images often, document their analysis significantly less than researchers who do a more intensive use of images (moderate and high use groups), as can be seen in Table 3.2.

3.2.3.5 Description

There is a slight tendency to associate annotations with individual images (44.9% always or frequently), instead of annotations in the set of images (37.1% always or frequently). No significant differences were found between the scientific areas. Researchers with low image use, do not annotate sets of images as often as those who highly use images do (Table 3.2).

Regarding the support where the annotations are made, the use of documents other than the image stands out. 32 (35.9%) of the respondents said that they always or frequently write it down on a paper document, about 36 researchers annotate (40.5%) on a digital document. Only 16 (18%) of the researchers write in the image, always or frequently. Exact Sciences and Engineering researchers are the least likely to take notes on a paper document other than the image in comparison mainly with Life and Health Sciences but also with Social Sciences and Humanities. These last two have the highest percentage of researchers that annotate on a paper document, 56.1%, and 27.8%, respectively (Table 3.1). There is no clear preference in the annotation support regarding the frequency of use of images in research.

It should be noted that 65 (83.1%) of the respondents say that they never or rarely use any application that helps them describe images. Only six researchers (6.7%) say they do use, always or frequently, an application to help them. No significant differences were found between scientific areas and frequencies of image use.

Regarding the metadata standards for image description, it is clear that few researchers use and know the topic. Of the three options given in the questionnaire (Dublin Core, Common European Research Information Format and EXIF), none showed a percentage of use above 1.1% for "always" frequency and above 6.7% for "frequently". In turn, the average percentage for the frequency "never" is 85.7%. When posed with the possibility of presenting standard options other than those of the questionnaire, only five researchers answered, showing a clear trend towards the non-use of metadata standards for the description of images. No significant differences were found between scientific areas and frequencies of image use.

When asked about the vocabulary/elements of description that researchers most use and consider relevant, the answers vary. Seen as most relevant are the elements "title" (68.5%) "author" (53.9%), "date" (57.3%) and "description" (51.7%). The following elements are the least seen as relevant, "rights" (39.3%), "format" (40.4%) and "capture instrument" (41.6%). Among the most used elements are "title" (83.1%) "author" (62.9%), "date" (62.9%) and "description" (64%). With the lowest utilization percentages are also the elements "rights" (24.7%), "format" (39.3%) and "capture instrument" (39.3%).

Figures 3.14 and 3.15 show the elements used and seen as relevant by the scientific area.

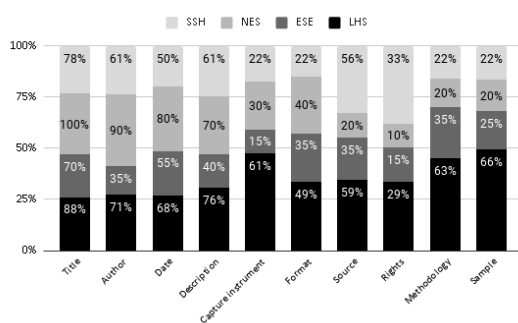


Figure 3.14: Elements used by research domains.

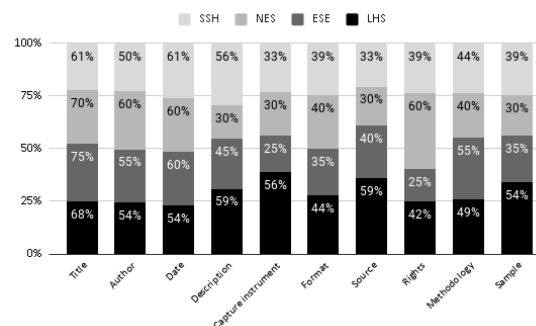


Figure 3.15: Elements relevance by research domains.

Exact and Engineering Sciences researchers are the least likely to use the *author* in comparison with Life and Health Sciences and Natural and Environmental Science as well as the *description* in comparison with Life and Health Sciences. In the opposite direction, Life and Health Sciences researchers are more likely to use the element *capture instrument* and *sample* in comparison with the Exact and Engineering Sciences and the Social Sciences and Humanities domains. Life and Health Sciences are also more likely to use the element *methodology* than Social Sciences and Humanities (Table 3.1). As for the relevance attributed to the elements, there were no significant differences between scientific areas and frequencies of use of images as research data.

3.2.3.6 Storage

Regarding storage location, the computer stands out (95.5% always or frequently), followed by external disk (73%), pen drive (41.5% always or frequently), and cloud (40.4% always or often).

No significant differences were found between scientific areas and frequencies of image use.

The most used formats to store are TIFF (74.2%), JPEG (83.1%) and PNG (61.8%). Although with very low usage percentages, RAW (13.5%), BMP (7.9%), SVG (2.2%) and PDF (3.4%) formats are also used. Figure 3.16 shows the storage format preference by the research domain. Regarding significant differences, Exact and Engineering Sciences researchers are the least likely to use the TIFF format to store their images when equated to Life and Health Sciences. Similarly, Life and Health Sciences researchers are less likely to use the PNG format than Exact and Engineering Sciences researchers (Table 3.1). Figure 3.17 shows the storage format preference by frequency of use. Regarding significant differences, researchers who rarely use images are less likely to use the TIFF format to store images than those who frequently use images (Table 3.2).

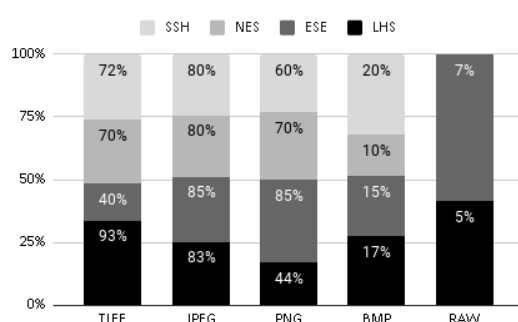


Figure 3.16: Storage format by research domains.

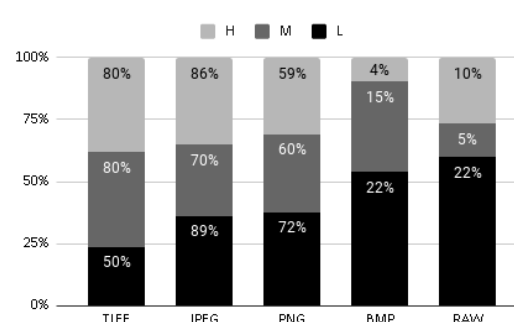


Figure 3.17: Storage format by frequency of use.

Regarding the volume of images stored during a research project, about 58.4% of the researchers said they were not able to quantify it. No significant differences were found in relation to scientific areas and frequencies of image use. When asked about the frequency with which they made backup copies, 40.4% responded monthly, 15.8% weekly, and 19.7% daily. About 18% of researchers do not regularly back up and 19.1% do not back up. When researchers make backup copies, they do it manually (65.2%), only 13.5% answered that they were done automatically. In these two questions, there were no significant differences in relation to the scientific area and frequency of use of images.

3.2.3.7 Sharing

Image sharing occurs mainly at the end of the investigation (61.8% answered always or frequently) or during the investigation (59.5% answered always or frequently). Despite this, 41.6% of the researchers responded that they often share the images sometime after the research project is finished. In this question, there were no significant differences between scientific areas and frequencies of image use.

When asked who promotes the sharing of images, the most frequent answer was the principal investigator (PI) (71.9%), followed by the investigator producing the image (44.9%), institution (31.5%), and funding entity (11.2%).

The PI is the one who promotes the sharing of images most significantly in the Life and Health Sciences relative to Exact Sciences and Engineering (Table 3.1). No significant differences were found between the frequencies of image use.

Regarding rights of use, the researchers replied more frequently that they leave the images restricted in access and use (35.4% always or frequently), followed by free to access and use (27.7% always or frequently) and free access but restricted use (20.7% always or frequently). When asked if they shared images in archives, institutional or thematic repositories, the researchers' response was never or rarely (85%). About 4.5% replied that they shared in scientific articles. In these two questions, no statistically significant differences were found concerning the scientific area and frequency of use.

Most researchers (77.5% never or rarely) do not mention the location where the images are stored in scientific articles. Only 14.6% answered that they always or frequently mention the storage location in scientific articles. Life and Health Sciences researchers are less likely to mention the place where images are stored in scientific articles compared mainly to Exact Sciences and Engineering but also Social Sciences and Humanities (Table 3.1). Surprisingly, researchers who use images in more than 75% of their research projects are also the least likely to mention the location where the images are stored (Table 3.2).

Regarding who they share their images with, 67.1% of the researchers that use images said they were always or frequently shared with the research group, 40.7% with the research institute/center, 30.3% with a restricted community of researchers and 13.1% with the public in general. No significant differences were found regarding the scientific area and frequency of image use.

3.2.4 Discussion

In this study, information was collected on the processes related to the use of images in research. It found that researchers often use them as research data, mainly in the Life and Health Science domain. Although they are widely produced and used in the research context, there are no guidelines that contribute to the standardization and orientation of their use. These conclusions are in agreement with a study carried out at Arab universities [42] focused on research data management in general. In the study of Arab universities [42], it was found that image was the fifth most generated research data format with a percentage of 56.4%. This study also pointed out that the causes for not drawing up a document to guide research data management (research data management plan) were as follows: lack of knowledge about what a data management plan was and how to produce it was seen as unnecessary.

In the creation process, there is a great deal of use of digital instruments, which may suggest that the digital revolution has been a key factor in the greater use of images as research data. The most significant use of the microscope in Life and Health Sciences and the traditional camera in Social Sciences and Humanities can be explained by the fact that this use is strictly related to the object of study and methodologies employed in each area.

The processes related to ensuring the quality of the photos are diverse and not all researchers carry out this activity. Likewise, the processing and analysis are heterogeneous, with no existence

of standards. This can be due to the multiple analysis options that an image may be subject to and the existence of different methodologies in the various scientific domains.

As far as description is concerned, researchers do not use metadata models to help them with this task. The same result was found in previous research[42, 128], where researchers said that did not use formal metadata standards, some researchers also reported that used informal metadata to describe their research data. When asked about the vocabulary/elements of description that researchers most use and consider relevant, the answers are varied. Only three elements were used and viewed as relevant by most researchers.

Regarding the storage location, the computer is the location chosen by the majority. This can be explained by easy access, familiarity, and recurring use of the computer in research projects. Elsayed and Saleh [42] found that most of the research stored their research data on their personal devices. It should also be noted that researchers are unaware of the volume of images produced and used during their research projects.

Although the researchers stated that they share the images used during their research, it was found that they do not do so by depositing the images in repositories that guarantee their preservation and sharing. A similar result was found by Elsayed and Saleh [42] in which the least preferred way of making data available electronically was open data repositories. These results are also in line with those obtained in a study [128] which concluded that researchers want to share their research data, but often find the process difficult.

3.2.5 Final remarks

This study identified patterns and habits in the creation, description, storage, and sharing of images. It was also possible to compare scientific areas in terms of practices and it was analyzed whether the habit of using images affected habits. The phases in which there was the greatest diversity were quality assurance and processing/analysis, due to the heterogeneity of the methodologies used by the different domains.

Since no studies are addressing this subject, the results presented are useful as they provide valuable insights into the processes and habits regarding the production and use of images in the research context. Although these are preliminary results, it was possible to see that the images are used as research data in all areas of research. Many practices are common to all areas and some differ by research area.

3.3 Exploring image management practices through interviews

As has been described in the previous sections, the formal management practices of these images are not aligned with their exponential growth. How do researchers deal with their images? Images are organized, analyzed, described, and shared according to previously established methods. This section arises to clarify what happens with data in image format. Interviews were conducted with researchers from different research domains addressing how they organize and manage the images

they use or produce in their projects. Results contribute to the definition of a current scenario, leading to proposing the conditions and the necessary steps to be followed.

3.3.1 Contextualization

This work followed on from the work described in the previous sections (3.1 and 3.2). In the first perspective, we have a practical example of a research group, in the second perspective an analysis of a diverse group of researchers that allowed corroborating results, in the third perspective we have the analysis through interviews, which allows a more direct contact with the researchers, making it possible to clarify doubts and validate the preliminary results with more certainty.

3.3.2 Methodology

To further investigate the responses and conclusions of sections (3.1 and 3.2), interviews were conducted with researchers from different domains. Although the questionnaires allow for reaching a more significant number of people, the interviews allow more profound analysis and validation of the results, provided by the direct contact that the interviews allow. Interviews are more flexible than questionnaires, allowing for a better understanding of previous conclusions and the collection of more detailed information. With the results obtained from both pieces of work (questionnaire and interviews), a holistic view of the findings can be developed, facilitating a comprehensive and general understanding of the phenomena.

The results of these interviews, added to the previous results, allow the development of a holistic view of the conclusions, that is, it facilitates a comprehensive and general understanding of the phenomena. Therefore, the generalization of all these results allows an understanding of similar processes that occur with researchers facing the same challenges.

The methodology used in this work is a qualitative method of investigation through the case study, namely exploratory research.

Exploratory research aims to improve familiarization with the case under study, so this method will investigate the topic so that it is possible to reflect on it and improve its understanding [119]. The exploratory research aims to generate ideas and hypotheses that complement this theme and help the work be developed, namely in identifying a set of habits and behaviors related to data management in image format. The case study is an integral part of the qualitative method of investigation, which aims to specifically analyze the topic in question and, in this way, create in-depth clarification about it.

In this case, qualitative research relies on primary sources in the field of empirical materials. For this, semi-structured interviews are carried out. Were recruited, by convenience, 15 researchers in 4 different domains: 4 researchers in Life and Health Sciences (LHS), 5 in Exact Sciences and Engineering (ESE), 3 in Natural and Environmental Sciences (NES), and 3 in Social Sciences and Humanities (SSH). These are the four main research domains used by the national funding agency for science, research, and technology in Portugal (country of study).

For recruitment, the choice fell on some researchers who had already participated in other projects and partnerships, as it is more agile, faster, and more likely to be successful. Table 3.3 shows the profile of the fifteen researchers interviewed. To maintain the confidentiality of the participants, we do not disclose all the information about the participants.

Table 3.3: Researchers profile

Research	Specific work area	Academic qualifications	Position in the institution	Age group
ESE1	Data science	PhD	Senior Researcher	30-40
ESE2	Enterprise systems engineering	MSc	Researcher	30-40
ESE3	Human-centered computing	PhD student	Research Assistant	30-40
ESE4	Enterprise systems engineering	MSc	Researcher	20-30
ESE5	Electrical and Computer Engineering	PhD	Senior Researcher	40-50
LHS1	Genome editing	PhD student	Research Assistant	30-40
LHS2	Animal health	PhD	Researcher	30-40
LHS3	Biological and molecular science	MSc	Researcher	20-30
LHS4	Reproductive genetics	PhD	Associate Researcher	20-30
NES1	Biology	PhD student	Researcher	20-30
NES2	Biodiversity	PhD	Post-Doc Researcher	30-40
NES3	Biology	PhD	Researcher	40-50
SSH1	Psychology	PhD	Senior Researcher	30-40
SSH2	Psychology	PhD	Researcher	40-50
SSH3	Anglo-American Studies	PhD student	Researcher	20-30

The interviews aim to give credibility and depth to this study, as they make it possible to indicate results based on real practices. Interviews can be divided into three types: informal conversation, interview guide, and semi-structured [69]. The latter was chosen because it does not require a rigid, fixed and standardized protocol, i.e. the questions do not have to be asked in an orderly manner and with a standardized formulation. So, this interview model allows the elaboration of a set of pre-defined questions, however, the guide is adaptable according to the direction of the dialogue between the interviewer and the interviewee [69].

Interviews were semi-structured by the phases of the images life cycle: Planning, Creation/Compilation, Quality assurance, Processing/Analysis, Description, Storage, and Sharing [125, 9, 131], as can see in Figure 3.19. This choice was due to the fact that this life cycle includes all the phases that we consider essential when managing data in image format. We intended to balance the importance given to all stages, showing that the success of image data management lies in the joint use of all these phases of the life cycle.

This choice was due to the fact that this life cycle includes all the phases that we consider important in image data management. We wanted to balance the importance given to all the phases, showing that the success of image data management lies in using all these phases of the life cycle together. We believe that if one of the phases is broken, the cycle cannot be completed sustainably, hence it is important to emphasize that all phases must exist in a balanced and joint way.

The interview script includes fourteen questions that are available in the INESC TEC data repository [111] and in Appendix A. Each set of questions was developed in order to understand the behaviors in each of the phases of the cycle, so that it was possible to concretely determine the actions, without mixing up the phases of the life cycle. Due to COVID-19, the interviews

took place remotely. The sessions were recorded in video and audio for later analysis. Data were anonymized, and all recordings will be deleted when the study is complete. The interviews took place from May to October 2020. For the development of the interviews, all the researchers 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 process and analyze the data from the interviews, we began by transcribing the interviews (after all interviews had already been done). A table was then drawn up for each question, where in the first column the identifier of each interviewee was placed (e.g., ESE1), in the second column, the description of the research domain, and in the third column a summary of the answer. The summary is made up of topics that include the main aspects of what was said, to make clear the specific behavior in that situation. Below, in Figure 3.18, is a summary of our methodological approach.



Figure 3.18: Our methodological approach

3.3.3 Results

This section summarizes the findings obtained in the context of the interviews, organized by the different phases of the life cycle of the images. Figure 3.19 visually represents the topics examined in the results.

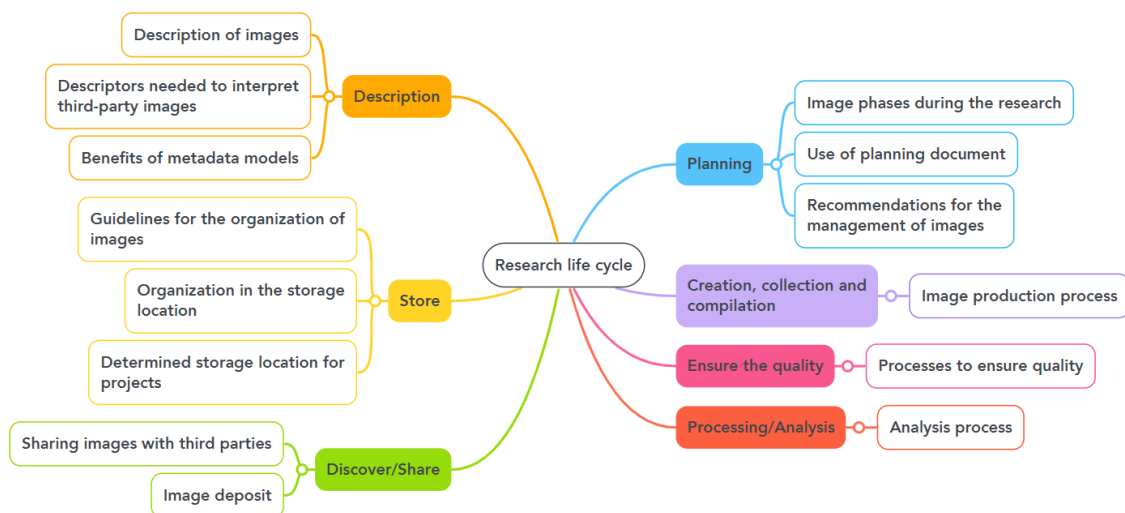


Figure 3.19: Mind-map of the topics examined in the results

3.3.3.1 Planning

The interviewees were asked whether they plan the phases that the image goes through, whether they draw up a document that establishes guidelines for image management and whether and what recommendations they would like to have for image management.

Image phases during the research

It was clear that most researchers do not formally plan their image management. However, in their research routines, they carry out practices that indicate informal planning. Were identified sequences of actions in the interviews. Table 3.4 shows these sequences, columns identify the researchers, along with their research domains, and rows represent phases of image management. Cells' values indicate the order of the stage in the overall sequence.

Table 3.4: Image management planning cycles identified by researchers

	ESE1	ESE2	ESE3	LSH1	LSH2	NES1	SSH1
collect	1				1		
select	2						
store	3				2	2	1
create		1		1		1	
treat		2		2			
use		3			3		
share		4				4	3
analyze			1			3	
investigate			2				
design			3				
assess quality				3			
refine				4			
index				5			
publish				6			
back up							2

As previously mentioned, in most cases, formal planning of image management does not exist. The images go through the cycle that the researchers, at that moment, considered appropriate. Sometimes, the image production is planned and described in advance. Other times, it appears in meetings, and then they are used for various purposes. Some researchers refer that when there is planning, the image is produced and worked collaboratively. Data observation is one of the stages that typically occurs when there is planning. In these cases, the image has a prominent role since it allows the interpretation and further processing of the data. For some researchers, if an image is reused, the image quality check is planned, and only then is it used for the research. One of the researchers said that this quality was also assessed through metadata, as this allows the information to be verified and the image to be better interpreted. Only one of the interviewees said that planning always takes place.

Of the seven cases presented in Table 3.4, it is interesting to highlight the researcher designated by SSH1. This is the one that most significantly differs from the sequences of the other researchers. This researcher, from the domain of psychology, mentioned that his practices are neither formal nor pre-established, but carried out according to what he considers to be "the most

qualified method for treating images". In the absence of a structured procedure, this researcher stores your images, then prepares a backup, as it believes that "data safeguarding is essential and its eventual loss could constitute a rupture in the research process" and, finally, shares the data with third parties, particularly in data repositories.

Use of planning document

Only two researchers said they always have a planning document, as it ensures that the team knows what has been defined and approved. Yet, the predominance is the absence of a document including standard guidelines for image management. Many say they don't feel the need for it. One of the researchers said that he knows only a few open-access guidelines and tools mainly aimed at metadata. Another researcher stated that the non-use of this document happens because there is no such culture in science, as it is not a habit in the context of research projects. Most researchers assume that they never thought about the possibility of developing this type of guide, but they consider it an advantage. One of the researchers notes everything planned and accomplished, but not in the form of a guiding document. There were no significant differences by research domain.

Recommendations for the management of images

Some researchers answered recommendations about image management are essential but do not know precisely at what stage they may be relevant. One stated that recommendations are important mainly at an early stage of the project. Another said recommendations are vital in the planning phase to affect the quality of the production and use of the images. Another researcher noted that recommendations are not essential and may stop being applicable as projects change. However, many researchers said these recommendations would be helpful. Most of them highlighted that a document with universal recommendations to guide researchers to prepare and publish their data clearly would be advantageous. They also spoke of the importance of these recommendations providing practical examples. Many researchers said that these recommendations could safeguard characteristics of the use of images and their benefits. Also, these recommendations can make researchers able to more quickly and easily assess the images they use or produce.

3.3.3.2 Creation/Compilation

At this stage, we asked how the image production process develops during the investigation.

Image production process

One of the ESE researchers said that he produces many photographs, maps, and graphs. The process involves documenting the code, rerunning the code, and generating the image again. Other ESE researchers said that this process could vary, but, usually, existing images are chosen and

treated. The image reference (name that clearly identifies the image) for the future is always kept. The researchers also mentioned that images can come in two forms: 1) they are produced spontaneously when the immediate need arises 2) they are planned in advance. In several cases, the images are produced collaboratively in this research domain.

In the domain of LHS, all the researchers said that the production process depends on the type of image. Production processes change with the types of images. Graphics are usually associated with data collection, data organization, image creation, image refinement, and treatment. Graphical abstract and infographics - root creation. Western plots - created photography, simple editing, and comparison of the bands. The creation of histograms, boxplots, funnel graphs and different types of graphs was also mentioned. Researchers also create explanatory diagrams and flowcharts, usually created with software.

The NES researchers mainly produce images in geographic information systems, where they create the map and add informational elements. However, this process depends on the type of investigation. This issue is often seen as a part of the research, where data collection and analysis processes are also considered. Satellite images are usually downloaded and pre-processed to identify defects, apply corrections, and georeference the image.

The SSH researchers say the images can be produced or reused, depending on the type of research and its purpose. In socio-cultural projects, as a rule, images are produced. When dealing with historical or conceptual projects, the images are reused from repositories, articles already published, or past projects. Regarding production, all the researchers in the area mentioned the importance of anonymization. For this reason, the production process consists of collecting, and blurring the characteristics of identification, storage, and use.

3.3.3.3 Quality assurance

In this section of the interview, interviewers were asked about the existence of processes to ensure the quality and the activities inherent to this.

Processes to ensure quality

Most researchers say there is little or no activity to guarantee or verify the quality of the images. One of the researchers explicitly stated that there is no such check, but said that if the image "looks good" and reflects what is intended, it is approved. One of the researchers noted that this verification occurs mainly in western plots. For some interviewees, verification only occurs when they fail to perceive the original image. Some researchers claim that, although images are critical to projects, they are not published and shared most of the time, so a raw image is sufficient to analyze what is intended. One of the researchers said he checked the quality when it comes to videos and photographs, as the details of these images can be critical. Some researchers mentioned adjustments to the code, cropping of images, treatment of image noise, care with contrasts, verification of the dimensions and source of images, aesthetic validation of the image, and verification of spatial resolution, temporal resolution, and georeferencing.

3.3.3.4 Processing/Analysis

For this topic, interviewers were asked about the image analysis process, namely how information is extracted from the images.

Analysis process

Most interviewees do not have a pre-established method of analyzing/processing the images. Most researchers said they did not analyze the images. Some researchers claim to do this frequently with methods that depend on the images and the project. One researcher stated that this analysis/processing depends on who performs it, as he will check what will be necessary according to the type of image. One researcher that works exclusively with maps said that this type of image is already the final product. Researchers identified the semi-automatic classification of images, the geostatistical processes, the extraction of direct or indirect information from the image, content analysis (especially for videos and photographs), information extraction (for graphics), visualization of the data, and the annotation of the image results as analysis/processing methods. Researchers also mentioned visual analysis, measuring distances, checking details, and cutting out irrelevant parts.

3.3.3.5 Description

Were asked to interviewees how they describe images, which information they need to understand and use third-party images, and the benefits and interest of a metadata model for image description.

Description of images

Few researchers describe their images. The description is not a frequent practice, and its realization depends on the context, the image, and the researcher who produces it. For one of the researchers, "less is more", that is, too much description can affect the purpose of the description and confuse whoever tries to interpret it. We identified the following examples of description: 1) description through technical details (for those who already know the context of the image) and general details (for those who do not know), 2) description through captions (most mentioned), 3) elaboration of metadata (most used to describe components of maps, production periods, dates and the purpose of the image), 4) description through annotations in a laboratory notebook. For researchers who describe images, the advantages include sharing metadata/captions for better interpretation and dissemination of the study.

Elements needed to interpret third-party images

The most mentioned elements are the source, rights of use, author, place of origin, title, context, characteristics of obtaining and processing the image, caption, scale, related resources, methodology, description, date, production method, treatment, and analysis processes, spatial resolution,

temporal resolution, abstract, keywords, and the area of study. Elements of authorship, date, context (abstract, description, keywords), methodology, production techniques, and study area were the most mentioned and are transversal to all research domains. Some elements vary according to the particularities of the research domain.

Benefits of metadata models

Researchers see metadata models as very important. Only one researcher said he wasn't sure, as he didn't know if the results would compensate for the time spent describing them. The other researchers responded positively to the possibility of using metadata models, stating that they will facilitate the storage of relevant information about an image. In addition, the metadata model would serve as a recommendation to describe, a description standard to follow, and would facilitate description for researchers. One researcher said that the model should be adapted to the type of investigation. Another researcher thinks it is difficult for a model to reach a balance of elements. On the one hand, there are general recommendations that have general applicability and are very relevant. On the other hand, a generic model can leave out important details, such as the context and the scientific area.

Almost all researchers mention that it is helpful to have guidelines for describing images. They also claim that it can be very advantageous to have a cross-cutting model for multidisciplinary research groups to harmonize and create common communication points. They highlight the importance of a core of generic and specific metadata depending on the area. Researchers stress that the metadata model can function as a working tool and help members of the research group follow the same working methods, avoiding deviations. The metadata model can also stimulate the publication and sharing of data in repositories and directories.

3.3.3.6 Storage

The researchers were asked if they define guidelines for the organization of the images, what is the organization of the images in the storage location, and if there is some central location where the images are stored.

Guidelines for the organization of images

No researcher defines guidelines for organizing their images. One said that he does not do it formally but pays special attention to the formats, the way they are organized, and the place where they are stored. Another admitted that this theme is a general failure, mainly because each researcher stores his own images, which is not safe and does not provide access to other research group members. Many researchers admit to having their method of organization and storage, but it is not formal or discussed in the working group. Everyone says they give importance to the way they organize the data, but it is not a procedure stipulated by the project's guidelines.

Organization in the storage location

All researchers claim to organize the images into project folders and subfolders on the computer. Some affirm organizing project folders by content (separating the content into categories that they create). Some researchers divide the raw data from the processed data. Others have folders for analysis and folders for scientific articles, each having different versions of images.

Determined storage location for projects

No ESE researcher has a storage location, except one researcher who said that the images are stored on the computer and accessible via the internal network. In the LSH domain, only one researcher claims to have a storage location for the project's images. This researcher said that the existence of this location depends on the project. If it is a small project, all the images are on the computers; if it is a large project, a folder is created in the cloud for everyone to access. Two of the NES researchers say that this central location exists. One said that it is on the computer or the external disk and that, sometimes, it is in both places at the same time. Another stated that they have a Network-Attached Storage (NAS), a device dedicated to the storage of network data, allowing homogeneous access to data. For the SSH domain, only one of the researchers stated that this location exists. He said that they have an external disk where they store all the data. This disk is kept in the group's laboratory and accessible to project members. Researchers also have a copy of this data on their computers.

3.3.3.7 Sharing

In this last phase, the interviewees were asked with whom they usually share their images, what motivates them to do so and whether they had ever uploaded an image.

Sharing images with third parties

Some researchers share their images with group colleagues in presentations to colleagues, at conferences, in articles, and on posters. Researchers also share the images with people associated with the project for specific tasks. Some share images with colleagues in the group to receive feedback on the image, to know if it is readable if it makes sense, and if it is reliable. One researcher never shares the original images. Another said that it only shares images in articles. Most researchers share images after being treated because others may not easily interpret a raw image. One of the researchers says he is careful to accompany the image of a caption with the context whenever he shares it. The reuse of data is one of the main reasons for sharing and the possibility of new collaborations, sharing new knowledge. Some researchers point to confidentiality and sensitive data as barriers to sharing. The fact that images are not central to the project also limits sharing.

Image deposit

Only some researchers have already deposited their images in a joint deposit that also included other types of data. One of the researchers said it was unusual and that he didn't normally hear about it. Another researcher deposits data but never deposited only images. This researcher said he did not see any relevance in depositing the individual images. One of the researchers said he knew that depositing is one of the data management requirements but has not yet done so with the images. One of the researchers said that sometimes deposit their data but consider that it is more beneficial to share the information that gave rise to the images and the results than the image itself. A part of the interviewees claim to see no use in depositing the images in isolation; it would only make sense accompanied by textual information. These researchers say that they only make deposits if it is a requirement of the conferences or scientific journals. Deposit normally occurs in repositories.

3.3.4 Discussion

Were interviewed a group of researchers about how they proceed and see the management of imagery data. It became noticeable that all researchers produce or use images, but these are not always essential parts of the study. Therefore, many researchers are not concerned with its management. The results coincide with those collected by questionnaires (Section 3.2), where the generality said they did not have a guiding document or formal practices for this purpose. Although images are widely produced and used in research, there are no guidelines to direct researchers to standardized practices regarding images.

In the creation process, it is visible that the methods differ by research domain. In this study, the use of digital instruments stands out, predominantly in ESE, LHS and NES, which was also confirmed in the previous survey (Section 3.2). At certain times, researchers pay attention to quality assurance, mainly in the domains of LSH and NES. However, there were no standard procedures to do so, varying with the type of study. In the questionnaires (Section 3.2), not all researchers perform this task, which also happens here. Regarding the processing/analysis phase, most researchers admit that it occurs according to the situations and context of the work, without any formal method or common practice in the work group. In the questionnaires (Section 3.2), it was identified as the same.

Most researchers do not use metadata models. They consider it relevant, but it is not a common practice. The description is usually done through captions. They believe that when the image is reused, metadata is essential for interpretation, and the use of a metadata model for projects would bring benefits. The previous study, through questionnaires (Section 3.2), also shows that researchers do not use metadata models. However, in this study, more comprehensive responses were collected regarding the elements that researchers consider relevant. All the researchers were able to find a justification for choosing a specific element. For example, the keywords were mentioned several times, as they are essential for a preliminary analysis of the relevance of the data. Regarding storage, researchers do not have formal habits or standardized methods of organization. The computer prevails as a storage location, as it is easily accessible. The same was evident of

the questionnaires (Section 3.2). Finally, the researchers showed that no sharing habits. Sharing demonstrations take place mainly in scientific articles and conferences. Internally within the project, sharing occurs frequently. They have also not been shown to have depositing practices in repositories or the like. The same results were obtained in the questionnaires (Section 3.2).

This work allowed confirming several results obtained previously with a questionnaire. The use of the interviews allowed close contact with the researchers. If the questionnaire makes it possible to collect factual information, the interviews were essential to determine each of the choices, obtain justifications for answers, understand the various topics, and understand the researchers' predisposition to initiate image management practices.

It was found that most researchers are aware of the RDM guidelines, but that this subject is not usually discussed at group/project meetings. Researchers also do not yet assume that imagery data is amenable to management, for many researchers as images are complements to the work and, therefore, their management is not essential. For most researchers, the time spent is still not worth the effort in data management unless that is an obligation. It became clear that the existence of practical and easy-to-implement standards and models (metadata model for image description, for example) is welcomed by researchers, as it is a way of seeing the work done with little effort. It was clear that the type of image is one of the most significant differentiating factors for the type of management carried out. The research domain brings particularities, but it does not always determine all differences. Sometimes, the type of images and the conditions leading to their acquisition are more differentiating than other characteristics.

During the research phase into the literature on research life cycles, it became apparent that none of the options included data security and privacy procedures as one of the main phases. However, issues related to data confidentiality, namely images produced or used in the research context, must be reflected. In fact, an image can indicate, for example, a place or a person, thus breaking the principles of confidentiality that are often desired.

The General Data Protection Regulation (GDPR), approved on April 27, 2016, launched new challenges to the various audiences that the sphere of privacy affects, from citizens, companies, organizations, educational institutions, and research centers. The protection of individuals concerning the processing of personal data is a fundamental right that needs to be defended, regardless of the context. With the validity of the Regulation, a solid and more coherent protection framework at the European level was demanded. Personal data are understood to be those that allow the identification of the data subject, such as name, identification number, location data, identifiers electronically, and data related to physiology, genetics, mental health, economic, cultural, or social situation. However, if an image shows or indicates any of these cases, it may compromise the protection of the author's data or the actors highlighted in the image.

In the context of the research, some guidelines have already emerged to prevent this problem. Data Management Plans (DMP) and Privacy Impact Assessments (PIA) are very useful tools in data protection. The DMP allows for structuring and organizing data from research projects, playing a central role in the development of good research practices, and contemplating privacy actions to be applied to data. PIA was introduced through the General Data Protection Regulation (article

35). The PIA aims to prepare a document with guidelines that intend to direct and control privacy risks. This document must foresee the risks and establish concrete scenarios associated with the research data, to anticipate problems and define strategies for the success of the management.

Therefore, it would be important to start a reflection on the clear introduction of a privacy phase in the research life cycles, so that practices related to this topic are mandatory for researchers who want to fully comply with security and data confidentiality.

3.3.5 Final remarks

There is still a way to go in the sphere of research data management. Researchers tend to agree that the RDM is vital in science, but a lack of knowledge and time is often a hindrance to progress. Within the various types of data, the text is the one we see most often associated with data management practices. However, the rest of the typologies should not be overlooked. Images, for example, can be valuable instruments in scientific production. Their capacity for representation can help researchers. Besides, the use or production of images is not limited to a domain of research. Allied with this, the enormous technological evolution and the new capture devices that allow to obtain and spread an image quickly.

It is visible that when confronted with this reality, researchers realize what habits they have regarding images. However, they do not usually set guidelines and plan methods for managing their images. Most of the behaviors highlighted in this work mainly occur informally and without conformity of the working group. The path is to continue to reflect on these issues, confront researchers with these issues and motivate data management practices to become recurrent. Once implemented in the working group, they are easily adapted to the various projects, bringing significant benefits.

The work we have been doing in the area of RDM at INESC TEC suggests that there is a growing awareness and interest in the development of image data management practices on the part of the research community. First, more and more researchers look to us to help them in the process of organization (through the development of data management plans), description (through the use of metadata models), and deposit/sharing (through the choice of repository and assistance in the filing process). Second, more researchers are concerned about reuse issues of their images, especially authorship issues, showing interest in knowledge about license and rights issues. The possibility of attributing DOI to the data has been an incentive to share. Finally, researchers almost always admit that a data management policy would be very helpful. This policy does not arise for lack of time or in-depth knowledge. In fact, some practices are beginning to be applied in projects, in order to initiate, even if superficially, behaviors that lead to good image management. Therefore, we believe that this study is necessary for the definition of best practices of research data management, namely image management, as it allows for establishing a scenario and, from there, defining a set of guidelines and offering tools for image management. In addition, this study makes it possible to strengthen the existing state of the art, through analysis of real practices.

Finally, talk about the challenges that images bring to RDM. Information systems are better equipped to handle text. Nowadays, technology allows extracting automated the content of a text

with some ease, but from an image, it is more difficult. Therefore, it is essential to have good description mechanisms that contribute to the search/findability of data in image format.

3.4 What does literature tell us about research image management practices?

All the work carried out in the previous sections has been essential in determining the behaviors and habits of researchers who use and produce images. However, the studies followed a methodological approach that allowed the results to be obtained from a range of stakeholders who were confronted with the topic, thus having a prior framework of the subject in question. Given this circumstance, it was necessary to explore this topic from the point of view of what the literature tells us about these behaviors and habits, in order to align results with previous studies and find results obtained without any framework of the topic of study, but rather from a structured search for it.

3.4.1 Contextualization

Science benefits from images since the possibility of accurate registration enhances research. In addition, images are helpful for visualization and communication in science. The abundance of images, diversity of origins, and diversity of holders create various practices regarding description, interpretation, and systematic use. It is essential to understand the behavior of researchers using and producing images and to know more about their habits. What kind of images are produced by researchers? How are images collected? Which instruments are used in this process? Are images described and deposited in repositories? These responses will allow, for example, to identify failures and propose good practices.

To study the use and production of images in research, 109 articles, from several domains where images were used either as data or metadata, were systematically reviewed. This work aims to answer several starting questions (SQ), namely:

SQ1: In the published works, is it possible to identify formal practices and standards in managing images as research data?

SQ2: Do research domains manage images differently?

SQ3: Are researchers aware of data management practices and policies?

SQ4: Do image-related habits and trends tend to change over the years?

This study aims to finalize the chapter on the informational behavior of researchers using and producing images. The aim was to complete this analysis using various methods and sources of information. From the study of the Generation.Mobi project (Section 3.1), through the questionnaires (Section 3.2) and interviews (Section 3.3), to this review that aims to report what the

literature says about the use and production of images in a scientific context, it is expected to validate preliminary diagnoses and consolidate results obtained, to give consistency and veracity to the study.

3.4.2 Methodology

To understand how images are used, produced, and managed in research contexts, a systematic review of the literature was carried out with slight adaptations in the way articles are identified and analyzed. After selecting the information sources and defining the search strategy, three main stages guided our methodological approach (Figure 3.20). The following sections detail these phases.



Figure 3.20: Methodological Approach

3.4.2.1 Article selection

The overall selection process followed the PRISMA statement [75] and is described in Figure 3.21.

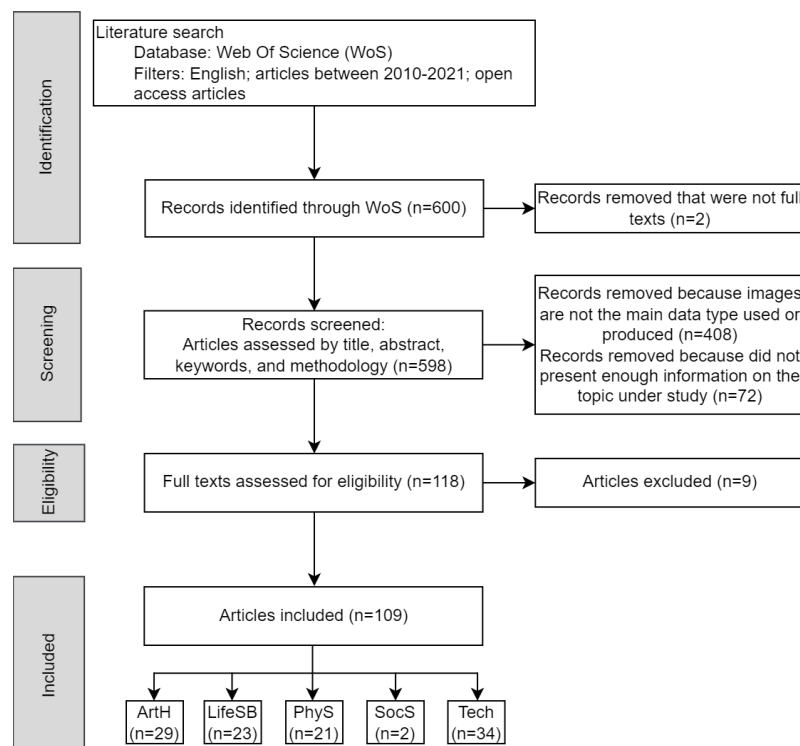


Figure 3.21: Flow diagram of study selection

It was decided to use Web Of Science (WoS) for its prestige, extensiveness, and organization. In this work, the five main WoS research areas were considered: Arts and Humanities (ArtH), Life Sciences and Biomedicine (LifeSB), Physical Sciences (PhyS), Social Sciences (SocS), and Technology (Tech). Instead of defining a search query, as is usually done in systematic reviews, the most relevant articles per domain of research were included. Note that it is not objective to collect articles related to a specific topic. Instead, it is intended to bring together the most important articles from each discipline and analyze their image management practices. To choose the most relevant articles for each domain, the 10 most cited articles on the Web of Science were selected by year between 2010 and 2021. In addition, articles should be in the English language and in open access. In total, 600 articles were identified.

The screening process involved reading the title, the abstract, the keywords, and the methodology of each of the 598 articles (two were excluded for not having the full text). In this reading, the aim was to see if the positioning of the images, namely their role as research data, was sufficiently relevant. Many articles used or produced images. However, images were often not the primary data type used or produced, meaning that the authors did not highlight the relevance of images in the research work. When this happened, the articles were excluded. It also happened that, in some cases, there wasn't enough information on the subject under study. That is, the topic of the images seemed to be relevant in the article, however, the present information did not provide sufficient details to consider including the article in the final sample. These details are, for example, the method of producing the images, the type of instruments used, the type of images, formats, quality and editing techniques, and the type of analysis involved, among others.

In the eligibility stage, articles were read in full to decide on their inclusiveness in the final sample. Nine of them did not include sufficiently comprehensive information for the study in question and were, therefore, excluded.

Of the 600 documents analyzed, 109 (18.2%) constituted the final sample. The articles that caused any doubt about the inclusion/exclusion were read almost in full. Of the 109 documents, 29 (26.6%) belong to the domain of Arts and Humanities, 23 (21.1%) of Life Sciences and Biomedicine, 2 (1.8%) of Social Sciences, 21 (19.3%) of Physical Sciences and 34 (31.2%) of Technology.

3.4.2.2 Article analysis

All 109 articles were downloaded and organized in folders by research domain and year of publication. In each folder, we included a document with the articles' titles, authors, number of citations in the primary collection of Web of Science, the total number of citations, and accession number.

A large number of articles, their multidisciplinary, and the need to facilitate the subsequent analysis of the selected papers made us define 13 dimensions of analysis shown in Table 3.5. Each of them is composed of a controlled vocabulary, defined as we analyzed the articles, considering the research and specific terminology. The controlled vocabularies functioned as an information tool that facilitated the description of the articles and the analysis process.

A spreadsheet was created to describe each of the selected articles. For each article, the title, domain of research, year, ID (assigned sequentially) and the 13 dimensions of analysis were included (Table 3.5). After reading each complete article, this spreadsheet was filled in. The complete spreadsheet will be made available in open access, through a dataset published in a data repository (after the article review phase).

Table 3.5: Dimensions of analysis

<i>Dimension</i>	<i>Possible values</i>
Image as data?	yes, no
Image as metadata?	yes, no
image type	animation, drawing, engraving, graphic, illustration, medical images, microscopic images, painting, photography, plant, portrait, video, computer-made image, map, print screen
collection technique	images produced by researchers in the laboratory, images produced by researchers outside the laboratory, images produced by the study participants, images collected by researchers from third parties, images collected by researchers from previous projects
programs used	ImageJ, Mathworks, Matlab, SPSS, other
instruments used	microscope, medical instrument, camera, manual writing, computer
analysis type	qualitative, quantitative
licenses	CC BY, CC BY-SA, CC BY-ND, CC BY-NC, CC BY-NC-SA, CC BY-NC-ND, Nature, APA, Elsevier, ACS, AAAS
image format	analogic image, digital image
use and production?	use, production
image editing	correction/editing, reframing footage, cropping, composition, adjust of scale/calibration, without editing
image description	structural metadata, non-structural metadata
image deposit	yes

3.4.2.3 Statistical analysis

The Statistical Package for the Social Sciences (SPSS) was used to analyze the data collected. As shown in Table 3.5, all the dimensions of analysis are nominal variables. A descriptive analysis was then carried out using cross-tabulation. For inferential analysis, we used the Chi-square test for equality of proportions for comparisons between domains, and Cochran's Q test for comparisons between dimensions analysis.

3.4.3 Results

After describing the methodology, we describe the results organized by the different phases of the research life cycle, as defined by the Data Documentation Initiative [125], DataONE [9], and UK Data Archive [131]. The stages of the research data lifecycle are planning, creation/compilation, quality assurance, processing/analysis, description, storage, and sharing. Subsequently, the results are discussed and our conclusions are presented.

The results related to the planning and storage phases will not be reported because none of the articles included information on this phase.

It is important to note that for the Social Sciences domain, only 2 articles were collected, so the conclusions related to SocS may need further analysis.

All analyzed articles are available in a dataset generated with all the information collected from each of the documents. This dataset is deposited in the INESC TEC data repository, in open access [[116](#)].

3.4.3.1 Planning

For this study, it was noted that none of the articles contained information on the planning phase. This phase is related to the moments prior to the execution of the projects, in which documents with guidelines for the course of the project are prepared. At this stage, you can build data management plans, documents to safeguard sensitive data, or others that involve structured planning of actions throughout the project. We can't say that such planning doesn't exist, we just know that it hasn't been reported by the authors of the article.

3.4.3.2 Creation/Compilation

The creation and complication phase is inherent to the types of images that are produced or used, as well as the instruments used in these moments. Furthermore, this phase relates to the role of images in research, namely whether they are originally produced or used by third parties. Finally, the creation and complication phase makes it possible to determine the scenarios in which the images were produced, that is, in which place and under what conditions their production took place.

Types of image

It was possible to identify the type of image used in each article. In total, eight types of images were identified, namely, computer-made images (images generated in the field of computer graphics, often referred to as computer-generated imagery (CGI)), drawings, illustrations, medical images (e.g., x-ray, tomography, magnetic resonance imaging (MRI)), microscopic images (images generated exclusively through the use of microscopic), maps, photographs, and videos.

It is important to note that an article may use/produce several types of images, so the number of types of images per domain may be greater than the number of articles per domain. Table [3.6](#), showed the domains that use each type of image, how frequently, and if this frequency is significantly different from other domains' frequency. In each domain, the significant differences between the use of different types of images were also analyzed.

Table 3.6: Images type used per domain and significant comparisons

Image type	Arth		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	% (n=29)	n	% (n=23)	n	% (n=21)	n	% (n=2)	n	% (n=34)	n	% (n=109)	
photography	16	55.17%	5	21.74%	18	81.82%	1	50%	9	26.47%	49	44.95%	PhyS>Arth*, LifeSB>Tech*, Arth>LifeSB*, Tech*
microscopic image	2	6.90%	16	69.57%					16	47.06%	18	31.19%	LifeSB>Arth*, PhyS*, SocS*, Tech>Arth*, PhyS*
medical image	1	3.45%	2	8.70%					9	26.47%	12	11.01%	Tech>Arth*, PhyS*
map	2	6.90%	3	13.04%	3	14.29%	1	50%			9	8.26%	PhyS>Tech*, LifeSB>Tech*, SocS>T*, Arth*
video	7	24.13%									7	6.42%	Arth>LifeSB*, PhyS*, Tech*
computer-made image	2	6.90%			1	4.76%			2	5.88%	5	4.59%	
drawing	4	13.79%									4	3.67%	Arth>Tech*
illustration	3	10.34%									3	2.75%	
Significant comparisons between image type	photo>micro*, medical*, map*, computer*		micro>all*		photo>all*				micro>map*, video*, computer*, draw*, illust*		photo>medical*, map*, video*, computer*, draw*, illust*, micro>medical*, map*, video*, computer*, draw*, illust*		

An * indicates significant results with $\alpha < 0.05$.

When no value is mentioned, this is equivalent to having 0%.

% are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

An analysis by domain shows that photography is the type of image most present in Arth, standing out significantly. In LifeSB, this highlight is in the microscopic images. Photography is again the most present in PhyS and SocS. In the Tech domain, the emphasis on one type of image is not as pronounced as in the previous ones, even so, microscopic images prevail. This last fact is surprising and, after analyzing the 16 articles in question, it was found that these studies concern the development, implementation and analysis of technologies that allow microscopic images to be captured and studied. In the domain of LifeSB, it would be expected that medical images would have a more pronounced presence, however, after analyzing the articles, it was found that in this domain the studies are closely related to the microscopic analysis of diseases or medical conditions and also to the proposal of medical approaches for decoding microscopic images.

Considering image types, it can be seen that photographs and microscopic images are the only ones that are significantly larger than other types. Another interesting fact is that videos, illustrations, and drawings only appear in the domain of Arth. In addition, the last two are associated with analog images. Finally, it is important to emphasize that photographs are the only image typology present in all research domains.

Types of instrument

As with the types of images, it was important to identify the main types of instruments associated with the production of the images under study. The following instruments were identified: camera, computer, medical instrument, microscope and paper. Table 3.7 shows the use of instruments by the research domain.

Table 3.7: Instrument type used per domain and significant comparisons

Instrument type	ArtH		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	% (n=29)	n	% (n=23)	n	% (n=21)	n	% (n=2)	n	% (n=34)	n	% (n=109)	
camera	17	58.62%	5	21.74%	18	85.71%	1	50%	9	26.47%	50	45.87%	ArtH>LifeSB*,Tech* PhyS>ArtH*, LifeSB*,Tech*
microscope	2	6.90%	14	60.87%					16	47.06%	32	29.35%	Tech>ArtH*,PhyS* LifeSB>ArtH*,PhyS*
computer	6	20.69%	3	13.04%	4	19.05%	1	50%	2	5.88%	16	14.68%	
medical instrument	1	3.45%	3	13.04%					9	26.47%	13	11.93%	Tech>ArtH*,PhyS*
paper	5	17.24%									5	4.59%	ArtH>LifeSB*, PhyS*,Tech*
Significant comparisons between instrument type	camera>all*		micro>computer*, medical*, paper*		camera>all*				micro>computer*, paper*		camera>computer*, medical*, paper*, micro>medical*, paper*		

An * indicates significant results with $\alpha < 0.05$.

When no value is mentioned, this is equivalent to having 0%.

% are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

It can be seen that, as expected, the type of instrument is directly related to the image type. In other words, images such as photographs and videos are produced with the camera; medical images with medical instruments; microscopic images with the microscope; drawings and illustrations are produced using manual writing and maps, graphics, and computer-made images with the computer. It is logical for researchers to know which instrument is used in the production of an image when they are the ones who produce it. However, when they don't produce it, they might not have this information. However, in the analysis of the documents, only two of the documents do not present information in relation to the instrument and this happens when the images are produced by third parties.

In the domains, it is possible to highlight that the camera has a more significant presence in ArtH and PhyS. The microscope, on the other hand, has this highlight in LifeSB and Tech. Paper only exists in ArtH. In terms of types of instruments, the camera, in general, has a more significant presence than the others.

Use and production

When planning a research project, the type of images and instruments to be used are not the only topics to consider. When it comes to data in image format, it is necessary to plan how the images will be acquired and included in the investigation. In a research project, an image can be the work of the project's researchers and can therefore be considered to have been produced by them, or it can be reused from third parties, in which case it is considered to have been used. All documents under analysis mentioned the type of image acquisition. The labels "production" and "use" were created to classify the images according to their acquisition. Table 3.8 shows the use and production of images by the research domain and the associated significant comparisons.

Table 3.8: Use and production of images in research domains and significant comparisons

Use and production	ArtH		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	% (n=29)	n	% (n=23)	n	% (n=21)	n	% (n=2)	n	% (n=34)	n	% (n=109)	
production	15	51.72%	23	100%	17	80.95%	2	100%	34	100%	91	83.49%	LifeSB>ArtH*,PhyS* Tech>ArtH*,PhyS* PhyS>ArtH*
use	15	51.72%			4	19.05%					19	17.43%	ArtH>LifeSB*, PhyS*,Tech* PhyS>LifeSB*,Tech*
Significant comparisons between use and production			production>use*						production>use*				

An * indicates significant results with $\alpha < 0.05$.

When no value is mentioned, this is equivalent to having 0%.

% are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

The ArtH domain presents a homogeneous distribution of results by use and production, the remaining domains have a significant prevalence of production. All papers from LifeSB, SocS, and Tech have produced their images.

The analysis of the articles made it possible to perceive that the researchers privilege their own production of the images, and when the images are used by third parties, there is not much information about the method used in this image reuse.

Collection scenarios

To systematize the different ways of collecting the images, two scenarios in which they may occur were studied, namely: images produced by researchers in the laboratory (inlab) and images produced by researchers outside the laboratory (outlab). Table 3.9 shows the methods for collecting images by research domain.

Table 3.9: Collection scenarios per domain and significant comparisons

Collection technique	ArtH		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	% (n=29)	n	% (n=23)	n	% (n=21)	n	% (n=2)	n	% (n=34)	n	% (n=109)	
inlab	8	27.59%	19	82.61%	4	19.05%	1	50%	26	76.47%	58	53.21%	LifeSB>ArtH*,PhyS* Tech>ArtH*,PhyS*
outlab	9	31.03%	4	17.39%	11	52.38%	1	50%	6	17.65%	31	28.44%	PhyS>LifeSB*,Tech*
Significant comparisons between collection technique			inlab>outlab*		outlab>inlab*				inlab>all*				

An * indicates significant results with $\alpha < 0.05$.

When no value is mentioned, this is equivalent to having 0%.

% are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

It is possible to verify that the inlab and outlab scenarios depend on the research domain. The inlab scenario is most representative in the LifeSB and Tech domains. In turn, the outlab inside scenario is more prominent in the PhyS domain. There are several examples taken from articles about inlab and outlab scenarios. Even so, it is possible to identify that studies involving microscopic data, medical data, and technological development are more associated with the laboratory. Georeferential data, observations, and astronomy-related data are more associated with scenarios outside the laboratory.

It is also interesting to note that five articles mentioned that their data were collected using study participants. Of these five cases, four belong to the ArtH domain and one to the LifeSB domain. Analyzing the articles, we realize that these cases fall mainly on ArtH, as they are related to social studies in which participants were recruited to determine behaviors and establish standards, through a series of questions raised by the authors of the studies.

3.4.3.3 Quality assurance

The growing production of images is undeniable. Nowadays, through image capture devices, this is available to anyone. However, it wasn't always like this. You only have to think back to the early years of the current millennium to see that most images were analog. Currently, the trend has reversed and the majority of the images produced are digital, as the available technologies allow the production or capture of an image quickly, allowing its storage or sharing at the same moment, without the constraint of places, times, or devices.

One of the methods most used by researchers to ensure the quality of their images is the format in which they are stored. In this work, the format of the images in the documents was analyzed. For this, the images were divided into two groups: analogic image and digital image. Table 3.10 presents the information regarding image formats by research domain.

Table 3.10: Image format per domain and significant comparisons

Format	ArtH		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	% (n=29)	n	% (n=23)	n	% (n=21)	n	% (n=2)	n	% (n=34)	n	% (n=109)	
digital	26	89.66%	23	100%	21	100%	2	100%	34	100%	106	97.25%	Tech>ArtH*
analogic	6	20.69%									6	5.50%	ArtH>LifeSB*,PhyS*,Tech*

An * indicates significant results with $\alpha < 0.05$.

When no value is mentioned, this is equivalent to having 0%.

% are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

It was possible to identify the image formats in all the documents analyzed. The predominance of the digital format is evident. With regard to documents combining analog and digital images, only one was identified. The ArtH domain was the only one to present analog images, namely three drawings and two illustrations. With both analog and digital formats, we have one drawing.

In the case of digital images, the most mentioned formats were JPEG (Joint Photographic Experts Group), PNG (Portable Network Graphics), PX (Pixel Image), TIFF (Tagged Image File Format), ICN (Windows Icon File), MNG (Multiple image Network Graphic) and PSD (Photoshop Document). In the case of analog images, the only medium mentioned was paper.

3.4.3.4 Processing/Analysis

The type of analysis and processing chosen in the context of research are determining factors in the data management process. The processing and analysis phase is related to the edits made or not to the images, as this information can be decisive in the treatment and study of the data since they can change the original characteristics of the images. This phase also addresses the type of analysis to which the images are submitted, as well as the programs used to perform this task.

Image editing

The production of images sometimes implies that they have to go through an editing process. The improvement of an image can include aspects such as cropping, and adjusting brightness, including the scale, among others.

This work analyzed how the images were edited. For each document, we checked whether or not the images had been edited. 84 documents did not include this information. For this reason, it is not possible to say whether the images were edited or not. In the others, we identified five types of editions: adjusting of scale/calibration, correction/editing, composition, reframing footage, and cropping. In addition to these, it was decided to include in the analysis documents that had unedited images and documents in which this information was not available. A document can have more than one type of image editing associated with it. Table 3.11 systematizes this information.

Table 3.11: Image editing per domain and significant comparisons

Image editing	ArtH		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	% (n=29)	n	% (n=23)	n	% (n=21)	n	% (n=2)	n	% (n=34)	n	% (n=109)	
without editing	9	20.69%	5	21.74%							14	12.84%	ArtH>PhyS*,Tech* LifeSB>PhyS*,Tech*
correction/editing	1	3.45%			2	9.52%			1	2.94%	4	3.67%	
adjust of scale/calibration					1	4.76%			1	2.94%	2	1.83%	
cropping	1	3.45%							1	2.94%	2	1.83%	
composition					1	4.76%					1	0.92%	
reframing footage	1	3.45%									1	0.92%	
Significant comparisons between image editing	without>all*		without>all*								without>all*		

An * indicates significant results with $\alpha < 0.05$.

When no value is mentioned, this is equivalent to having 0%.

% are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

The adjustment of scale/calibration, composition, and reframing of the footage appeared only once. Cropping appeared twice. Correction/editing appeared most frequently and therefore stands out from the other types of editing. Fourteen documents explicitly identified that the images were not edited. This factor is important, as for some types of studies it is important to keep the image intact, as only its originality will achieve the expected results. An edition could alter the "truth" of that reality. Images without editing were only identified in the ArtH and LifeSB domains. However, it is important to remember that it is not possible to say that in other areas this has not happened, there is just no information about it.

Analysis type

Another process related to image processing/analysis is the type of analysis that can be performed. The analysis used for the images is an important feature to verify. In the context of this study, two types of analysis were determined: qualitative analysis and quantitative analysis. Table 3.12 presents this information.

Table 3.12: Analysis type per domain and significant comparisons

Analysis type	ArtH		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	% (n=29)	n	% (n=23)	n	% (n=21)	n	% (n=2)	n	% (n=34)	n	% (n=109)	
quantitative	4	13.79%	22	95.65%	10	47.62%			9	26.47%	45	41.28%	LifeSB>all* PhyS>ArtH*
qualitative	20	68.97%	1	4.35%	9	42.86%	1	50%	6	17.65%	37	33.94%	ArtH>LifeSB*,Tech* PS>LifeSB*,Tech*

An * indicates significant results with $\alpha < 0.05$.
 When no value is mentioned, this is equivalent to having 0%.
 % are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

The ArtH and SocS domains are those in which qualitative analysis prevails. This result was to be expected, since studies in the social and human domain are closely associated with qualitative analysis techniques. In the cases under study, the works were related to filmed interviews, image content analysis, and the determination of psychological pathologies through images obtained by patients and study participants. In LifeSB, PhyS, and Tech, quantitative analysis prevails. The first one with a lot of prominences, which is also an expected result. In PhyS and Tech, the distribution is more balanced, however, given the character of these domains, more emphasis on quantitative analysis would be expected.

When analyzed in detail, it appears that the articles, in several cases, do not present the information related to processing and analysis. In fact, any process that involves data, as a rule, has an associated analysis. However, this information is not always reported in the articles, as they tend to have summarized information, as has been seen in several cases throughout this study.

The type of analysis performed on the images can often be related to the scenario in which these images were collected. For this reason, it is interesting to see what kind of correlation exists between these two parameters. When checking the relationship between the type of analysis carried out on the images and the collection scenarios, it emerged that when it comes to images produced by researchers in the laboratory, the analysis tends to be quantitative. When the images are produced by researchers outside the laboratory, the type of analysis tends to be qualitative (Table 3.13).

Table 3.13: Collection technique vs analysis type

		qualitative	quantitative
collection technique	in lab	9	31
	out lab	22	11
Total		31	42

Programs

Also, the programs that are used during research are important in the processing/analysis phase. It is necessary to understand what kind of programs are used for tasks such as editing or analyzing images. In the context of this work, 45 programs were identified. Note that a document can have more than one associated program. Table 3.14 presents the analysis of the programs.

Table 3.14: Programs used per domain and significant comparisons

Programs	ArtH		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	% (n=29)	n	% (n=23)	n	% (n=21)	n	% (n=2)	n	% (n=34)	n	% (n=109)	
other	5	17.24%	15	65.22%	5	23.81%			8	23.53%	33	30.28%	LifeSB>ArtH*, PhyS*,Tech*
ImageJ			1	4.35%					3	8.82%	4	3.67%	
Matlab	1	3.45%	1	4.35%					2	5.88%	4	3.67%	
Mathworks	1	3.45%							1	2.94%	2	1.83%	
SPSS	1	3.45%	1	4.35%							2	1.83%	
Significant comparisons between programs	other>all*		other>all*		other>all*		other>Mathworks*,SPSS*		other>all*				

An * indicates significant results with $\alpha < 0.05$.

When no value is mentioned, this is equivalent to having 0%.

% are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

The 4 programs with the highest incidence are ImageJ (computer program for image processing, publicly accessible and developed in Java), Matlab (high-performance interactive software focused on numerical analysis, matrix calculation, signal processing, and graphing), Mathworks (mathematical computing software that supports data analysis and simulation) and SPSS (software with application analytics, data mining, text mining, and statistics). In total, 67 of the documents did not contain information on the programs. The ImageJ program is the one that appears most persistently in the analyzed documents. Although these are the most representative programs in the sample of articles under study, it is clear that there is a set of other programs used in the most diverse domains.

It was found that in 100% of the cases collected, the programs are associated with digital images. In Section 3.4.3.5, the results related to the description tasks that occurred in the images under study will be presented. However, it is interesting to note that 7 of the 19 occurrences of description that were verified, happen when there is the identification of an associated program. Further ArtH, there will be results inherent to the deposits/publications of the images under analysis, in Section 3.4.3.7. However, still in the context of programs, it is interesting to note that in 11 of the 13 documents where the task of depositing the images was explicit, there is also evidence of the use of a program.

This indicates that they make the description in a more formal way than the others that only identify annotation tasks. Of all the documents analyzed, 90 did not present information regarding the description of the images. Taking into account the potential of data management, this value is high, as it would be expected that more researchers would describe their data. However, as with other topics, the fact that it is not mentioned in the article does not necessarily imply that it did not happen.

3.4.3.5 Description

The description is a big challenge in data management. Metadata plays an essential role in this task since they allow the inclusion of important information for the interpretation of data by third parties. In addition, they allow data reuse and preservation. However, the data description process is often seen as demanding and time-consuming. For this work, it was important to understand

whether the researchers described their images and how this task was carried out. Description can be done using various techniques. In the documents analyzed, nine notes/annotations, one tag, four metadata/element, three labels, one classification and one keyword were identified.

In order to systematize all the information, these description techniques were divided into two categories: structural metadata (metadata, element, label, classification, keywords) and non-structural metadata (notes, annotations, tags). Table 3.15 shows the relation of the description of the images by research domain.

Table 3.15: Image description per domain and significant comparisons

Image description	ArtH		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	% (n=29)	n	% (n=23)	n	% (n=21)	n	% (n=2)	n	% (n=34)	n	% (n=109)	
non-structural metadata	2	6.90%	3	13.04%					3	8.82%	8	7.34%	ArtH>PhyS*
structural metadata			5	21.74%	3	14.29%			2	5.88%	10	9.17%	PhyS>ArtH*,Tech*

An * indicates significant results with $\alpha < 0.05$.

When no value is mentioned, this is equivalent to having 0%.

% are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

ArtH and Tech privilege descriptions at a non-structural level, using more informal methods, such as annotations or tags. However, in ArtH there is no use of structural methods, which leads one to wonder if the socio-humanistic domains do not usually use planned description approaches. LifeSB and Phys, on the other hand, use structural metadata more frequently, with more formal characteristics. The PhyS uses exclusively structural description, which may be related to the accuracy of description that is necessary for studies in this domain.

The distribution of results between the two categories is very balanced. However, non-structural metadata stands out lightly. It should be noted that only four documents explicitly present the terminology "metadata"/"elements", which indicates that they make the description in a more formal way than the others that only identify annotation tasks. The first case is in the LifeSB domain, where microscopic images were collected and metadata was associated with each set of images. As well as being present in the description, the authors of the article state that the metadata was also deposited and shared along with the datasets. In the second case, in the PhyS domain, a description of the data was made, establishing a set of elements to be filled in. One of the elements was filled in with an image showing the type of infrastructure ("photo used"). The remaining elements were textual ("green infrastructure" and "type scale level"). The last two cases are in the Tech domain. In the first case, a list of elements is presented at the end of the article, which systematizes the article and allows for a better interpretation. It is possible to see this list as a metadata model for describing the data, including images. In the second case, the dataset contains a set of photos uploaded by amateur photographers with searchable metadata and keywords.

Of all the documents analyzed, 90 did not present information regarding the description of the images. Taking into account the potential of data management, this value is high, as it would be expected that more researchers would describe their data. However, as with other topics, the fact that it is not mentioned in the article does not necessarily imply that it did not happen.

Still in the sphere of description, we tried to see if the type of images had any relevance in the description. Only 5 types of images appear associated with the description: photographs, microscopic images, medical images, computer-made images, and videos. Table 3.16 systematizes these results.

Table 3.16: Image type vs Image description

Image type	Total images with associated description
photography	10 (55.56%)
microscopic image	6 (33.33%)
medical image	2 (11.11%)
computer-made image	1 (5.56%)
video	1 (5.56%)

3.4.3.6 Storage

For this study, it was noted that none of the articles contained information on the storage phase. This phase is related to the way and method in which data is organized and stored. In this phase, it can be defined where and for how long the data will be stored, specify how and for what periods backups will be carried out, clarify how access to the data will be managed and how data security will be guaranteed (contingency plan if necessary), and establish the location and method of storing the data after the project has been completed.

3.4.3.7 Sharing

Researchers are increasingly encouraged (and sometimes required) to make their research data findability, accessible, interoperable, and reusable. Sharing is one of the essential phases of these practices. In this topic, deposit and licensing issues were analyzed.

Data deposit

Data deposit practices are increasingly recurrent. Researchers find here an opportunity to disseminate their work, share knowledge and see their work reused and cited. However, many researchers still show some resistance to depositing and sharing their data openly. This topic presents an analysis of the documents that present information about depositing or publishing your images. Table 3.17 presents this information. This table only presents information for documents that claimed to deposit their images. None of the remaining documents analyzed stated clearly that it did not deposit the images.

Were identified 16 documents in which the images were deposited. 92 documents did not present this information, so it is not possible to know if they were deposited or not. The LifeSB domain showed the most cases. The remaining domains show few results for deposit.

As for the verified deposit locations, they are diverse, but they fall mainly on open access archives, institutional open access repositories, general open access repositories, and on GitHub.

Table 3.17: Image deposit per domain and significant comparisons

Image deposit	ArtH		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	% (n=29)	n	% (n=23)	n	% (n=21)	n	% (n=2)	n	% (n=34)	n	% (n=109)	
yes	1	3.45%	12	52.17%	1	4.76%			2	5.88%	16	14.68%	LifeSB>ArtH*, PhyS*, Tech*

An * indicates significant results with $\alpha < 0.05$.
 When no value is mentioned, this is equivalent to having 0%.
 % are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

This fact leads one to believe that there may be other works whose data are deposited, however, it is not possible to say whether they do it or not. None of the analyzed documents showed the access policy of their deposits, therefore, it is not possible to identify whether the deposited images were shared with open access or if they are kept private. The access policy, as a rule, may depend on the phase in which the project is taking place, on the need to share it for publication in scientific journals, or, for example, for the guarantee of authorship.

Licenses

Licenses are essential tools in the scientific production and communication process. They secure copyrights and identify the access policy. In this work, all licenses mentioned by the authors of the documents were identified. Nine licenses have been identified, four belonging to the Creative Commons Licenses group. Nature, APA, Elsevier, AAAS, and ACS licenses were also identified. Note that only 55 of the 109 articles under analysis included information regarding the license. A brief explanation of these licenses follows, in Table 3.18.

Table 3.18: Licenses description

License	Description	Link
CCBY	It allows others to distribute, adapt, and develop others' work, even commercially, as long as they credit the original author. This is the most flexible license.	shorturl.at/mtvBO
CCBY-NC	This license allows others to adapt and develop works by other authors on a non-commercial basis. It does not require the licensing of derivative works on the same terms as the originals.	shorturl.at/dkloG
CCBY-NC-ND	It is the most restrictive license as it only allows others to download works and share them, as long as credit is given to the original authors. It is not allowed to change or use the works for commercial purposes.	shorturl.at/dirtB
CCBY-NC-SA	This license allows others to adapt and develop the work of the original authors on a non-commercial basis, as long as they credit the authors and license the new creations under the same terms.	shorturl.at/rsyIQ
Nature	License applied to all journals published by Nature Portfolio, including Nature and journals with "Nature" in their titles. These policies are compatible with the vast majority of funders' open access and self-archiving policies.	shorturl.at/msxJL
APA	License that addresses licensing and certification issues in psychology.	shorturl.at/goxOQ
Elsevier	Works under an Elsevier license are copyrighted. Users may access, download, copy, translate, and extract texts and data, for non-commercial purposes, under a set of conditions. However, it is not possible to redistribute, display or adapt the works.	shorturl.at/pFKOT
AAAS	This license allows AAAS exclusive rights to publish the work. Authors can choose between two Creative Commons licenses (CC BY-NC and CC BY).	shorturl.at/wFLR9
ACS	Intended for core feature sets and add-on licenses based on incremental features. Allows the author to choose 4 different typologies. Requires a base license file to be applied to each ACS server. Has Large Deployment and CTS feature licenses.	shorturl.at/JPQ28

Table 3.19 indicates the use of licenses by the research domain. From a general point of view, licenses from the Creative Commons group are the most used, especially CCBY, with the exception of the PhyS domain where CCBY-NC-ND is the most used.

The Tech domain has a greater diversity of licenses. Given the broad spectrum of types of work in this domain, it was expected. In ArtH, 11 works have an associated license, while in LifeSB, 18 works have an assigned license, proving to be the best-positioned domain in terms of

licenses. PhyS, on the other hand, only presents 8 documents with information on the licenses. In SocS, the two documents that make up the domain present information on the licenses.

Table 3.19: Licenses per domain and significant comparisons

Licenses	ArtH		LifeSB		PhyS		SocS		Tech		Total		Significant comparisons between domain
	n	%(n=29)	n	%(n=23)	n	%(n=21)	n	%(n=2)	n	%(n=34)	n	%(n=109)	
CCBY	7	24.14%	17	73.91%	2	9.52%	1	50%	5	14.71%	32	29.36%	LifeSB>all*
CCBY-NC-ND	3	10.34%			5	23.81%	1	50%	2	5.88%	11	10.09%	PhyS>LifeSB*,Tech*
Nature									4	11.76%	4	3.67%	Tech>LifeSB*
CCBY-NC-SA					1	4.76%			2	5.88%	3	2.75%	
CCBY-NC									1	2.94%	1	0.92%	
APA	1	3.45%									1	0.92%	
Elsevier			1	4.35%							1	0.92%	
AAAS									1	2.94%	1	0.92%	
ACS									1	2.94%	1	0.92%	
Significant comparisons between licenses			CCBY>all*		CCBY-NC-ND>all*								

An * indicates significant results with $\alpha < 0.05$.

When no value is mentioned, this is equivalent to having 0%.

% are calculated in relation to the number of articles in the domain

<10% 10%-29% 30%-49% >49%

Whether to produce a resource or reuse it from third parties, it is important to indicate the intellectual property relating to that resource. Scientific transparency allows for resource sharing and, consequently, a healthy open access policy. For this reason, Table 3.20 presents the relationship between the use and production of the images and the respective licenses.

Table 3.20: Image acquisition vs licenses

	License									Total
	CC BY	CC BY NC-ND	CC BY NC-SA	CC BY NC	APA	Elsevier	AAAS	Nature	ACS	
Use	3 (60%)	2 (40%)	0	0	0	0	0	0	0	5 (9.1%)
Production	29 (58%)	9 (18%)	3 (6%)	1 (2%)	1 (2%)	1 (2%)	1 (2%)	4 (8%)	1 (2%)	50 (90.9%)
Total	32 (58.18%)	11 (20%)	3 (5.45%)	1 (1.82%)	1 (1.82%)	1 (1.82%)	1 (1.82%)	4 (7.27%)	1 (1.82%)	55

Of the 19 documents in which the images are used, 5 (26.32%) mention the license. In all 5 cases, it was found that the intellectual property belongs to someone else and therefore there may be no knowledge of the license. However, when images are collected from platforms such as repositories, in most cases the license has to be associated with the resource. However, it is also common for researchers to obtain resources through direct contacts, such as e-mails, and therefore this information may not be transmitted.

Of the 91 documents in which the images are produced, 50 (54.95%) mention the license. When the images are of their own authorship, researchers tend to indicate the license, as it works as security in relation to resources. Still, in the analyzed documents, just over half indicated the license information. This does not exclude that the license exists, it may just not have been included in the final version of the document.

3.4.4 Discussion of results

After a detailed analysis of the results obtained from this study, it is important to discuss those that were the essential topics that motivated the accomplishment of this work. For this, it is necessary to understand whether answers were obtained for the research questions presented in the Section 1.

Regarding SQ1, we can say that in the published works it was not possible to identify precisely and without any doubt the existence of formal practices and standards in the management of images as research data. In several cases, we were able to identify patterns by research domain or practices that are more associated with a certain type of image or research profile. With the analysis of all the works, we were able to conclude that there are several behaviors associated with image management, however, these behaviors do not result from a formal protocol, pre-established and designed to standardize the method of image data management. It became clear that in research routines, researchers tend to find methods, techniques, and tools that they consider more favorable for the context of their work, but they do not formally document these practices in the work they publish.

With regard to SQ2, it was quite clear that the research domain, in several cases, determines how the images are managed. Most of the time, the different research domains manage images differently. First, the image types prevalent in each domain are different from each other. This difference means that images are often managed in different ways. If in other types of data, the differences are not always so significant, in data in image format, their typology determines very different characteristics. A photograph has characteristics that distinguish it from a map. Just as a drawing has different characteristics from a microscopic image. These particularities reveal a need for its own management, whether due to its format, form of analysis, organizational and storage capacity, or even publication. Furthermore, the profile of researchers in each domain is different. Whether due to working conditions (in the laboratory or outside it, for example), previous knowledge, academic education, and previous interaction with the dynamics of RDM.

Regarding SQ3, it is possible to notice that a good part of the works indicates that the researchers are aware of some data management practices and policies, although they do not explicitly say so in their works. This conclusion can be obtained through the analysis performed. We believe that the management practices they plan and execute are not documented in the articles they publish, other than mandatory topics, such as licenses. In many cases, a thorough reading of the work and an analysis of the procedures were necessary to find and interpret these practices and policies. They exist in many cases but are rarely documented in the articles. This is also a particularity in this analysis because in most cases where there is no information about a particular topic, it cannot say whether that practice exists or not, there is just no information about it.

As for SQ4, it cannot say with certainty that habits and trends related to image tend to change over the years. This study covered a time span of eleven years (2010-2021), so during this period, there were no major variations or certain specific characteristics that it can identify and say are related to the passage of years.

This work arose from the need to understand the habits and behaviors of researchers who

use and produce images. When we began this study, we found that there were no studies in the literature that would allow us to clearly analyze this issue. The theoretical material found never focused on this subject as the main theme of the studies and this gap limited future work. For this reason, over the last few years, we have developed some works that explore these dynamics, trying to understand in detail the issues of images in science.

With this work we intend to close this research cycle, combining the results of the literature with the results obtained in contact with researchers. These four works made it possible to perceive the consistency of some practices and habits, determine limitations and challenges and give robustness to the results achieved.

3.4.5 Final remarks

This study allowed us to reflect on how research projects and researchers prepare and plan the use of images. In fact, it became clear that the use and production of images can happen many times, however, this frequency decreases when images acquire the main role of research.

Clearly, there are still tasks associated with research data management that are left out by researchers. This happens with the most common data typologies, such as textual and numeric data, but it is even more accentuated when the data is of an imagery nature. There is no clear reason for this. In fact, the exponential growth of image production, associated with technological advances that allow the capture and dissemination of images, has led to a greater awareness of this type of data. Perhaps because the images are still seen as complements to the research, they make them more susceptible to lack of management, unless they are imperatively the main topic of the study.

Practices related to the definition of the typology of images and capture instruments, as well as the establishment of the use and production of images, their type of analysis, and methodology, are in several cases reported in the articles. However, details related to programs used, descriptions of images, and deposits are often omitted. One of the problems of this careful analysis, thorough analysis dimensions, involves realizing whether that information is really non-existent or just not reported in the article.

The idea is that two fundamental pillars are missing: 1) alerting to the diversity of data typologies, associated with the explanation of their characteristics, potentials, and challenges; 2) encouraging, sensitizing, and teaching about research data management and its impact on research projects.

Future work involves developing several tools that facilitate the work of researchers in the task of data management, namely when these data are images. We are in the process of publishing a metadata model and controlled vocabularies for the description of images. We believe that this will significantly contribute to a better understanding of the image description process. We are also working on an experiment that aims to establish a scenario of how the process of interpreting images in repositories occurs, with the objective of determining gaps and proposing improvements in the deposit and publication process. Previous work, the present work, and what is defined as future work will be fundamental for another objective: to prepare a guiding document with

guidelines for the management of data in image format that systematizes a series of good practices, suggests research and management policies and clearly explain its advantages, in order to raise awareness of this topic.

Although it follows a systematic process, this work has specific characteristics that make it deviate from a pure systematic review. Since we wanted to consider articles from different scientific domains, we did not use a search query and selected works by their impact (assessed through their number of citations). We limited the selection to the top 10 most cited works per year to allow a manual analysis of the results. Given this, future replications of this methodology may consider a different set of papers. However, besides changes caused by the natural evolution of the current practices, we don't expect this change in the included articles will impact our results. We believe our 109 articles statistically represent the actual state of image management practices.

Chapter 4

Image as data

Increasingly, images are produced and used as research data. To analyze the dynamics surrounding images as data, as well as to develop new hypotheses, three studies were conducted. The first led to the development of a metadata model for describing data in image format since there was a need to aid the data description process. In addition, this study enabled the development of a set of controlled vocabularies for some of the model's elements. The second study addresses the issues of automatic image description and the use of artificial intelligence tools as a complement to the description. The third study presents a comparative analysis of various technological solutions for data sharing and storage, as it is necessary to reinforce the importance of these solutions in the research data management process. At the end of this chapter, there is a section presenting guidelines for the use and production of images in a research context, highlighting some practices and habits that can guide the good use of images, taking into account the principles of research data management.

This chapter is aligned with RQ2, aiming to give depth to the question, trying to raise hypotheses and record conclusions that lead to its answer.

4.1 Image data description: proposal of a metadata model and controlled vocabularies

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, elements that are unique to a certain domain can be included to satisfy the needs of description (example: *chemical compound* element for the *chemistry domain*). In addition to the model, controlled vocabularies have also been proposed for the elements in the metadata model.

4.1.1 Contextualization

The exponential production of research data increasingly drives science. And a more driven science further enhances the production of research data. Research Data Management is an essential requirement for research projects.

Involving researchers in data description is problematic, as the lack of knowledge and the absence of RDM practices in research groups are common [23]. 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.

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 elements 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 [25].

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 [5]. 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 element lets the researcher realize the possible options for a given field and avoid confusing, ambiguous, or even wrong descriptions. Allied with this, it streamlines the description, as the researcher can perform this task more quickly and effectively.

4.1.2 Methodology

The methodological approach was organized into five steps that will be described in the following subsections. Figure 4.1 systematizes these phases.

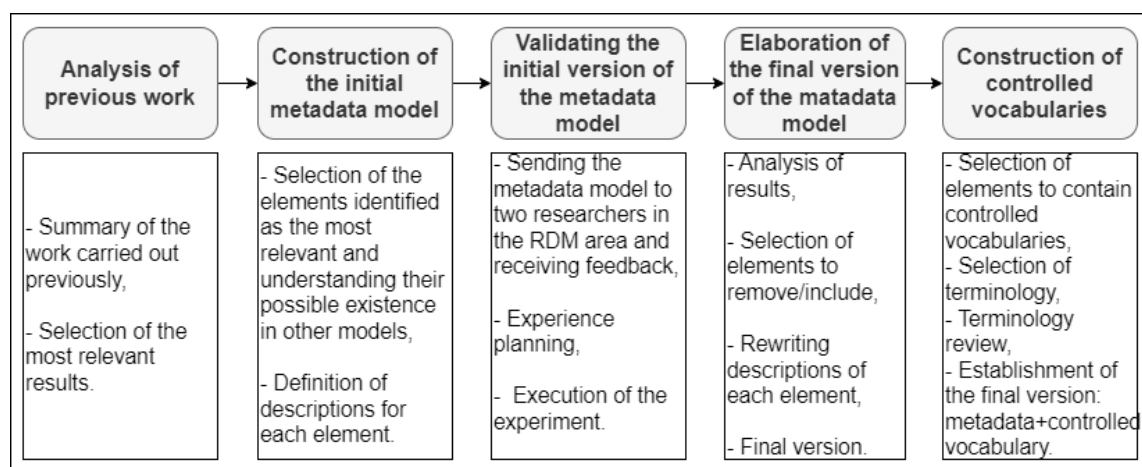


Figure 4.1: Stages of the elaboration of the metadata model and controlled vocabularies

4.1.2.1 Analysis of previous work

After an exhaustive search, no specific metadata model was found 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 the previous studies that explored the behaviors and practices of researchers regarding the management of research data in image format. These studies were essential for elaborating the initial metadata model that is presented and evaluated in this study and allowed raising the requirements for the elaboration of the initial metadata model.

4.1.2.2 Construction of the initial metadata model

Based on previous work, a set of requirements was established to develop the metadata model. These requirements included: elements to be included; existence of elements in other models; adaptability of the model to different research domains; adaptability of elements to repositories and directories and model flexibility.

4.1.2.3 Validating the initial version of the metadata model

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

For the description sessions, a total of 11 participants were recruited, 2 (18.18%) from ESE, 3 (27.27%) from ESL, 2 (18.18%) from ESN, and 4 (36.36%) from CSH. Since the HSC domain had the lowest incidence in the interviews, to have a greater number of participants from this domain 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 them asking for their cooperation and explaining the study in question. If they replied in the affirmative, a second email was sent with the date and time of the experiment and they were asked to send the two images to be included in the experiment. After this, a personalized Google Forms form was drawn up with the images of each of the participants (the metadata template was the same for all the forms, but the images varied). The day before the session, a new email was sent to remind them of the experiment, with the access link and 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 experiments allowed the development of the metadata model.

COVID-19 did not allow this experiment to take place in person, so a remote experiment was designed. A Google Forms document with two pages was used (one for each image to be described). At the top of the page, the objective of the task was described. Next, the elements were listed and, for each one, a caption was included to clarify its purpose. Mandatory elements were flagged. Figure 4.2 presents a Screenshot of the form used for the experiment.

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 este(a) a reutilização: este descritor não deve ser utilizado
A sua resposta

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

Tipo *
Natureza ou gênero da imagem. Exemplo: fotografias, imagens médicas, imagens microscópicas, desenhos, ilustrações, retratos, vídeos, gravuras, imagens feitas em computador, mapas, gráficos, animações, pinturas, plantas 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 "autoria própria"
A sua resposta

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

Recursos relacionados
Link ou apontamento para um recurso estritamente 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, composição, recorte, edição de brilho, requadramento. Se a imagem for produzida por terceiros não preencher este descritor: a menos que saiba concretamente 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 enfoque 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
Especifica os valores numéricos ou textuais definidos na escala de medida
A sua resposta

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

Seguinte Limpar formulário

Figure 4.2: 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.

All the sessions were recorded. There was no interaction between the participant and the person conducting the session, so there was no influence on the answers. The person in charge of the session took notes on the hesitations, the spaces left blank, and the time spent on some elements. She used these notes and seven other questions to guide a discussion at the end of the session. The questions will be detailed in Section 4.1.4.

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 (after the defense and publication of this thesis). Details about the elaboration of vocabularies can be read in Section [4.1.7](#).

4.1.2.4 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 describing data in image format was drawn up. For each decision, the RDM guidelines were always taken into account. Details about the realization of vocabularies can be read in Section [4.1.6](#).

4.1.2.5 Construction of controlled vocabularies

In the final model, to make the model more robust, reliable and capable of aiding description, controlled vocabularies were developed for some of the elements in the metadata model. Firstly, the elements that could have a controlled vocabulary were identified. Next, an exhaustive search was made of the terminology related to each of these elements. In the end, a dataset was built containing the metadata model and the controlled vocabularies. The aim is to make all the information relating to this metadata model available to the scientific community. Details of how the vocabularies were created can be read in Section [4.1.7](#).

4.1.3 Initial metadata model

The initial metadata model was drawn up on the basis of the results obtained in previous studies (covered in chapter [3](#)), always taking into account the opinion of researchers from different domains. The initial model includes the elements most frequently mentioned by researchers. In addition, it was necessary to understand which elements are most used in the general research landscape and which are usually included in repositories, directories and the like.

The main objective is to propose a generic metadata model that can be used in all research domains, with a special focus on images. However, these elements must be able to include all relevant information. To obtain a more flexible model, a set of mandatory elements (important for interpretation) and optional elements (not fundamental for interpretation, but which give depth to the data) were defined. The model also has to be flexible to the point of allowing the inclusion of new elements, that is, having a domain of "others", where it is possible to include relevant elements, in addition to those already stipulated. Finally, the model should not be too long. It 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 elements, participants tend to describe it more often. In Table [4.1](#), are included the elements of the initial metadata model.

Table 4.1: Image description metadata model elements - before description sessions

Element	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 element should not be used.
DC:Temporal coverage ²	Point or time period to which the image content relates.
Type ¹	Nature 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 element, 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 element.
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]Mandatory, [2]Optional

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

4.1.4 Evaluation of the metadata model

The description sessions were analyzed according to the notes taken during the experiment and the questions put to the participants at the end of the session. Whenever it was necessary to classify a detail, the recordings made during the sessions were used.

4.1.4.1 Observation of the description

There were some differences between the participants in the description times. Table 4.2 shows how long each participant took to describe each image and the number of elements used per image. Was identifies the participants with an "R" followed by the acronym of their research domain and a progressively assigned number. 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 elements 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 elements in the first image than in the second, 9.1% (1) of the participants used more elements in the second image than in the first, and 36.4% (4) used the same number of elements in both images.

It is important to mention that the experiment took place with two images per participant. A large set of images was not used to simplify the experiment 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 4.2: Time each participant took to describe images and number of elements 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

4.1.4.2 Empty elements

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

4.1.4.3 Hesitations

To identify less clear elements, the hesitations that occurred during the sessions were also analyzed. Most of the participants (9 out of 11) struggled to distinguish the *date of production* and *date of acquisition*. Participants often "jumped" between elements to understand the differences between the time elements. Two participants (one from SSH and one from LHS) hesitated to fill in the methodology element. One of them didn't 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 elements *quality*, *edition* and *sample* were also associated with hesitations and were not always filled.

4.1.4.4 Results summary

To obtain detailed information on each element, we analyzed the time participants spent on each one, the length of the description, the number of times they were used, the average time participants spent on each element and the average time between elements. Table 4.3 presents 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 element. The *time lapse* refers to the average time

between the end of filling out one element and the beginning of filling out the next. The *char* is the number of characters used in each element. The *used* refers to the percentage of times each element was used.

Table 4.3: Elements usage during the sessions: average start and end times, average time spend on each element, average time between element the, average number of used characters and percentage of times used

elements	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 elements 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 element *Related resources* has a longer filling time. During one session, it was noticeable that participants used documents they had saved on their computers to check these resources and copy their information into the element. This made it take longer to fill in in most cases. *Methodology* was seen as one of the elements 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 elements. Given its possible length, *Description* could be a element that takes longer to fill in. However, it turned out that this is one of the elements that participants find easiest to fill in, so it's quicker. *Time coverage*, *Date of production* and *Format* stand out for the shortest description time. These are elements that, as a rule, have a more direct response, without the need for much detail. In addition, they are elements that don't use many characters.

It was found that the participants usually described in sequential order. There was no complete reading of the model before filling it in. On average, it took the participants 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 element and the beginning of filling out the next one, the longer period of time is between the *Date acquisition* - *Production date*; *Related*

resources - *Methodology*; *Methodology* - *Edition Statement*; *Instrument* - *Spatial coverage* and *Scale* - *Sample*. The longer time between elements can be justified because they were the ones that caused the most hesitation among the participants, as they took some time to interpret and realize what content would be included in that field.

Description, *Related resources*, *Methodology* and *Quality* are the elements 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 elements with more characters, it is placed within seventeen seconds of completion, which suggests that although it is a long-filling element, the participants have information regarding the completion.

It was found that some elements were used more frequently than others. The *Sample* and *Scale* are the elements that were least used, with 18.18% and 27.27%, respectively. Also, other elements were not always filled. *Related resources* was filled in 90.91% of the time. *Production date* and *Source* were filled 81.82% of the time. *Quality* and *Temporal coverage* were filled in 72.72% of the time. *Edition statement* was filled in 54.55% of the time. The reasons given by participants include: the purpose of the element was unclear, the element was not relevant and the information to fill in the description was not available/recorded.

4.1.5 Feedback from the participants

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

4.1.5.1 What difficulties were there?

Some topics were highlighted by participants as the greatest difficulties, namely: 1) distinction of the data elements, 2) not knowing many quality improvement techniques for including in *Quality* element, 3) understanding 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 the inclusion of the *Related Resources* element, since, although it is very pertinent, it is not a quick element to fill out and that often implies resorting to external search.

4.1.5.2 Did you miss any elements?

Most participants pointed out elements that could be included in the metadata model. One participant from SSH said that he was not familiar with data description, although he considered it

important. For this reason, he said he was unable to identify new elements, but said that this issue could be discussed in group meetings. An NES participant said that, at the generic level, no element 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 elements. The most requested elements 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/funders*: considered bureaucratic but essential, because many projects are required to show their funders/partners (2 times).

4.1.5.3 What elements did you not use and why?

The participant that used all the elements said the metadata model seemed relatively simple to fill out and the elements useful. The remaining participants mentioned mostly the *Sample*, the *Scale*, the *Edition Statement*, and the *Quality* as less useful elements. 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 elements empty, everyone thinks they are relevant, they simply did not apply the images in question, but they may be important to others.

4.1.5.4 Do you agree with the obligation defined for elements?

Seven of the participants agree with the necessity to define mandatory fields. 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 elements should be rethought. Another said he would not make the acquisition date mandatory and would only put the elements "data" (which would include the images produced and acquired). Another participant questioned whether any element 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 never knows what 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 ideal, but never with too many mandatory elements.

4.1.5.5 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 elements 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.

4.1.5.6 Would some elements 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 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 elements would be more easily filled in by other participants. One of the participants 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.

4.1.5.7 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 says 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.

4.1.6 Final metadata model

After analyzing the results of the description sessions, the final version of the metadata model was drawn up. This new version has 26 elements, 9 mandatory, and 17 optional.

In this new version, the time coverage element has been removed. It was found that the date elements were sufficient for the description, and it was not necessary to include more than the production date and the acquisition date. In the description of the production date and acquisition date, the format in which the date should be written was included. In the previous version of the

metadata model, the acquisition date element was mandatory, however, it was realized that this information does not always exist, so this element is now optional.

Four new mandatory elements 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 elements, 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, experiments with researchers and projects, and awareness of international guidelines for data management, made us believe that these two elements are essential. Without a contact element, 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 elements were also included: research group, project, contacts, partnerships, funders, 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 considered. As well as being mentioned by some participants, it was clear from research on the subject that this is an important feature. In other words, if the participants feel it necessary, they can include more elements specific to their domains in this metadata model. The material element was added on the initiative of the authors of this work because they consider that this element can be very relevant, especially when it comes to analog images, as the material used to make the image can be important information to mention. 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, the descriptions associated with the elements were also improved to clarify their purpose. The contextualization was detailed, more examples were given and it was indicated if more than one value was accepted (e.g. *several Edition statement can be indicated*). In the date elements, the expected format for the date was specified. Table 4.4 shows the final metadata model. A dataset was thus generated, which was described and deposited in the INESC TEC Data Repository, in open access for understanding the data in detail [112].

This metadata model is not intended specifically for the description of an image but for datasets of images. The 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 could be described in sets, reducing the time spent and grouping the images into 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.

Table 4.4: Image description metadata model elements - after description sessions

Element	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 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 element 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 element, 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 element. Several quality procedures can 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.	
Funders ²	Entity or institution that financed resources for obtaining or producing the set of images.	

[1]Mandatory, [2]Optional

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

4.1.7 Construction of controlled vocabularies

The desire to make the model more interoperable, easier to understand and easier to use has led to the development of controlled vocabularies associated with some of the elements in the metadata model. It is hoped that these vocabularies will facilitate the process of describing images.

The first step in developing controlled vocabularies was to select the elements that should have an associated vocabulary, i.e. those that have a limited set of possible values. For example, on the one hand, the *Author* element has an infinite set of possibilities, and it is not possible to pre-identify all possible names. On the other hand, the *Type* element 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 sources of information that would clarify in detail the terminology associated with each element. Some elements were immediately excluded, as the countless possibilities of terms did not enable the construction of vocabularies.

Ten elements were selected to build controlled vocabularies: *Research domain*, *Type*, *Format*, *Instrument*, *Material*, *Use rights*, *Methodology*, *Edition Statement*, *Quality*, and *Scale*.

The next step was to define the specific terminology that would make up each of the controlled vocabularies for describing the images. This phase is fundamental, as a poor choice of terminology can affect the credibility of the model, so choosing good sources of information (e.g. studies, scientific articles, dissertations and theses) is essential. As explained below, an individual study was carried out for each element.

For the *Research Domain*, followed the guidelines of the Foundation for Science and Technology (FCT)¹, 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 *Type*, research was carried out into all types of analog and digital images that could be captured or developed, whether for recreational or professional purposes. 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 element, the elaboration of vocabularies made it clear that formats and supports can be included in this element because when it comes to analog images speak of support, that is, support is what sustains something, the basis of physical support. This vocabulary contains one hundred and twelve terms.

For *Instrument*, for each type of image, the instruments that could be used to capture or develop it were identified. This vocabulary contains fifteen terms.

The element *Material* was challenging because, before the vocabularies began to be developed, material and instrument were confused. In fact, material and instrument are different things and, in the field of image, they have different roles. So, for this element, all the materials where it was possible to see or print the images were identified. As this is physical material, this element and its controlled vocabulary only apply to analog images. This vocabulary contains twenty-one terms.

For *Rights of use*, a study of existing licenses was carried out to support research. The web resources of the various accreditation and licensing institutions available for scientific production, such as Creative Commons and Elsevier, were analyzed. This vocabulary contains thirty-five terms.

As participants had difficulty filling in the element *Methodology*, six main methodological scenarios were defined. In the event that the information in the vocabulary is not sufficient, the participant still has the option of giving more details about the methodology in the description or selecting the "other" option and specifying other methodological approaches. This vocabulary contains six terms.

The various image editing and analysis programs and methods were studied to create controlled vocabularies for the elements *Edition Statement* and *Quality*. This vocabulary contains forty one terms.

¹<https://www.fct.pt/>

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

Due to the large number of terms in the controlled vocabularies, it is difficult to present them in this document. Therefore, a set of data was generated, described and deposited in INESC TEC's Data Repository, in open access for understanding the data in detail [112].

4.1.8 Final remarks

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 for participants in the data management process. The benefits are based on the possibility of using a valid, robust tool that can respond to their needs, such as metadata models, especially when they have elements that can guarantee an effective description of the data. In addition, templates reduce the time spent, as they present participants with a set of terms that are relevant to describing their data, without them having to look for the elements in external sources. It has therefore become clear that having a metadata model means less work and this is a valid argument that the participants are looking for, since the speed of the day and the deadlines to be met mean that they have to select the most important tasks to carry out.

Describing data in image format is still strange. Many participants associate 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 elements. The inclusion of subtitles in the elements was considered essential for a faster description, making it easier to understand what needs to be included in the field. Furthermore, the support of examples in the caption is seen as very useful and even essential in certain elements, 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 elements that cover all the characteristics of the image. However, it allows for specificity through the possibility of including new elements, when necessary. Furthermore, through 26 elements, are given several possibilities for describing images, precisely with the purpose of contextualizing this data typology.

It also became clear that the model could not have only mandatory elements, since 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, and the purpose of use or production of the image are some of the many characteristics that can dictate which elements to use in the description moment. However, it is very important to include optional elements, to recall and identify description possibilities that otherwise might be forgotten. The choice to determine the mandatory elements 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.

4.2 Automatic description of research images: utopia or reality?

This section deals with the topic of automatic image description. Based on the assumption that technological advances are leading us to increasingly sophisticated tools and instruments, it is important to reflect on the position of human description.

This question is explored here by comparing the labels assigned to a set of images by: 1) the researchers who produced the images; 2) researchers not associated with the context in which the images were produced; 3) an automatic image processing tool (Vision AI).

These three perspectives were intended to study two essential phenomena in the process of image description: formal analysis and semantic expression analysis.

4.2.1 Contextualization

The automatic description of images is a significant advance for the visualization and interpretation of images. It allows us to go beyond the human vision which, although very accurate, is fallible. Image processing tools enable us to observe in greater detail and can also be decisive in diagnoses, assessments, and decision-making [55, 54].

We believe that we are closer to achieving the expected results when it comes to analyzing the visible content of an image, but when it comes to analyzing the semantics behind that image, namely its messages and meanings, the road ahead seems longer.

It is necessary to ensure that all layers of the image are interpreted. Can this guarantee only be achieved with automatic processing tools? Or is human curation necessary to validate and complement the description? This is the purpose of this study, which aims to explore these questions and raise hypotheses.

4.2.2 Description of images

The description of images implies a set of tasks aimed at obtaining results that allow a faithful representation of the images. Therefore, image description must involve the reliable creation of access points to information and the use of controlled languages, based on authority control.

The use of controlled languages is useful because of a very justifiable benefit since an information system that does not have control of access points does not offer the user the guarantee that he retrieves all the information that actually exists and is relevant to him, even though, at first glance, the search seems to be effective.

When we talk about controlled language, we are referring to the sensible use of vocabulary expressions that correctly represent the content without creating double interpretations. It is important to note that there is no specific vocabulary for the representation of concepts and, although it must be effectively controlled, it cannot deviate from the original content of the document under analysis [101, 58, 104].

The representation of concepts can be performed through indexing terms, access points, descriptors, elements, metadata, or classifications, for example. However, regardless of their typology, these vocabulary expressions must be able to represent concepts, the selected terms must not correspond to more than one concept and their presentation can be made as a simple term or a compound term (formed by two parts: nucleus and modifier) [81].

4.2.2.1 Formal analysis

Formal image analysis is, above all, founded on two major pillars that allow it to be carried out properly. The first is the identification of the elements that are general properties of the image, namely the environment and the composition of the image, the spaces and scenarios (whether natural, urban, or peri-urban), the human relations in that space, the static aspects (such as gesture, body expressions, and clothing), and the objects and living beings. The second is about choosing the concepts that best represent the content of the image. These selected labels or expressions should be unambiguous and well-known terms. This selection is challenging because it is necessary to guarantee the interpretation of the image (does not select a few concepts) and that the user does not have an overload of information regarding the image that may be redundant, duplicated, or not necessary (does not select too many concepts) [96].

4.2.2.2 Semantic expression analysis

The reading of images has a polysemic character, that is, the multiplicity of meanings existing in this source of information can be varied. As in the written document, the images require analysis, establishing comparisons, similarities, coincidences, and repetitions [126]. In the analysis of images' semantic expression, it is necessary to determine the aspects that are not visible in a literal and raw reading of the image, such as ideological messages, advertising messages, satire, and others [80].

A large portion of society manages to make only a superficial reading of an image, understanding its objective and superficial characteristics, leaving aside what is expressed in a semantic way [126]. For this reason, a greater understanding of the image and its importance is necessary [80].

The idea that non-specialists have the right to read images like someone who reads text is increasingly present. Manguel [78] even says that images are like stories waiting for a narrator. And the spectator must discover the explicit or implicit stories. However, images are not always easy to read, so much so that, in order to try to read them, it is necessary to know how to see and interpret them. Therefore, one cannot work with the image as if they were totally transparent, but understand them as a language, produced within a socio-historical and cultural context.

4.2.3 Methodology

The sample of images used for the study consists of 27 images, from five researchers from different projects and research domains. Of the total, 22 are in *.png* format and 5 in *.jpg* format.

The interaction with the researchers took place through an online form (Google Forms), which included the images that the researchers previously sent and which they should describe down according to the content of each one.

We created a dataset that compiled all the images and their descriptions. It was understood that the analysis of the results would work if the first descriptions (which were classified as the launch of the study) were made with descriptions made by automatic processing tools and by researchers who were not familiar with the images. For the first, we use Vision AI², a tool that integrates computer vision models and Machine Learning in order to automatically obtain image labels. For the second, we selected researchers who had no previous knowledge about the images, nor who belonged to the same research domain. When constructing the dataset, all labels were normalized (capitalization of the first letter and reduction of all labels to the singular) to make the analysis more complete. The dataset is available at INESC TEC's data repository [117]. Figure 4.3 shows an example of an image and its description.

²<https://cloud.google.com/vision?hl>

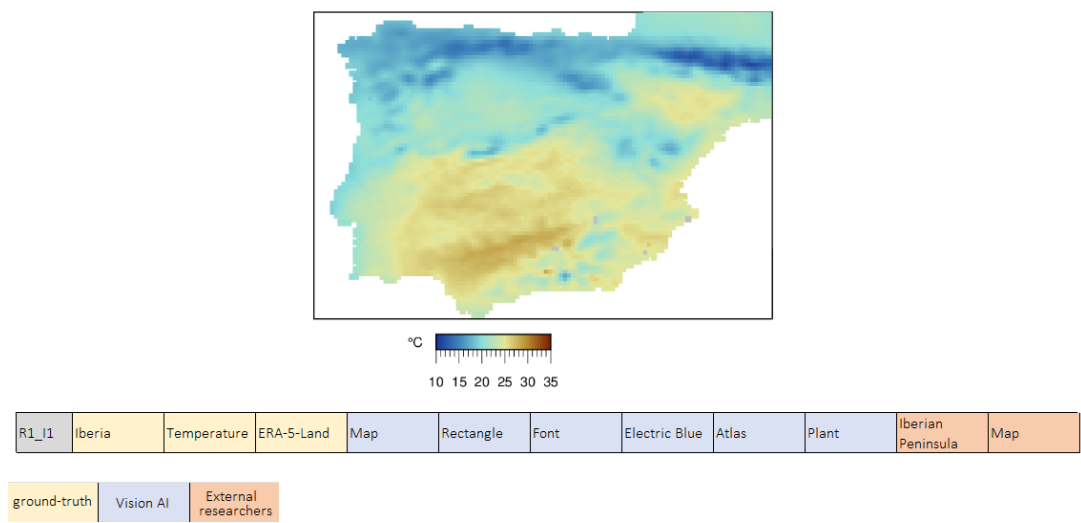


Figure 4.3: Example of an image description

4.2.4 Results

Based on the ground-truth, it was noticed that the description obtained through Vison AI, in most cases, did not correspond to what had previously been identified. As can be seen in Figure 4.4, only image R5_I4 had a term that completely coincided between the researcher and the tool. This term was "map". All the others did not converge. Although Vison AI collected more labels than the original description most of the time, this did not guarantee a closer match.

In view of the results, it was decided to see if there were any labels which, although not completely the same, had some similarities, known as partial matches. This was verified in three images, namely R4_I5, R6_I1, and R6_I2. In the first case, the researcher annotated the "line chart" and the Vison AI "line". In the second case, the researcher annotated the "mouse cage" and the Vison AI "cage". In the third case, the researcher annotated "animal welfare" and the Vison AI "animal".

Descriptions by researchers from outside the projects where the images were produced present different results. Regarding the labels that are fully compatible with the original ones (named by full matches), eleven labels were found to be equivalent, and these labels pertained to nine images (R2_I1, R2_I4, R2_I5, R3_I4, R4_I5, R4_I6, R5_I3, R5_I5 and R6_I2). For R2_I1, both researchers identified "ArchoOnto" and "CIDOC-CRM". For R2_I4 and R2_I5 "Linked Data". For R3_I4 "comment" and "anotation". For R4_I5 "Publications per year". For R5_I1 and R5_I5 "digital twin". Para R5_I3 "blockchain". For R6_I2 "hen".

As for the partial matches between the ground-truth and the external researchers, 16 of the 27 images under analysis have a coinciding label. This means that in more than half of the images under analysis (59.26%), the external researchers were able to find a label that matches the original description. Among the many examples are "B2Note tool use" and "B2Note"; "Metadata set" and "Metadata"; "DMP creation" and "DMP", "class" and "class diagram"; "map" and "european map", "Iberian" and "Iberian Peninsula", among others that can be seen in the published dataset.

It was also found that in images R1_I1 and R2_I3, there was a coinciding label in each of them when comparing the descriptions of AI Vison and the external researchers. In the first case, the term used was "map" and in the second "diagram" (Figure 4.4).

image ID	Ground-truth		Vision AI		
	#labels	#labels	number of full matches labels	number of partial matches labels	average number of matching (full and partial) labels
R1_I1	3	6	0	0	0,0%
R2_I1	10	9	0	0	0,0%
R2_I2	13	11	0	0	0,0%
R2_I3	10	10	0	0	0,0%
R2_I4	11	13	0	0	0,0%
R2_I5	12	20	0	0	0,0%
R3_I1	4	27	0	0	0,0%
R3_I2	5	12	0	0	0,0%
R3_I3	5	13	0	0	0,0%
R3_I4	6	18	0	0	0,0%
R3_I5	6	15	0	0	0,0%
R3_I6	5	32	0	0	0,0%
R4_I1	3	8	0	0	0,0%
R4_I2	3	9	0	0	0,0%
R4_I3	5	15	0	0	0,0%
R4_I4	3	8	0	0	0,0%
R4_I5	3	14	0	1	33,3%
R5_I1	3	14	0	0	0,0%
R5_I2	2	7	0	0	0,0%
R5_I3	3	11	0	0	0,0%
R5_I4	3	7	1	0	33,3%
R5_I5	4	8	0	0	0,0%
R6_I1	5	22	0	1	20,0%
R6_I2	5	12	0	1	20,0%
R6_I3	5	17	0	0	0,0%
R6_I4	5	8	0	0	0,0%
R6_I5	6	28	0	0	0,0%
Total	158	374	1	3	1,9%

image ID	Ground-truth		External researchers		
	#labels	#labels	number of full matches labels	number of partial matches labels	average number of matching (full and partial) labels
R1_I1	3	2	0	1	33,3%
R2_I1	10	3	2	0	20,0%
R2_I2	13	2	0	1	7,7%
R2_I3	10	2	0	0	0,0%
R2_I4	11	4	1	2	27,3%
R2_I5	12	4	1	1	16,7%
R3_I1	4	7	0	0	0,0%
R3_I2	5	3	0	1	20,0%
R3_I3	5	3	0	1	20,0%
R3_I4	6	4	2	1	50,0%
R3_I5	6	5	0	1	16,7%
R3_I6	5	7	0	0	0,0%
R4_I1	3	2	0	1	33,3%
R4_I2	3	3	0	1	33,3%
R4_I3	5	2	0	1	20,0%
R4_I4	3	2	0	1	33,3%
R4_I5	3	2	1	0	33,3%
R5_I1	3	7	1	1	66,7%
R5_I2	2	3	0	0	0,0%
R5_I3	3	4	1	0	33,3%
R5_I4	3	8	0	0	0,0%
R5_I5	4	3	1	0	25,0%
R6_I1	5	3	0	1	20,0%
R6_I2	5	2	1	0	20,0%
R6_I3	5	4	0	1	20,0%
R6_I4	5	3	0	1	20,0%
R6_I5	6	3	0	0	0,0%
Total	158	97	11	17	17,7%

Figure 4.4: Correspondence of labels between the three types of description

As for the average number of labels, with the exception of Researcher 2 (R2) who has an average of 11 labels per image, the others are similar, having an average of 4 labels per image. As for Vison AI, it obtained the highest numbers, with an average of 14 tags per image. As for external researchers, they show very similar numbers to ground-truth, with an average of 4 labels per image. In the last column of the tables in the Figure 4.4 (average number of fully and partially matching labels), can see that Vison AI presents averages with little similarity, with only 4 images having matching labels (1.9%). The external researchers show more similarity of labels, and only 6 images do not have any matching of labels (17.7%).

4.2.5 Final remarks

Although images have an undeniable capacity for representation, having access to the information they make available is no guarantee of full knowledge of what they translate. People are familiar with operations related to written text, but there are still many things to be studied when reading images, be it a painting, a film, a photograph, a graphic, or another. It is necessary to understand that the image is no longer just art and has become information and knowledge.

With technological advances, namely, the countless possibilities provided by Artificial Intelligence, much progress has been made in the sense of collecting and representing the content of images in an automatic way. It is undeniable the important role that these tools have in the most diverse areas, however, it is necessary to analyze all the challenges and limitations they include.

This work has shown that automatic image processing tools achieve better results when used on images such as photographs, as they more easily pick up the features present in the image. However, when it comes to graphics or diagrams, the scenario is different. In addition to the text inserted in the image, other labels are less consistent. It was found that this analysis is done in a very literal way, obtaining labels such as "rectangle" or "number" when it was a rectangular graph with associated numbers, leaving the theme and the variables associated with it undescribed.

It was also noticeable that formal analysis is much more present in automatic processing tools, rather than semantic expression analysis. This only confirms that these technologies are not yet fully available to delve into the semantic layer and meaning inherent in images. We had an example where one of the images was intended to reinforce the idea of "animal welfare", however, this could not be identified with a formal analysis of the image and was therefore not a concept identified by Vision AI.

Is necessary to raise awareness of the fact that human inclusion in the description process is important, as it allows validation of the results, guaranteeing their reliability. Automatic processing tools can be great helpers for researchers looking for help in the first layer of description, even more so when the volume of images is large, however, these should always be included in the workflow, as their knowledge of all the details inherent to the images is fundamental to a good description.

4.3 Solutions for data sharing and storage: a comparative analysis of data repositories

Sharing and reuse are the last steps in RDM process. It is essential to ensure that the data stored in digital repositories is well preserved in the long term and that its adequate interpretation and future reuse are guaranteed. Following this debate, questions arise related to the interoperability of systems and the suitability of platforms. In this section, we study how data management platforms can solve the problems associated with description, preservation, and access in digital media, making their usefulness evident. We identify some of the most relevant repository platforms in the scope of research data management, offering an aggregating view of the various solutions and their main characteristics, thus aiming at a better understanding of them for their appropriate choice.

4.3.1 Contextualization

Currently, the number of articles and datasets produced in science is increasing [7]. Allied to this, there is a growing awareness of the importance of producing. This fact motivated the publication and sharing of data to gain greater importance for researchers who want

to see their data properly organized, stored, and described, in order to promote their sharing, reuse, and citation. The management of these resources has become a concern for researchers and research institutions. Proper data management contributes to the reproducibility of science, the reduction of duplicate efforts in data production, and the possibility of comparing results with those of peers [63, 76].

One of the main results of this practice is that researchers use platforms to manage their data, share them with the entire research community, and contribute to their preservation over time [7, 59]. The diversity of platforms available makes it difficult to choose, as it becomes more complicated to choose the one that specifically meets identified needs. Some of the obstacles are in defining metadata for description and long-term preservation [7].

The abrupt growth of data production is one of the reasons that promote repository platforms to implement functionalities to describe the datasets that are deposited on them [8]. The various stakeholders in this process, whether research institutions, individual researchers, or curators, contribute to the description of the data produced. In this context, the role of curators is intrinsically linked to maintaining the accessibility, quality, and integrity of data over time [7]. Thus, the description that researchers make of their data, combined with the metadata of the datasets themselves, creates the necessary conditions for their future reuse and citation by other researchers. The support of protocols such as OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) for retrieving metadata from different sources and creating an interface for displaying indexed resources makes the dissemination of data more effective [73].

Taking into account all the aspects mentioned, in many cases, stays difficult to conveniently choose the platforms that support the data repositories. Thus, a summary of the characteristics and functionalities of each one of them can be a good starting point for institutions to make an informed choice [8].

4.3.2 Repository analysis

The selection of platforms took into account their relevance in the context in question, recognizing their influence and usefulness in the context of research data management. Based on statistics from the Registry of Open Access Repositories (ROAR)³, eight software solutions were selected, but this was not the only criterion. Priority was given to platforms that support interoperability features and that also support repositories of research and government institutions. Issues associated with the description of datasets or the definition of metadata were also taken into account. Thus, the following platforms were considered: CKAN, Dataverse, DSpace, ePrints, EUDAT, Figshare, Zenodo, and Islandora.

Each of the platforms was analyzed according to six criteria. These criteria were derived from the documentation of the platforms and the experimentation of demo instances. They are divided into six categories whose structure is inspired by the Open Archival Information System (OAIS) model [88] and can be seen in the next subsections.

³<http://roar.eprints.org/view/software/>

4.3.2.1 Infrastructure

In Infrastructure, the fact that most platforms are open source stands out, giving them the advantage of being able to be updated and customized without changing the main features. The possibility of installing the platform on servers, such as DSpace, Dataverse, Islandora, ePrints, and CKAN, gives institutions full control over the deposit of their data. Table 4.5 presents this information. Basic attributes that support the functioning of the platforms.

Table 4.5: Comparison of platforms for the infrastructure component

	Infrastructure			
	Open Source	Storage	Installation	Payment system
CKAN	✓	Local or remote	Installation	x
Dataverse	✓	Local or remote	Intallation or service	x
DSpace	✓	Local or remote	Installation or service	x
ePrints	✓	Local or remote	Installation or service	x
EUDAT	✓	Remote	Service	x
Figshare	x	Remote	Service	x
Zenodo	✓	Remote	Service	x
Islandora	✓	Local or remote	Installation	x

4.3.2.2 Ingestion

At Ingest, the focus is on the general acceptance of OAI-PMH protocol platforms for interoperability between repositories, while simplifying data dissemination and the SWORD (Simple Web-service Offering Repository Deposit) protocol for API support, as a means of data deposit. In the case of CKAN, the OAI-PMH acceptance functionality is only available through an extension to the platform⁴. Table 4.6 presents this information.

Table 4.6: Comparison of platforms for the ingestion component

	Ingestion		
	Interoperability	API of deposit of data	Automatic import of metadata from files
CKAN	OAI-PMH	SWORD	x
Dataverse	OAI-PMH	SWORD	✓
DSpace	OAI-PMH	SWORD	✓
ePrints	OAI-PMH	x	✓
EUDAT	OAI-PMH	SWORD	✓
Figshare	OAI-PMH	SWORD	✓
Zenodo	OAI-PMH, REST	SWORD	✓
Islandora	OAI-PMH	SWORD	✓

⁴<https://extensions.ckan.org/extension/oaipmh/>

4.3.2.3 Content organization and control

In the Organization and control of content, it should be noted that the platforms present informational content structured hierarchically or in a linear way. Access to data is also limited, either through authorizations granted to each particular user or to specific groups, as is the case with DSpace and CKAN. All platforms offer the possibility of embargo, licensing, and elimination of deposits. It is relevant to emphasize the issue of maturity of deposits, as it is important to give the possibility to deposit, regardless of the maturity state of the data. Tables 4.7 and 4.8 present this information.

Table 4.7: Comparison for the content organization and control component - part 1

	Content organization and control					
	Organization of contents	Access granularity	User profile	Authentication	Support of attached documents	Embargo
CKAN	Linear	Grups	✓	✓	✓	✓
Dataverse	Hierarchical	File, User	✓	✓	✓	✓
DSpace	Hierarchical	Groups	✓	✓	✓	✓
ePrints	Hierarchical and linear	User	✓	✓	✓	✓
EUDAT	Linear	User	✓	✓	✓	✓
Figshare	Linear	Institution, Publisher, Researcher	✓	✓	-	✓
Zenodo	Hierarchical	User	✓	✓	✓	✓
Islandora	Hierarchical	User	✓	✓	✓	✓

Table 4.8: Comparison for the content organization and control component - part 2

	Content organization and control					
	Volume and size	Eligibility of depositor	Language	Data maturity	Deposit elimination	Licensing
CKAN	-	All allowed	All allowed	Any state	✓	✓
Dataverse	database, files and records: 2GB-10GB	All allowed	All allowed	Any state	✓	✓
DSpace	-	All allowed	All allowed	Any state	✓	✓
ePrints	-	All allowed	All allowed	Any state	✓	✓
EUDAT	max 20GB per record, max 10 GB per file	All allowed	All allowed	Any state	✓	✓
Figshare	max 20GB per record	All allowed	All allowed	Any state	✓	✓
Zenodo	max 50GB per record	All allowed	All allowed	Any state	✓	✓
Islandora	-	All allowed	All allowed	Any state	✓	✓

4.3.2.4 Metadata

In Metadata, special attention is given to the flexibility aspect of metadata schemas. DSpace, Islandora, CKAN, EUDAT, and Zenodo accept several schemas in addition to the default ones. With the exception of CKAN and Figshare, all platforms support the export of schemas in commonly used formats. It should also be noted that most platforms include the validation feature, which adds a guarantee of good and correct descriptions. Table 4.9 presents this information.

Table 4.9: Comparison for the metadata component

	Metadata			
	Schema/ Standard/ Model Flexibility	Schema Export	Validation	License registration
CKAN	✓	x	x	✓
Dataverse	x	DDI, DC	✓	✓
DSpace	✓	QDR, MARC, MODS	✓	✓
ePrints	x	-	✓	✓
EUDAT	✓	DC, MARC, MARCXML	✓	✓
Figshare	x	DC	x	✓
Zenodo	✓	DC, MARC, MARCXML	✓	✓
Islandora	✓	DC, DDI, MODS, METS	✓	✓

DDI: Data Documentation Initiative; DC: Dublin Core; QDR: Qualification Dataset Register; MARC: Machine-Readable Cataloging; MARCXML: Machine-Readable Cataloging XML; MODS: Metadata Object Description Schema; METS: Metadata Encoding and Transmission Standard

4.3.2.5 User interface

In the user interface, the ability to customize the design of most platforms is highlighted. EUDAT is a ready-to-use service and offers its own interface that cannot be changed. As modules of this platform can be installed on the institutions' servers, it should be noted that the focus of this study is EUDAT as a ready-to-use service offered by the entity itself and not the installable version of the modules. Design optimization for mobile is an interesting topic of discussion given this is a future trend. Table 4.10 presents this information.

Table 4.10: Comparison for the user interface of component

	User interface	
	Customization of design	Design for mobile
CKAN	✓	✓
Dataverse	✓	✓
DSpace	✓	✓
ePrints	✓	-
EUDAT	x	-
Figshare	-	-
Zenodo	✓	-
Islandora	✓	✓

4.3.2.6 Articulation with other services

In the articulation with other services, it is evident, in general, that there is an acceptance of the platforms in view of the features of analysis of the raw data presented. At this point, the possibility

of recognizing values corresponding to geographic data in the datasets and thus aggregating the remaining data by their location was highlighted. Table 4.11 shows this information.

Table 4.11: Comparison for the articulation with other services component

	Articulation with other services				
	Media viewing and reproduction	Tabular data graph	Georeferenced data analysis	Diverse data types	Data access via API
CKAN	-	✓	✓	✓	-
Dataverse	-	✓	✓	✓	✓
DSpace	x	-	✓	✓	x
ePrints	✓	-	-	✓	✓
EUDAT	✓	✓	✓	✓	✓
Figshare	✓	✓	-	✓	✓
Zenodo	-	✓	-	✓	✓
Islandora	✓	✓	✓	✓	✓

4.3.3 Final remarks

The analysis of the repositories proved that the selection of a platform for data management can be a difficult task, as it is necessary to assess the concrete needs in each situation. An important aspect to focus on is the fact that repositories tend to be increasingly prepared to be in line with data interoperability and accessibility guidelines. Of course, data sharing is one of the goals of data repositories, however, it is necessary to ensure that, when accessed, the data are properly interpreted, as this will guarantee their reproducibility. Data reused by third parties guarantee credit to authors, through citations and references.

It is important to emphasize that repositories are not mere guardians of data. The competent description needs to be promoted through varied metadata models and domain-specific vocabularies. It is necessary to guarantee different conditions of access, such as full access to data or only access to metadata, but with the safeguard of the possibility of contacting the authors. Authors must be safeguarded by associating a DOI (Digital Object Identifier) to the data while facilitating the citation and reuse process. There are several challenges, however, the advantages will outweigh all efforts.

4.4 Guidelines for a standardized practice in the use and production of images in research

This section identifies guidelines to help researchers using or producing images in a research context. Throughout the various experiences carried out during this work, it became apparent that, although they considered images to be important parts of their projects, most researchers had never reflected on the subject more consistently, nor had they thought about formalizing practices to apply to their projects. The aim of these guidelines is precisely to offer some tips, based on practices

mentioned in the literature and consistent with the principles of research data management. The recipients are intended to be researchers, regardless of their field of research and their scientific maturity and experience. We believe that there is no rigid standard of guidelines, but establishing some habits can benefit the scientific process and open up greater debate and scientific production on the subject.

4.4.1 Contextualization

As explored in our work, images in research can be important parts of the research process, whether they are data or metadata. Images should not be captured or chosen in such a way as to be manipulated and used inappropriately, even if it is to serve the research objective a priori.

Depending on the assigned purpose, images can assume different roles and importance, can be produced or used at different times in the research process, and can be associated with different collection, editing, and analysis methodologies.

These guidelines were designed taking into account the practices highlighted in the literature, personal experience throughout this doctoral journey and always taking into account the principles that govern the management of research data. It is important to emphasize that these are not general guidelines for the use and production of images, but rather guidelines for the use and production of images in a research context, based on data management.

The guidelines will be presented along with some examples. It's important to note that these are only some of the possible examples to elucidate the guidelines given at the time.

Finally, leave a comment on this section. At the beginning of the doctoral journey, we expected these guidelines to be a significant contribution of this work, as they aim to give clear indications regarding the use and production of images in science. The gap in the literature on this topic further reinforced this idea. At this point, this conviction remains, however, it has become clear that, although images have specific characteristics that differentiate them from other types of resources, it is impossible to dissociate some of the general RDM guidelines. That is, during this section (4.4), it will be possible to see guidelines very oriented towards images and in particular on some facets of them. However, at certain times it is possible to verify a relation with the general guidelines for other types of data. With this, we do not believe that the guidelines lose strength, on the contrary, we believe that they allow us to reinforce the importance of the principles of research data management, highlighting the phases in which the fact that they are images requires more targeted guidance.

4.4.2 General guidelines

There is no detailed standard for the production and use of images in research. However, this work made it possible to determine which guidelines should be followed to guarantee the integrity of the images and the investigative process.

Images can take an analog or digital format. Most of the images currently produced are digital (very motivated by technological development), however analog images continue to be produced in different contexts and their importance should never be overlooked.

In a scientific context, when creating images it is of great importance to guarantee their quality. In the case of digital images, it is necessary to check that the image capture instrument is properly calibrated and aligned with the object. In the case of analog images, it is necessary to ensure that they follow the defined production requirements, taking into account the context for which they are intended [32].

It is possible to say that digital images have challenges that are different from analog images, as the immense possibilities of processing, analyzing and editing images, based on increasingly advanced tools and software, call attention to other precautions. For this reason, in the use and manipulation of images, especially digital images, there are some ethical considerations that should guide their use [31]. According to Cromey [31], these considerations can be seen as assumptions or guidelines, namely:

- Scientific images can be compromised by improper manipulations;
- Manipulations of digital images should always be performed on a copy of the original image;
- Simple adjustments to the entire image are acceptable;
- Image cropping is usually acceptable, as long as it does not alter the content of the image;
- Images that are to be compared to each other must be captured under identical conditions and undergo identical processing;
- Manipulations performed only on specific areas of the image are questionable;
- The use of filters to improve image quality is not recommended;
- Cloning or copying objects from other images and composing a new image is very questionable;
- Avoid using image formats with compressions that affect image quality;
- Intensity measurements should be performed on a uniformly processed image calibrated to recognized standards;
- Resolution and magnification are important;
- Caution is required when changing the size, in pixels, of the digital image.

4.4.3 Guidelines for research image management

In addition to the more technical and general issues, such as those presented above, the investigation process that involves the use and production of images must consider other guidelines for the

management of these images, especially in the sense of following the guidelines indicated in the scope of Research Data Management. It is really crucial to standardize processes, establish methods, and ensure uniformity in image management, as this will make it possible to avoid failures and redundancies.

Throughout this work, the research life cycle has been used several times as a tool for analysis. In the case of these guidelines, it was considered that this life cycle could be an ally in order to systematize the information that is intended to be conveyed since it groups the various guidelines by life cycle stages. Figure 4.5 presents the research life cycle used to establish and organize these guidelines

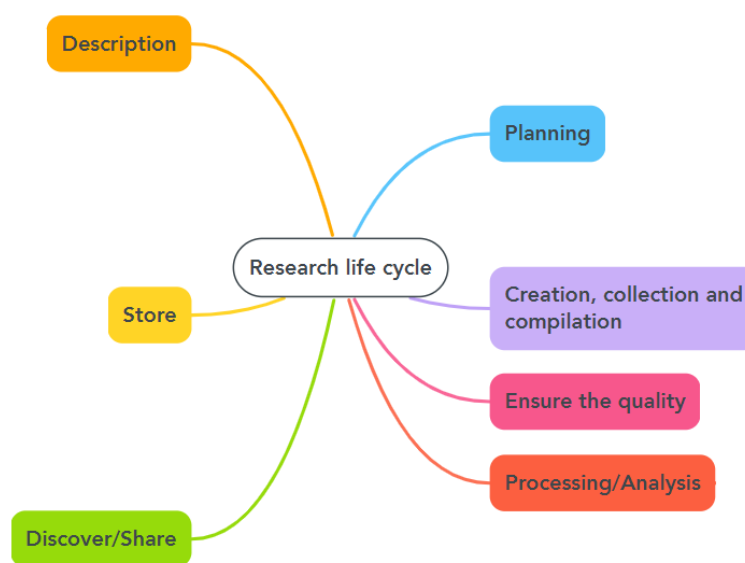


Figure 4.5: Research life cycle

In order to create a set of properly structured guidelines based on good practices, we followed suggestions from the Consortium of European Social Sciences Data Archives (CESSDA)⁵, in its 2019 document "Adapt your Data Management Plan: A list of Data Management Questions based on the Expert Tour Guide on Data Management" [92]. This document offered some general ideas that allowed us to organize these guidelines in the context of the images that we consider important to mention.

4.4.3.1 Planning

The planning phase is essential whenever we talk about data management, regardless of the type of data. As far as guidelines for image management are concerned, this should be supported by the use of a planning document. This should be seen as a fundamental tool for organizing and structuring all the procedures inherent in the use and production of images throughout the research process, i.e. from the moment it is planned until it is completed.

⁵<https://www.cessda.eu/>

There is no need for a document that applies exclusively to images. The document can be generic for the project, but include a section dedicated to image management. A good research practice is the use of Document Management Plans (DMP), where the entire process relating to the project and its data is documented. The DMP can include a section where image management is explained.

This planning document (whether it is a DMP or not), in addition to explaining the role of data, including images, should contain a set of good research practices. The planning document should take place at the beginning of the project, to guide all the following stages. By carrying out this procedure, the planning phase of the life cycle is guaranteed.

Specifying the component of the document relating to the images, it is important to note that, in addition to all the information on the phases of the life cycle of the images, this document must guarantee the presentation of general information about the images. This information is detailed in Table 4.12.

Table 4.12: Guidelines for the planning phase

Guidance	Example
<i>General information about the images</i>	
Origin of the images	Captured in the laboratory via microscope
People involved	Principal researcher and research assistants
Definition of the project where the images are inserted	Experience "2.0" within the scope of the "XPTO" Project
Copyright issues	Copyright: exclusive right of the author to reproduce the resource. License: CC BY
Financing	Funded by the research foundation "XYZ"
Contacts	\color[HTML]{333333} example@project.com
Costs	200 €
Dates	May 2022 - September 2022
Purpose of images	Elaboration of a case study using direct observation of a phenomenon via microscope, with image capture
Type of images	Microscopic images
Format of images	*.png
Edition	The images were cropped to measure 210mmx148mm
Quality	Resolution: 1080p
Sample	45 images

The most specific aspects of each image or set of images must be specified later in the detailed description of the project (e.g., collection methodology, analysis techniques, processing tools, among others).

4.4.3.2 Creation, collection and compilation

Issues relating to the creation, collection and compilation of images should also be considered and followed according to certain guidelines.

In addition to the more technical aspects of images, this phase of the life cycle must include image organization and storage issues. When one or more images are taken, various details need to be documented. Table 4.13 shows the guidelines that should be followed in this regard:

Table 4.13: Guidelines for the creation, collection and compilation phase

Guidance	Guidance description and examples
Formal characteristics of images	name, type, size, dimensions and format of images (e.g., "Figure A", photograph, 68,1 KB, *.jpg, W: 538 pixels x H: 473 pixels)
Organization of images	- Location where the images will be stored, - Structure of folders and files related to the organization of the images
Collection of images	In the case of image use: e.g., obtained through the "XPTO" data repository In the case of images production: e.g., obtained by researcher X, in Furnas do Enxofre, on Terceira Island, between April 22, 2019 and April 24, 2019, using the Canon EOS Kiss X2 Reflex 12.4 camera, in rainy weather conditions, with a temperature of 11°C, wind 18k/h and relative humidity of 52%).
Purpose of images	Specify how existing images will be used in the project e.g.
Reuse of images	In the case of reusing images, specify where the images are located, how their search is planned, what the access terms and conditions are, what the expected process and time span for obtaining access to the images are, and what the expected cost relative to this access.

4.4.3.3 Processing/Analysis

The way an image is processed and analyzed is decisive in obtaining reliable and trustworthy results. It is not enough to collect an image, it is necessary to determine the procedure and method inherent in the way it will be processed and analyzed. Table 4.14 sets out the guidelines on this topic.

Table 4.14: Guidelines for the processing/analysis phase

Guidance	Example
Definition of a strategy for the development of image versions	All images must be backed up, assumed as "backup_figure x". Each time an image is edited, it must be saved as a copy (always keeping the original) and renamed, by including a version identifier (example: "sceloporus magister_version 2")
Who will be responsible for their maintenance and elaboration of criteria for the distinction of versions	Principal researcher
Determination of issues relating to interoperability, namely the use or not of ontologies and terminologies	Use of the OpenCyc ontology
Use of software, and coding issues	Use of the Pix4D (professional photogrammetry and drone mapping)
Determination of image analysis approaches, methods and procedures	Content analysis using qualitative methods

4.4.3.4 Ensure the quality

Quality issues have very specific characteristics in images. If, on the one hand, the content of the image is essential, more technical characteristics, such as the format in which it is stored or even its dimensions, can be essential in guaranteeing the quality of an image. Quality, with the use of editing, can be related to trying to solve some problem with the image or to improve some characteristic. Quality information must be taken into account. Table 4.15 presents the guidelines for this topic.

Table 4.15: Guidelines for the ensure the quality phase

Guidance	Guidance description and examples
Editing and image quality issues	- Identify if there were problems with the image that require editing if these problems are not identified, and explained the reason for the editing; - Mention important quality issues for the study (e.g. resolution, color, brightness, saturation, etc.)
Determine the minimum quality required for images	The minimum quality required for images must be determined in terms of origin, content, size, format and methods
Explain how the quality of the images will be evaluated	Determine in advance which criteria will define whether the image has sufficient quality (example: clear identification of all elements present in the image, correspondence of the colors present in the image with reality, specific image format, etc.)

4.4.3.5 Store

There are various data storage solutions, but the way and method in which data is stored is decisive for aspects such as organizing data, retrieving information and safeguarding resources, for example. With this in mind, it is important to determine the practices to follow when it comes to storage. Table 4.16 presents guidelines with corresponding examples.

Table 4.16: Guidelines for the store phase

Guidance	Example
Define where and for how long the images will be stored	The images will be stored in the cloud on the project's resource management platform. Backups will be stored on the project's external disk. The images will be stored for up to 6 months after the total completion of the project except those present in publications, communications, and platforms such as repositories and directories (for these see the last line of this table)
Specify how and for what periods backups will be carried out	Image backups must be carried out at the same time as the original images are stored in the cloud. Every time a new version of the image is made, a new backup must be made, to be included on the external disk where the other backups are located.
Determine an image recovery strategy in case of loss	If the original image is lost, it must be backed up and a new copy made to the backup disk.
Clarify how access to images will be managed and how data security will be ensured (contingency plan if necessary)	Access to each of the project elements will be determined. The principal investigator and research assistants will have access to all the images and will be able to manage the sharing of some of the resources with other elements. The other elements will have access restrictions. Access will be guaranteed using passwords.
Location and method of storing the images after the project has been completed	After the completion of the project, images that are not part of project reports or publications/communications arising from the project will be deleted (6 months after the total completion of the project). Images that are present in publications, communications, datasets or platforms such as repositories and directories will be kept, always indicating the copyright and licenses for access and use.

4.4.3.6 Description

We confirmed that there is no model or set of elements applied solely to the description of images as data. For this reason, it was essential to develop the metadata model and controlled vocabularies highlighted in Section 4.1.

For this specific topic (description), within the guidelines, it was realized that it is essential to explore the practices to follow a little more. The specification of what documentation will be produced in relation to the images could be used as a guideline, as well as what description characteristics will be applied, namely the use of metadata and metadata models. However, it

is important to go beyond this and reinforce the idea of the dual role of images: as data and as metadata. This information is described in Table 4.17.

Table 4.17: Guidelines for the description phase

Images as data	Images as metadata (explored in Section 5)
The guidelines to be followed should be guided by the good practices standardized by the principles of the RDM	Images should be seen as the agent of description. That is, the value to be included in the descriptive fields
The metadata model proposed for the description of images, although not a fixed and watertight tool, can be fundamental in the description of image data, especially when it comes to specific characteristics of this typology, such as issues associated with quality, editing, and scale, for example	If, for textual and numerical values, the indication is a description oriented towards due interpretation, using clear language, focusing on the main information to be presented and without interpretative duality, the same applies to image values. It must be ensured that the images or set of images allow a true reading of what is intended. It is not intended that the image is just a decorative element that accompanies the data, it must be able to be "read" and understood
The use of controlled vocabularies should be seen as an important guideline, as it makes it possible to establish a set of vocabulary elements that restrict the description values, making it possible to standardize the description and improve interpretation	The image must be accompanied by other values, such as text, whenever necessary
When it comes to images, the description of the data must go beyond what is visible in the present content of the image, i.e. the formal analysis must be accompanied by the analysis of the semantic expression (spoken in Section 4.2.2)	The person performing the description must be able to distance themselves from their condition of an unequivocal connoisseur of the image and put themselves in the role of someone viewing it for the first time. They must therefore be able to check that the image contains all the elements necessary for its descriptive function and add other elements whenever they consider that the imagery is not sufficient

4.4.3.7 Discover/Share

Increasingly, data discovery and sharing practices are a concern for researchers. Scientific production aims to disseminate it to raise new hypotheses and test ideas. Today, properly cited data or publications are fundamental to the reputation of researchers, their projects, and their research centers. The dynamics of open science promote publication practices that take into account ethical issues, copyright, and transparency, for example through the use of DOI's and specific reuse and citation conditions. Table 4.18 shows the guidelines identified as important for this topic.

Table 4.18: Guidelines for the ensure the discover/share phase

Guidances for discover	Guidances for share
Guarantee that the images will be archived in a reliable data repository	Ensuring the use of a persistent identifier for images
Specification of how the images will be made available (open access or not)	Specifying the formats in which the images will be shared
Number of images to publish (all or just a few?)	Determining the need to use specific software to process the images? If yes, define whether this can be deposited/shared together with the images,
Identify the licenses required for images	Determining how images should be cited when reused
Definition, if necessary, of an embargo period for (all or some of) the images	Specify the existence of agreements or access restrictions that need to be considered
Determine whether there is a risk of delaying the publication/availability of images (in whole or in part). If so, explain what needs to be done to avoid this.	Specify the existence of legal/ethical restrictions that prevent the publication and share of images. Explain whether these restrictions will require action to be taken before images are made available,

4.4.3.8 Image protection

It became clear that the life cycle used should include a phase dedicated to the protection and privacy of data in image format. As this is an emerging issue and the subject of many regulations, it was decided to include a set of guidelines to guarantee this component in image management:

- define whether or not it is mandatory for an ethics committee to approve the use and production of images,
- determine whether it is necessary to request informed consent for the use and production of the images, if so how this permission will be obtained, and how the images will be organized and stored. In the case of informed consent, it must follow the model: "Informed, Free and Clarifies Consent to Participate in a Research Project in accordance with the Declaration of Helsinki⁶ and the Oviedo Convention⁷". This consent allows the participants to attest 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 publicize the identification of participants.
- specify how access to images that constitute personal and sensitive data will be controlled,
- define how the security of this sensitive data will be guaranteed,

⁶<http://epidemiologia.med.up.pt/pdfs/Helsing.2013.pdf>

⁷<http://dre.pt/pdf1sdip/2001/01/002A00/00140036.pdf>

- provide for the existence of a requirement to notify the privacy officer of images that contain confidential information or special treatment,
- clearly specify if there are images that, due to their privacy conditions, require special treatment or access limits during and after the project,
- how images that constitute personal or sensitive data will be protected during and after the project,
- determine how permissions and restrictions will be reinforced,
- specify whether there are any ownership/copyright issues associated with the images, including clarifying what permission is needed to collect and reuse images and whether, in the case of reuse, rights will be transferred to another organization,
- provide for any restrictions that may occur in terms of privacy.

Chapter 5

Image as metadata

When working on the dynamics of images in the context of research data management, it became apparent that there was a large gap in theoretical material on this topic, especially when it comes to images as metadata. This aspect sparked an interest in exploring it further, paving the way for more material on the subject to be disseminated to the scientific community.

This chapter addresses the issue of images as metadata. This topic was essential, as it allowed us to explore a new hypothesis that translated into a new perspective for data description, using images as metadata.

This chapter is aligned with RQ3, aiming to give depth to the question, trying to raise hypotheses and record conclusions that lead to its answer.

5.1 A new perspective for describing research data

Through studies and work developed over the last few years, we propose a new approach to description, where images can have a preponderant role in the description of data, assuming the role of metadata. In this section we present several pieces of evidence, point out their challenges and determine the opportunities this new perspective can have in the research.

5.1.1 Contextualization

Research Data Management (RDM) includes activities that organize and manage the life of a research project and is crucial for consistent work performance [51]. Some activities are related to the description, which is a fundamental step, since it allows data to be properly documented and interpreted, promoting their subsequent reuse and sharing. There are guidelines for research data management, highlighting the most important aspects in each phase. In the description phase, one of the most important of RDM, the advantages of data description, as well as assisting tools are indicated. The tools suggested for this phase include metadata templates, which are essential for researchers who want to see their data properly described. However, this description is always oriented towards using text as metadata values. We believe that the exclusive use of textual description can be discussed, because, although the text is fundamental in the description and,

therefore, should never be dispensed with, its use together with images can bring added benefits. This work puts into perspective how images can support research as metadata, contributing to the description and interpretation of data.

Given the importance of the image and its frequent inclusion in research, a set of general objectives have been defined: 1) show that the use of images as metadata can improve the description and interpretation of data, 2) present an innovative perspective for the description of data, highlighting the role of images in cognition and science.

5.1.2 Methodology

This work presents a new perspective for the description of data through the use of images and has chosen to follow consistent methodological guidelines. We opted for an approach presented by the Journal of Biogeography [91], as it seemed to us that it well illustrated the objective of the work.

As can be seen in Figure 5.1, we began by identifying the topic under discussion and analyzing its current position. Next, a set of hypotheses was defined to support the new perspective. These hypotheses, as well as the need for a new perspective, emerged from the work carried out in recent years and support the idea that it is beneficial to include images and their potential in the data description process. Once the new perspective was defined, the evidence that made it viable was identified. This evidence is supported by examples and relevant literature. As a final point in this process, it is important to clarify the strengths of the new perspective, without omitting possible weaknesses or added challenges. The opportunities and advantages identified served as a basis for feeding new positive implications into the research.

The use of this methodology makes it possible to clarify the current situation, elucidate the existing limitations, show many evidences and examples, and, thus, strongly support the new perspective.

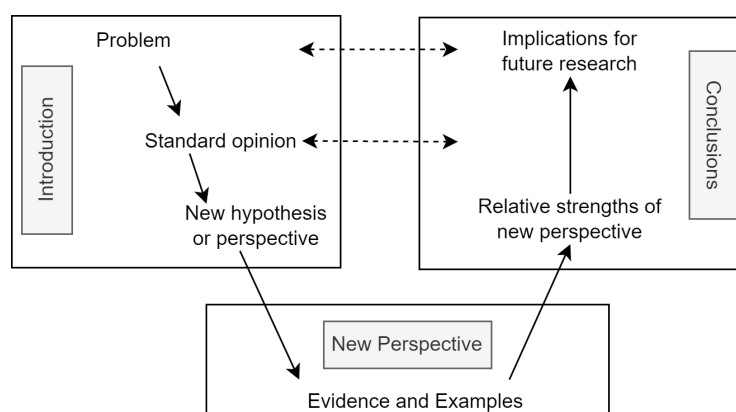


Figure 5.1: Methodological approach to the new perspective

Source: Adapted from the *Journal of Biogeography*, 2021
[91]

5.1.3 Data description in RDM: Standard Approach

In this section, the current scenario regarding the dynamics of using images to describe data will be presented. It was the lack of references to the topic of images as metadata that opened the door to the new perspective, as it was clearer to understand the importance of the new perspective suggested.

5.1.3.1 Motivation

As discussed in greater detail in the section 2.1, research data management is a growing topic of discussion in the scientific context. RDM includes several phases that allow you to plan, organize, process, store, preserve, and share data. One of these phases includes data description. In 2016, with the dynamization of the FAIR principles, new ideas were discussed that could act as a guide for the management of research data [51]. Like other types of resources, images can be a part of these principles, whether in the data organization process, above all in the implementation of guidelines for accessing research results, fostering the entire Open Science ecosystem. The ability to make data findable, accessible, interoperable, and reusable can be promoted by including new perspectives, namely the use of images to describe data and improve its interpretation.

5.1.3.2 Data Description

As mentioned in particular detail in the section 2.3, a huge amount of data is produced daily, in different research contexts, and the description can be a fundamental piece in the project organization. Firstly, project researchers manage to have their data properly annotated, improving communication within the group and adding context to the data. Secondly, the description promotes the sharing, citation, and reuse of data, as interested parties are able to retrieve/search more efficiently the data that are shared by their authors and can also more easily understand the context of the data and specific characteristics that only the analysis of raw data did not allow. These conditions encourage other researchers to cite the data more and reuse them for other projects [7].

The current requirements for open data in the European Union (EU) are increasing the awareness of researchers concerning data management and data publication [7]. Metadata is essential in research data management, namely in data discovery and reuse. The specialization of researchers makes them natural providers of domain-specific metadata, even more so in the long-tail of science [102]. As data creators, they have unique knowledge of the data production context, including domain-specific concepts and configurations used in the production of the data. Current practices tend to either leave metadata definitions to researchers or curators. The former typically results in ad-hoc elements, while the latter follows standards but lacks specificity.

Data descriptions produced by researchers were also found to be more focused on the details rather than on general features when compared to those made by information specialists, highlighting the difference between descriptions intended for personal archives and those meant for data repositories, with reuse in mind [134]. The relevance of RDM is visible in the growing body

of studies in diverse areas, some with a focus on researchers' perspectives and practices regarding data organization and sharing, while others have a closer look at researchers' description practices.

Directives call attention to the need to describe data. The description process is fundamental in research data management. Many standards are used today to describe data. There are some of a more generic nature, such as the Dublin Core, and others specific to the domain, such as the Darwin Core and the Genome Metadata, within the scope of Biology.

Metadata models are important pillars in the description. They contain a series of elements/metadata that act as reference points on the data and allow circumscribing information about them. Metadata allows for classifying and organizing data, enhancing search and interpretation which, in turn, promotes citation and reuse. Metadata models can follow a pre-defined structure, however, new elements can be used in order to complement the description, until reaching the desired level of maturity [134].

In all cases of description, the options given to depositors refer to elements/metadata with textual or numerical values. Doing thorough research on the various deposit platforms this fact is easy to verify. At no time is the possibility included in one of the fields of the metadata model to include a value in image format. This is not to say that the text should be replaced, but rather to use the unique characteristics of the images as a complement.

5.1.3.3 Problem Identification

Images can be fundamental to describing data and should not be discarded. However, we believe that images can also be used as a metadata value, with text not being the exclusive element for data description. Images are present in the context of research, so why not use this material to add value to the description of the data?

It is expected that this new perspective associated with images as metadata in research data management is extremely innovative and brings numerous advantages and challenges. RDM cannot be watertight. It must be evolutionary and changeable. If the tendency is to produce more images and if this production is increasingly facilitated by new technologies, why not use this to promote such an important stage of RDM, such as the description?

5.1.4 Images as Metadata: A New Perspective

"A picture is worth a thousand words". Everyone has heard this phrase. The scientific character attributed to images dates back to the 19th century and was essentially seen as a replica of reality that was transformed into a "perfect imitation of its object of duplicity" [120].

Roland Barthes [12] sees the image as the only way to reach the veracity of something, because "in it, a pipe is always a pipe, infallibly". For the author, the image document allows the confirmation of the real, it is, deep down, proof of what happened. The image has grown, it is not watertight and its growth has allowed the conquest of freedom of expression and respect for visual language. Still, Barthes says that the objective of the image has always been, above all, to

guarantee diversity in communication with others, so it is not strange to say that the image is a communication organ as capable as any other.

In science, too, the image can play an important role. Many times there is real value in using images in science as a means of promoting scientific content. The role of images in research is positioned in helping to understand, as they allow drawing attention to details and explaining concepts, theories, and methods that are difficult to explain in words [102, 66].

The problem presented in Section 5.1.3.3 is the starting point for the presentation of this new perspective: using images as metadata in the description process, in research data management. As mentioned in that section, the expectation is that there is a way to be discovered to take advantage of the potential of images in the description of data.

In this section, the new perspective under analysis will be presented, presenting the evidence that supports it, as well as the challenges and opportunities it holds.

5.1.4.1 Evidence 1: Visual Processing and Memory

A considerable part of the human brain is devoted to visual processing, hence the human being is considered a visual being [3]. The inspiration and appeal that images have are related to cognition and, consequently, to the ability to pay attention to details [10]. Contrary to what happens with other typologies, images are agents that attract attention very quickly and easily. In fact, the human brain processes images at a very high speed [3]. Let's see: when you see an image, our capacity for analysis occurs in a short period of time, which allows us to very quickly understand the meaning and scenario of the image in question [55, 10, 54]. It is even faster if the object in the image is familiar, because, if so, the human brain recognizes the object in 100 milliseconds. One of the essential components in this cognitive process related to images is colors, as the brain is programmed to react to color, which is why it is evidence that the most active human sense is vision [10].

Still, in the field of visual processing, it is possible to verify that impressions and memories are based on vision, since the eyes and brain work with images constructively, superimposing them at the time of their analysis [10, 47]. In research projects, which tend to cover a long period of time, memory is an important aspect, as researchers need to document various characteristics of the research process. Many times, the moment of documentation does not occur simultaneously with the moment of visualization of a procedure, discovery, data collection, or other [102]. Hence the use of images can be valuable in stimulating memory. The visual system processes sensory information about shape, color, distance, and movement of objects more quickly, in addition to enabling language functions, which is a necessary skill at different times in the scientific journey [10, 66].

5.1.4.2 Evidence 2: Science and Society

When looking for data on platforms such as repositories, we were looking for a central place where we could find what we wanted without too much difficulty. Metadata is a key element in this process, as it opens the door to data discovery, offering important details and the possibility of finding

a context that facilitates data interpretation [134]. In some cases, a more exhaustive description process is not used because the data management workflow can, sometimes, be conditioned by the difficulty of communication between the data producer/researcher and the data curator, which eventually limits a more detailed description. focused on specific metadata of different types, such as images. However, the need to assign context to data is increasingly visible in datasets that, in many cases, have other description elements, such as ReadMe files, for example.

The images can function as a facilitator in this search and interpretation. Reading, although fundamental, implies more time, and, therefore, the images can be allied at that moment. In addition, images are intrinsically related to memorability, and if the data has images in its description, its content will be memorized more efficiently [55, 54]. Images, such as photographs, drawings, maps, graphs, diagrams, illustrations, medical images, and microscopic images, among others, enable the interpretive process to become more fluid.

Nowadays, many research institutions are promoting their scientific work through online services such as blogs, using images as tools to disseminate studies and new scientific discoveries. Other examples are the web platforms to increase the visibility and impact of scientific production, such as kudos¹). In these cases, images act as a vibrant aspect to attract attention. In addition, they function as a faithful representation of reality, promoting a culture of truth in scientific work. Researchers and institutions see these as valuable tools for scientific communication. Exist examples like Biocanvas², which provides photographs that report scientific beauty over time. The Great British Biosciences blog (BBSRC)³ also provides the possibility to view bioscience images. Or even the blog Biomedical Picture of the Day (BPoD)⁴ where can be seen see biomedical images.

The information and communication society, in which we currently live, tends to record all transactions [120]. The image should not be demoted from its prominent place in this society, as this documentary typology constitutes a social object, with materiality and associated ideas. In science, this also happens. Currently, motivated by new technologies, researchers tend to use image capture devices, such as tablets or smartphones, to record project moments, namely the data collection or production process. However, the studies that have been carried out within the scope of this work demonstrate that there are no specific criteria or guidelines for the appropriate use of these images [47, 70].

Researchers state that images work to remember specific details, to reassess a scenario at the time of data analysis, to document or prove a situation, and for use in articles or conferences. Images constitute a benefit for the research process since they mediate the guarantee of communicational diversity and assist in interpretation. Lacerda [72] points out the capacity of images to record actions and information, and their competence to be a carrier of materiality and different communication and expression resources.

Current research is characterized by an unprecedented growth in the volume of data being produced, as powerful computational capabilities are available to even small research groups: the

¹https://www.growkudos.com/?utm_source=navbar-logo

²<https://biocanvas.net/>

³<https://bbsrc.tumblr.com/>

⁴<https://bpod.mrc.ac.uk/>

so-called long-tail of science [63]. Usually, small groups or individual researchers have minimal resources to ensure the long-term availability of their data. As such, they need adequate research data management practices supported by practical tools so that the datasets they produce can be made available to others [7]. This is especially important as more research funding agencies adhere to the European Commission's Guidelines on FAIR Data Management in Horizon 2020, which advocate for a set of principles to make data Findable, Accessible, Interoperable and Reusable [51].

In Figure 5.2, a simulated work task example of a data deposit can be seen in which the description model includes the use of images. The model is authorial.

The figure displays a simulated work task example of a data deposit on the INESC TEC website. The main content area shows the dataset 'Gamma radiation monitoring at the Azores ENA-ARM station (Graciosa Island)'. The dataset description states: 'Gamma ray total counts in counts/minute (cpm), every 15 minutes, from the Gamma Radiation Monitoring campaign at the ENA site (Graciosa, Azores). Data were collected to study the temporal variability of the property of the noble gas radon (Rn-222), aiming to examine how it is influenced by meteorological conditions and its association with the atmosphere's ionization and aerosols concentration. This dataset has Jupyter notebook for visualization.' The 'Data and Resources' section lists two datasets: 'dataGRM_ENA' (Time series of gamma total counts every 15 minutes) and 'GRM-ENA' (Dataset notebook). The 'Additional Info' section provides details about the dataset, including the field, source, author, last updated, created, DOI, and dc: Spatial Coverage. The sidebar on the left shows the organization's profile and social media links. On the right, a detailed metadata view is shown, including coverage, date, format, extent, language, publisher, relation, type, time series, and instrument name.

Figure 5.2: Simulated work task example of a data deposit

Source: Own authorship

5.1.4.3 Evidence 3: Practical Examples in Research Context

Palmer et al. [93] analyze the importance of creating a structure of principles and processes that help to articulate and support the description of data. One point of consensus was the crucial role of images in documenting the investigation. For researchers, the images have a double purpose: they provide context and as a vital medium to record the object of study. This approach leaves room for positioning an image in RDM as a fundamental asset in the description and interpretation of specific domains. Palmer's study is evidence of the need to bridge the gap between researchers and data curators, while guidelines and tools to support metadata and RDM are being created.

Images appeared in science in the nineteenth century when they began to be produced in Astronomy projects [122]. Its potential was quickly realized, given its high capacity for representation [120]. Silva [122] states that images allow different brain actions of memory and recognition. In this study, it is reported that when confronted with images or videos, users tend to remember more quickly what had been forgotten. Supporting this thesis are the studies by Czerwinski and

Horvitz [34] and Elswiler, Ruthven, and Jones [43], who claim that although humans forget many details of the tasks they perform, with the help of images they easily remember what they haven't remembered.

In History, the use of images often serves as a basis for the development of the process of analysis and interpretation. With the digitization of images and the creation of digital collections easily accessible via the Web, historians now have access to images that could not be accessed in their physical format, as well as tools that enable the discovery of new details in images [118]. In Health Sciences, X-rays and ultrasound are high-value images being used as clinical data and anatomical diagrams to communicate critical information. The production of images, in this area, has an important role in clinical practice, surgery, histology, physiology, and forensic medicine [122].

Al Nasar, Mohd, and Ali [6] talk about the role of human behaviors in the life cycle of images. They mention the awareness and education of institutions for the practices of interpreting research projects with the use of images and comment on the importance of this for eliminating description mistakes that the automatic systems contain and the forgetting of some details by the researchers. Also, Kang, Bederson, and Suh [70], Suh and Bederson [127], Sinha and Jain [123], and Mota et al. [84] focus on the issues of automatic processing of images, reflecting on how it should be combined with human interaction for evaluation and error correction.

5.1.5 Challenges and Opportunities for RDM

Research data management is increasingly a concern for researchers who need support in the organization of their data [7]. This constitutes a valuable contribution to information management, since, through the description, data keep their meaning and are more easily interpreted. Also, data management can be fundamental in the communication between researchers and the scientific community, where images can play a central role in the "teaching-learning" process [1].

In many cases, the description of the data appears at the time of its publication/deposit, with a view to proper storage, data sharing, or even following requirements that some journals and scientific conferences.

One of the great challenges of using images as metadata in the context of RDM arises precisely in the publication/deposit phase, especially with regard to existing technological solutions. Many of the data deposit platforms, like repositories, provide the possibility to select one or several elements and metadata models to assist in the description process.

Applying a single metadata model may not fully match description needs in a specific context. In these cases, the collection and selection of elements from one or more models are carried out to build a combined scheme, without losing the interoperability of the base schemes. The final result is the application of a profile characterized by allowing to specify of the controlled vocabulary, format, and relations of the elements, refining definitions of the standards used, and combining different models to specify their use and thus support a particular application, function, community, and environment [24, 62, 39].

Even knowing the benefits of a flexible description model, deposit/publication platforms do not support the functionality of including other types of elements other than text. In fact, there are

several hypotheses that can be implemented in the most diverse platforms, such as the possibility of uploading images stored on personal computers or even direct searches on the web. However, none of them have been applied.

One of the great opportunities for images is related to the use of current technologies for capture and dissemination. Currently, a simple mobile phone allows you to capture an image in a few seconds and store or disseminate it quickly. The data can, therefore, be easily reached, without the need to invest in large means of collection, unless the quality of the images is a key point. But even in this, we find an opportunity: the exponential growth of tools for image production is notorious, with a high impact on quality and editing. In addition, we are increasingly finding programs that allow us to analyze images in detail, in different formats, and of different types.

A major challenge of this new perspective for RDM is the concrete understanding of the various roles that images can play. If, on the one hand, we have images as data, generated, obtained, or used during the course of the research project and that serves as a tool that supports scientific evidence, on the other hand, we have images as metadata that provide a better interpretation of the data, attributing context. In fact, we believe there is another challenge associated with this one: images can often take on both roles. That is, at a given moment they can be seen as research data, fundamental elements in the collection of facts, but at another, they may not have that characteristic of data, but of metadata, insofar as it provides that other data are properly described and enabled to greater detail. And this is the beauty of the images because, after great potential, we managed to verify another. Its characteristics allow its use to be extended to several dimensions, in order to obtain the best results.

As previously mentioned, one of the advantages associated with the image is its ability to represent and prove reality. This fact is unequivocal. However, it should be noted that with the new artificial intelligence techniques, namely deep fake, this feature can be rethought. We believe that this does not compromise this characteristic of the images, however, it is always necessary to consider issues related to editing and image quality, in order to guarantee the integrity of the results.

5.1.6 Final remarks

It is always necessary to remember the potential of images, as a result of their specific characteristics. This does not mean that they should be overvalued in relation to other resources. The text, the sounds, the figures, among others, have particularities that make them unique in certain contexts. The image, due to its unique capacity of faithful representation and the impact it has on human cognition, awakens to countless challenges and opportunities that can be a valuable tool for researchers who want to see their projects and, consequently, their data, properly managed.

We consider it very important to reinforce the idea that this new perspective is not intended to replace the text or other resources, but rather to complement their use, through the unique characteristics of the images. However, the producers of the images that will be used as metadata, must be aware that they must fulfill their purpose - properly describe - and, therefore, must guarantee

their quality, that they do not cause interpretation doubts, and that they are always accompanied by text that is necessary.

In Section 5.2, will present a work where several researchers, from different research domains, are confronted with this new perspective. This case study corroborates the idea presented in this work, showing a practical side of this perspective that we believe is a great challenge, but also a giant opportunity for the scope of description in RDM.

5.2 Images to describe research data: a case study on the use of imagery metadata

The description is one of the most important phases of the RDM process. Taking this idea into consideration, we used a qualitative method of investigation through an exploratory case study where we conducted 16 semi-structured interviews, with researchers who have produced, described, and published research data, to understand how images can assume the role of metadata in data description.

5.2.1 Contextualization

In the description phase, the advantages of data description, as well as assisting tools are indicated. Metadata models are fundamental pieces for researchers who want to see their data properly described. This description is oriented towards using text as the metadata value. However, it is believed that other types of data can be used to complement the text.

This work follows on from the new perspective presented in Section 5.1, in order to validate the hypotheses raised about the use of images as metadata. This exploratory case study puts into perspective how images can support research as metadata, contributing to the description and interpretation of data.

5.2.2 Methodology

With this work, we raise the possibility of describing research data through images and assess how researchers assess this possibility. In this case, the images would not function as the element itself, but rather as the value assigned to it. We used a qualitative method of investigation through an exploratory case study [138]. We opted for the semi-structured interview since the methodological instrument selected for this approach was the semi-structured interview. To ask a group of participants a set of pre-defined questions, while still being adaptable to the course of the conversation and allowing for deviations from what was previously established. This method of collecting information makes it possible to obtain answers with more specificity and direct the conversation when it moves away from the central topic.

With this methodology, the aim was to reflect on it and improve its understanding. The use of images as metadata for innovative and exploratory research, through a case study, allows the

generation of new ideas and hypotheses that can trigger new dynamics in the field of research data management.

We recruited 16 researchers from different fields for convenience. To select the participants, we contacted 26 researchers who had deposited their datasets in the research data repository of the authors' research institution. Contact was made via email, explaining the study and its aim. Information was also given about the estimated time of the interview, which could be conducted in person or remotely, and that an informed consent form would be drawn up which, among other details, ensured the anonymization of the data. Of these 26 contacts, 16 responded positively to the interview invitation. After a positive response, the interview was scheduled. The informed consent form was then drawn up and sent. All the sessions were recorded using sound and images. All the data was anonymized and the recordings will be deleted at the end of the study. The interviews took place between November 2022 and January 2023.

The interview consisted of eight open questions. The first three questions addressed the topic of images as metadata from a generalized perspective. The next three questions addressed issues related to the deposits and descriptions of each of the participants, to understand the applicability of the topic to specific data sets. The last two questions reflect on the role of images in the research data collection process, particularly in the description phase. The interview script can be viewed in full in the Appendix B. A dataset was developed that includes the interview guide and the summary of the interview transcripts. This dataset is available in the INESC TEC Data Repository [113].

To support the interview, a mockup (Figure 5.3) was created of the data repository where the participants have their data published, with a simulated example of a data deposit in which images were used as metadata. This mockup was shown to the participants to help them get a concrete idea of what they were studying.

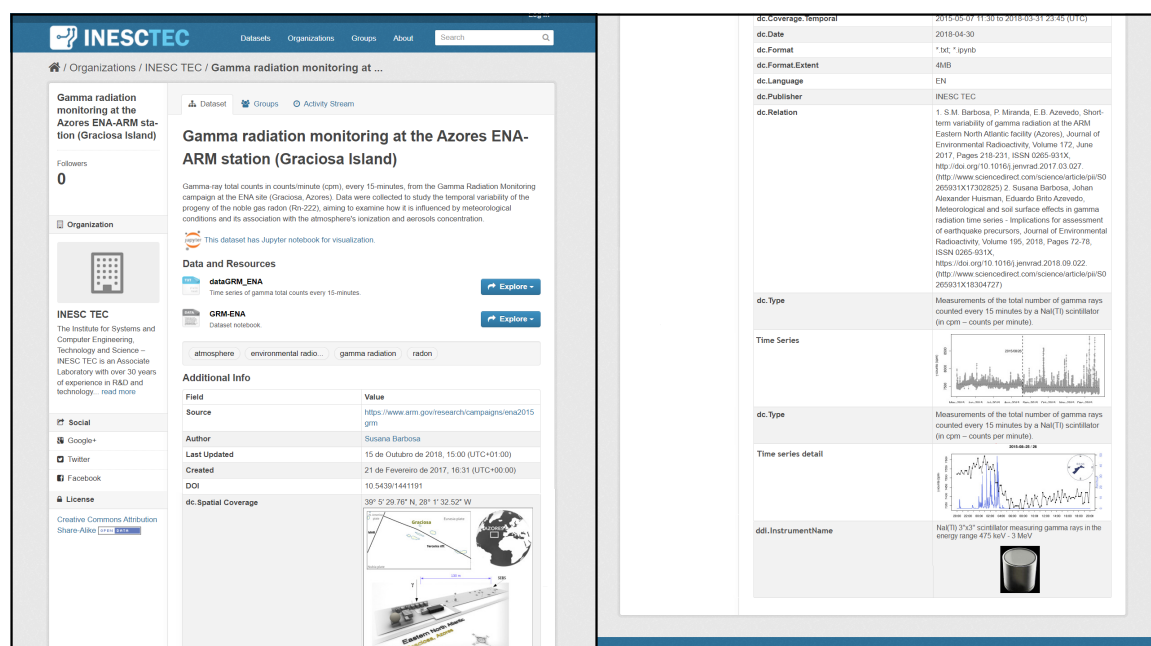


Figure 5.3: Mockup of the data repository

Table 5.1 shows the profile of the 16 interviewees. To maintain the confidentiality of participants, all information about them will not be disclosed.

Table 5.1: Profile of interviewees

Researcher	Specific Work Area	Qualifications	Position in the Institution
R1	Computing	PhD	Senior Researcher
R2	Computing	PhD	Auxiliary Researcher
R3	Electrical Engineering	PhD	Researcher
R4	Intelligent Systems	PhD	Senior Researcher
R5	Bioengineering	PhD	Auxiliary Research
R6	Industrial Robotics	PhD	Research Assistant
R7	Industrial Engineering	PhD	Senior Researcher
R8	Energy Systems	PhD	Researcher coordinator
R9	Energy Systems	PhD	Auxiliary Researcher
R10	Energy Systems	PhD	Senior Researcher
R11	Chemistry	PhD	Researcher
R12	Chemistry	PhD	Researcher
R13	Multimedia	PhD	Researcher
R14	Urban mobility	PhD	Researcher
R15	Information Science	PhD Student	Research Assistant
R16	Information Science	PhD Student	Research Assistant

5.2.3 Results

In this section, the results obtained in the context of the interviews were summarized and organized by the themes of the questions that make up the interview guide.

5.2.3.1 Images as metadata? Yes or no?

When confronted with the question about a scenario of describing data with images, ten of the respondents stated that they had never thought about this possibility. However, they all say that it makes sense because "the descriptions would become richer and more complete". Some of the researchers even say that they never thought about this possibility, although they think "it makes sense" (R2). One of these ten interviewees stated that he always thought that "the inclusion of images for the dissemination of the studies would be interesting" (R5), however "never particularly in the description of the data"(R5).

The remaining six interviewees said they had already thought about the possibility of data description with images. Four of them reinforce that this inclusion, although it can happen at any time, makes even more sense at the time of deposit and sharing because they see the description as a task that "helps others more than themselves" (R13). Most of the researchers regret the fact that deposit platforms do not encourage the inclusion of images as metadata. One of the researchers also stated that "images are needed to better visualize the data. Especially for series, maps, graphs" (R1).

5.2.3.2 Third-party research data described with images

The expectations of researchers who seek data from others to develop their own work or to study a particular topic can vary. In this section, we address the possibility and willingness of the interviewees to see third-party data described with images.

Advantages

All respondents said that they would like to see research data from other researchers depicted with images. Without exception, all researchers consider that images as metadata allow an overview of the data. They believe that it will not be possible to have a full understanding of the data in a first glance, but rather to obtain contextual information, which facilitates and accelerates the process of data perception. One of the researchers believes that with the use of images, it will be easier to understand certain elements, which would make the data more explicit and understandable. In addition, the images facilitate access to more information regarding the dataset, giving the charts as an example of an image for this situation.

Other researcher characterized most repositories as “raw deposit spaces, with the bare necessities” (R8). This interviewee thinks that the images can play a decisive role here due to their “more aesthetic than interpretative nature” (R8), but he still sees this as a great advantage, as they can be very useful instruments to appeal to the data and offer initial information.

One of the interviewees reinforced the idea that images, unlike text, help in a faster perspective of data, as they are “more efficient at transmitting information” (R15). Although the images do not allow a complete description, they make it possible to highlight and summarize specific moments of the project and data.

One researcher made it clear that images as metadata are useful, however, this utility is not related to the search/retrieval of data, as the repositories are not prepared for this functionality. For this researcher, more than for any other reason, images are important to understand whether that particular data is of interest for a given purpose, help in decision-making and make it possible that in certain cases it is not necessary to download the data. Basically, the images would work as a “prior analysis mechanism” (R3). Two of the interviewees focused on the idea of knowledge graphs, as image tools for understanding data. In addition, this type of images as metadata would help, not only in the description but also in the organization of information about the data. For these researchers, the visual impact on the description of datasets is fundamental.

Disadvantages

It was necessary to find out whether the interviewees would like to see other researchers’ data described with images. About this, not many disadvantages were identified during the interviews. Only 4 of those interviewed were able to list disadvantages. The first has to do with the possible poor quality of the images, namely about size and format, as they believe that the images could be small or in a format not supported by the user and, consequently, become imperceptible. Another disadvantage is related to the potential neglect of text if researchers start to use images. The third

disadvantage mentioned was the fact that the deposit platforms do not have a component/option for the inclusion of images, in fact, this possibility is not even given in the deposit platform guidelines, which limits this possibility from the outset. Finally, the work and time required to produce the images to be included as metadata was identified as a disadvantage. The researcher who mentioned this fourth disadvantage said that “the images should be the most prominent factor and the textual description should be seen as an additional description. The only problem with the images is that you have to make them and not all the data have images to associate. As not all researchers have time to do them. (R10)”.

5.2.3.3 Description of research data in publishing platforms

All respondents answered that they would use images in describing their data if the publishing platforms allowed it. Many of the arguments presented in this question coincide with what was said in question 4.2.1. Advantages. In this context, it would be possible to put on the platform more details about the data. These details, by their visual characteristics, would complement what is textually included and give more context.

One of the interviewees said that one of the main reasons for using images was to help replicate the study, as images are more attractive than text. Also, images cause a sense of truth because the information in the text, or in the results of the dataset, can be proven (“with the images, people will be able to verify that this is indeed the case” (R16)). Another researcher also went along with this idea. He says he would use them as long as it was possible to create sufficiently descriptive images. One of the advantages would be that eventually there would be more researchers interested in opening the shared data and thus replicating the study.

Two of the interviewees pointed advantages regarding visualization. One of them gave an example he appreciates: the AI Assets (assets catalog where can browse, search and download all the AI assets currently indexed in the AI4EU AI-on-demand platform, including AI libraries, datasets, containers, and more), which has a simple list with illustrative images, to access the resource it is necessary to enter the link. In this case, the images would serve as a “gateway to the data”. He also gave the example of IEEE Access (multidisciplinary, online-only, gold fully open access journal, continuously presenting the results of original research or development across all IEEE fields of interest) to talk about what he considers to be the ideal scenario: the use of graphical abstracts. He thinks that repositories should provide: the title of the dataset, the graphical abstract (which could contain images of different types), and the project name. If users wanted to know more about the dataset, they should access its full content. According to this researcher, “images are very important for dissemination (...) for interpretation the text is fundamental” (R13).

The other researcher also addresses the issue of visualization, stating that “images would be the preview tool” (R14), eliminating the cases in which it is necessary to download data because the textual metadata is too basic and universal and does not allow a clear understanding of the dataset. This researcher believes that the images could be as valuable as some keywords, so it would not be necessary to use several terms, many of them for similar expressions. Two respondents pointed out that while they would like to use this method of description with images, it is necessary to

ensure that the platforms support the images you want to include, regardless of format or size that these can be had.

5.2.3.4 Scenarios where images would be useful as metadata

As for specific scenarios where images could take on the role of metadata, the answers were diverse. Seven interviewees said they had encountered description scenarios that fell short of expectations. However, only two of them identified that this problem was due to the lack of images as metadata. One said that the use of "too much text in a description, in some cases, makes it difficult to understand all the important information of the dataset" (R4) and the other that, at times when this happened, he realized that "the lack of complementarity of types of description, particularly images, may not help in sharing data" (R6). Of the remaining respondents, five said that they were not able to identify some scenario where the data description fell short because they could not use other values for description besides text. However, all respondents stated that the interview allowed them to reflect on the subject, as they believe that if they had already been confronted with this problem, at the time of the interview these scenarios could be more easily identified. The remaining four interviewees also said that they were not able to identify scenarios where images would be useful as metadata, but because for them, the moment of description boils down to "always following the indications that the platforms indicated", without questioning other hypotheses (R11).

5.2.4 Inclusion of images in the description process

The following three topics of analysis refer to three questions asked to respondents about the datasets they deposited and described in a data repository. This moment was important for establishing a real link between the theme of the interviews (images as metadata to describe data) and concrete cases, well known to the interviewees.

Existing elements where images could complement the textual description

All respondents answered that in various circumstances images could be used to complement the description. Twelve of the interviewees said that it is difficult to identify one or several particular elements because "what may make sense in the context of one dataset may not make sense in the context of another". All stated that although they work within one research domain, the data can vary greatly depending on the project, the method of data collection, and the treatment of the data. All twelve, without exception, clearly state that images can play a key role in supplementing description but believe that their positioning in the metadata model may vary. Still, respondents identified the elements seen in Table 5.2.

Table 5.2: Existing elements where it would be possible to include images

Element	Justification
Description	"Would place an image referring to the structure and architecture of the data model that is being described. This would make more explicit the way the model is organized and the links between the different parts of the model." (R16)
Daily Load/ Flexibility Profiles	"You could have an illustrative image of the profiles. Since it includes many time series (for several nodes), it would not be possible to simply illustrate all the content. Also included in these is a description of electricity distribution or transmission networks that although they could be translated into images of network diagrams, are rarely available." (R6)
Spatial Coverage	"A clear identification of a place through, for example, a georeferenced image, is much better to interpret. Besides, it's faster, because we don't have to go and look for the coordinates" (R1)
Methodology	"We could eventually using photos and videos of the moments of data collection and production would certainly make it easier to interpret, to report the reader to that moment." (R15)
Instrument	"There are many instruments that have small variations and, in several cases, it is difficult to report these small differences. A photo of the instrument in question helped clarify any doubts." (R13)

Existing elements where images could replace the textual description

For this issue, fifteen of the sixteen interviewees stated that it is not possible to replace textual elements by images. Everyone said that the image can be very important in complementing the textual description, but not completely replacing it. There were several arguments pointed out by each of the researchers. One of the interviewees said: "what for me is the interpretation of an image for another may not be and that is where the text helps to clarify" (R10), for another interviewee "if there are only images, descriptive elements are lost that only the text provides" (R14) and for yet another interviewee "complete replacement by images would always remove key information that researchers need to know when looking for data" (R1). Only one researcher pointed out a scenario in which it was only possible to have images as a metadata value, namely when the images are graphs. This researcher gave an example of metadata he often uses: time series. The interviewee believes that, in these cases, graphs are "better tools for analysis and interpretation than the text" (R4).

New elements where images could be included

It was noticeable that researchers who deposit data tend to follow a formal pattern of description, not giving much thought to alternatives, although they think there is room for improvement. For this reason, the interviewees always have doubts when they are asked to propose alternatives because they do not have a deep knowledge of data description guidelines. Still, some of the interviewees, revisiting the dataset they have deposited, were able to identify new elements where it would be possible to include images. Those elements and their analysis can be seen in Table 5.3.

Table 5.3: New elements where it would be possible to include images

Element	Number of times mentioned	Justification
Location	7	Emerged mainly to identify concrete data collection locations, through images. (R1, R3, R4, R7, R9, R12, R13)
Context	5	Researchers consider that often the description or the abstract is too brief and focuses on the project in general and not on the data in particular. (R2, R3, R5, R14, R16)
Equipment/ Instrument/ Instrumentation	4	Researchers consider that this element (mentioned in different ways but with the same meaning) would be useful with images. Two of the respondents say that it should be mandatory to include an image whenever the methodological process includes the production of the data. (R6, R7, R13, R16)
Example	4	Researchers consider that an image can illustrate the data and give an example of what can be found in the dataset. (R2, R8, R10, R15)
Methodology	3	Researchers consider that a good understanding of the methodological process helps to understand the data even better. For the researchers, the methodology, besides textually, can be described with photos or videos. (R13, R15, R16)
Profile	1	The researcher who spoke about the profile works in the field of energy and, for him, images can help in the description of specific profiles. (R6)
Pathology	1	For the researcher who talked about pathology, images can be relevant when it comes to medical data. According to this interviewee often the data reports a disease or condition. In these cases, an image in the dataset description could redirect directly to the location of the pathology by signaling in the image. (R11)

5.2.4.1 Description time with images as metadata

Confronted with the possibility of time being a factor to be taken into account when using images as metadata, only two researchers stated they considered that a description model with images

would not have an impact on time expenditure. For these, from the moment that "if it were a well-understood and correctly applied methodology, time would not be an issue" (R12).

The remaining fourteen interviewees said that the description with images would probably involve a greater expenditure of time, but all, without exception, say that this is not an obstacle or problem. One of the interviewees stated that "by using images and uploading them to the research data description platform, the description process could be more time-consuming, namely with regard to the selection of images referring to each one of the elements. However, this time would make a difference in the specification detail of the different research data, making them more explicit and easier to understand" (R16). Another researcher justifies this greater expenditure of time with the creation of images. For this interviewee, "the description with text and images (or even just images) would imply spending more time in the process. This happens because the specificity of the shared data does not favor the immediate creation of images that summarize well what the data translates" (R16).

One of the interviewees stated that, although he does not consider the extra time a problem, he does not know the opinion of other researchers. Therefore, he believes that this description model with images "should be optional, because it really can take more time and lead to researchers stopping publishing the data. But we think it is important to have this functionality. If it starts to become the norm, soon no one will think about whether it takes longer or not, they will do it because it is the right practice" (R9).

The role of technological systems in this process of description was also discussed by one researcher. For this, "if the systems were optimized for these dynamics, it could even be faster because there would already be a defined process for this task" (R7).

5.2.5 Discussion

During the interviews, we found that researchers are sensitive to the idea of using images as metadata. All were able to identify the advantages of including images in the data description process. However, it was clear that, for the majority, this idea was new.

In fact, many of the researchers who use and produce data are not aware of all the RDM guidance as well as the opportunities that exist. Many follow institutional procedures or technical guidelines to ensure they comply with some data management good practices. This shows the importance of having a professional provide guidance for proper and effective research data management. Not only showing the procedures to follow but also opening new possibilities for discussion, with new theses, new panoramas, and new perspectives.

One of the most discussed points was the alignment of publishing platforms on the dynamics of including images as metadata. The description process must occur in a preliminary phase of the project to properly document the procedures and work phases. In these cases, the description occurs internally such as laboratory notebooks or personal storage platforms. However, it is clear that researchers' greatest concern with regard to description occurs in the data-sharing phase. It is at this point that greater importance is given to properly described data so that they can be duly recovered, interpreted, and reused. This is where deposit platforms, like repositories, play a key role.

Even if the data are not open access, the metadata must be, enabling the dissemination of results. Therefore, publishing platforms should allow the inclusion of images in fields assigned to metadata values. This does not happen, which prevents this description option. In a paper published in 2022, a comparative analysis was made of the most widely used publishing platforms [109]. One of the evaluation components concerned description and, as described in the article, it was found that this scenario, identified as necessary, does not exist.

The limitations of the platforms do not end there: it is necessary to ensure that in the same way that we can search for text, we can also search for images. Once again, this does not happen. Therefore, it is necessary to feed this theme, showing its potential, in order to promote public discussion.

5.2.6 Final remarks

It became clear that the topic of images as metadata is not widely discussed in the research community. However, most see it as a major challenge, opportunity, and advantage for research data management. It is necessary to make room for new conceptions and practices. RDM does not include any procedure that limits the researcher's field of action to the point of being limited to the guidelines mostly followed by the scientific community. Images pose a huge challenge. They have specific characteristics that can be useful in data management, namely their ability to faithfully portray a reality. In 2016, with the dynamization of the FAIR principles, new ideas were discussed that could act as a guide for the management of research data. Like other types of resources, images can be a part of these principles, whether in the data organization process or, above all in the implementation of guidelines for accessing research results, fostering the entire Open Science ecosystem.

Chapter 6

Conclusion

The work carried out has made it possible to study and understand in depth the role that images play in the scientific context. In fact, one of the great challenges of the work has been to raise awareness of the importance of image resources, without devaluing any other type.

Images are present in everyone's daily life. Easily captured or reused, seen as mediators of communication, and necessary in many contexts. However, their positioning in research is dubious. Was it possible to determine a clear scenario for the use and production of images? Can they somehow promote the research process? All of this was verified. In the same way that each of us uses a photograph, a video, or a drawing to represent and highlight a moment, this also happens in scientific contexts.

In addition to the various aspects of research data management, namely the roles of data and metadata, this work has allowed us to delve into the various dimensions of images. From the types, which can be extremely varied, such as microscopic images, medical images, maps, photographs, engravings, among many others (all of which can appear in a research scenario), to their formats, collection scenarios, editing possibilities, descriptive capacity, processing and analysis methods, among others.

This work has made it possible to explore challenges and point out trends and opportunities. Research data management in itself has a huge range of potential, as does imaging. Combining them has opened up many possibilities. All the stages of this work point to very concrete results. From analysing the habits and behaviors of researchers who use and produce images to exploring the role of images as research data, to the new perspective presented, in which images are seen as a description tool, assuming the position of metadata.

It's important to make a note of the heterogeneity of scientific practices that the term "science" makes opaque. Science is not unitary, it is permeated by differentiated epistemic practices, even when we divide it up and enclose it in large boxes like the Social Sciences and Humanities, for example. This discussion is not new and has already been promoted by authors such as Karin Knorr Cetina in "Epistemic cultures: How the Sciences Make Knowledge" [27] and also addressed by Blaise Cronin, a big name in Information Science, in "The Hand of Science: Academic Writing and Its Rewards" [33].

It is possible to say that the three periods highlighted in the methodology (Section 1.1) have been fulfilled, guaranteeing each of the proposed objectives. Although this work has been done through the COVID-19 period, this has not prevented the interaction component with researchers from taking place as planned, even using different contact methods.

One of the conditions seen as important in consolidating this work would be validation by the scientific community. This objective has also been achieved, insofar as all the components of the work have resulted in publications and communications with the national and international scientific community, guaranteeing their validation by peers, making it possible to discuss them, and opening doors for further studies.

This work has also undergone a phase of change when it comes to data privacy and ethics. With the arrival of the GDPR in 2018, countries are no longer bound by their own directives but are obliged to follow international regulations, with their national commissions as a means of verifying and ensuring compliance. Research must also be aware of all the requirements at this level, including being alert to the challenge of personal data, sensitive data, ethics, and intellectual property. Here too, images play an important role, as they have characteristics that make them susceptible to breaches of privacy, such as identification (of people, places, opinions, religious, political, and sexual positions, among others) or even data transparency, through the countless possibilities for editing and altering components in images. Throughout this work, this issue has always been taken into account and often explored.

We also would like to conclude by reflecting on human dynamics versus technological dynamics. Nowadays, the possibilities for automatic processing are diverse. Technologies and the evolution of artificial intelligence tools have opened doors in many ways and images are part of the challenges that have arisen from this. We don't believe that human dynamics should be in opposition to technological dynamics. Both are essential and complementary. Neither takes the place of the other. That's why we often talk about the role of the data curator in the management process, as an intervening and active agent.

Technology is wonderful, but it doesn't live on its own, which is why this work fits in the Social Science field. As Jim Collins said in his book "Good to Great" [29], "No technology, no matter how amazing (...) can by itself ignite a shift from good to great (...) No technology can turn the wrong people into the right people. No technology can instill the discipline to confront brutal facts of reality (...) No technology can supplant the need for deep understanding (...) Those that do not (...) will spiral downward or remain mediocre. This is the big-picture difference between great and good (...)".

6.1 Informational behavior of researchers using and producing images

The truth is that all this work of discovering and analyzing the images, from their most conceptual, and historical basis, allowed us to create the pillars that supported all the work that followed. Reaching the current level would be difficult without understanding the foundation. It was this

exploration that made it possible to validate some of the decisions made, as it allowed us to check whether that choice was the ideal one given the conditions that make up the images. The same happened with the objective but in-depth analysis of research data management. This is an area of great interest to the scientific community, as it has fostered the inclusion of consistent practices in data management, helping professionals, favoring institutions, enhancing projects, and leveraging methods.

Therefore, the chapter on the information behavior of researchers who use and produce images was essential because, on the one hand, it made it possible to determine a more concrete scenario for the defined topic, demystifying what these researchers' habits are in general, and on the other hand, it made it possible to understand the greatest needs, what the patterns are and thus facilitate the decisions that followed. Ultimately, this chapter was decisive, as it worked as an ally in helping us decide which tasks to carry out. Without this foundation, the following experiences would certainly have been more limited. In this way, it was possible to design the experiments based on previously acquired knowledge, guaranteeing compliance with what had been verified. This is why this phase was called "exploration".

It was clear that researchers do not have standardized practices or behaviors when it comes to scientific images. In the first instance, an attempt was made to see if the domains of research could bring about considerable differences, but this was not the case. Of course, different domains bring different particularities, but nothing substantial enough to point to when it comes to information behavior.

Although these behaviors don't exist in a consistent and standardized way, it was found that researchers do have management habits and tendencies, but they are not always documented or taken as a procedure for the projects themselves. In the experiments carried out in this chapter, it was noted that this happens many times. However, this only became clearer when talking to researchers or exploring in detail the practices inherent in their research processes.

Researchers are open to the standardization of practices, as well as aware of their importance for the projects they carry out. The obstacle is often time, which tends to make them prioritize tasks, leaving the establishment of these practices in the background. Awareness needs to be raised so that research data management and all the procedures relating to it are applied to research contexts, in particular recognizing the role that this management can play when it comes to images.

6.1.1 RQ1. How does the production and use of images occur in science?

This research question is aligned with Chapter 3, where the dynamics of the behaviors and habits of researchers who use and produce images in a research context are analyzed.

It is possible to say that this question has been answered. The various experiments carried out, from the initial study of one project (Generation.Mobi) to the literature review that demonstrates what the literature tells us about these behaviors, through to the questionnaires and interviews, reveal several indications about how the production and use of images occur in science. Although we have concluded that there are few standardized procedures within research groups or centers,

the practices do exist. This is why it was essential to organize the results by the various phases of the life cycle, as it allowed us to analyze each phase individually and explore its details.

In exploring this research question, it also became apparent, particularly through the questionnaires and interviews, where more direct contact was made with the researchers, that they are open to the idea of including images in their research processes. Furthermore, it was clear that they use and produce images frequently. This is largely due to the numerous possibilities for producing and capturing images, such as mobile devices (cell phones, tablets, cameras, etc.), which are increasingly becoming a trend. What's more, these devices themselves make it possible to share an image or even store it quickly and efficiently in a matter of seconds.

It was possible to conclude that there is no specific pattern per research domain. In fact, the studies carried out were aimed at checking whether the research domains involved considerable variations. It became clear that there are domains that focus more on one type of image than others, for example, microscopic images are more related to the life sciences, while drawings are much more represented in the humanities. Just as there are research domains that tend to have a more qualitative analysis and others a more quantitative analysis of images. Domains related to the exact sciences lean more toward quantitative analysis, while domains within the social sciences lean more toward qualitative analysis. However, these variations do not have a direct impact on the definition of practices and guidelines for image management in science. For this reason, the contributions of this work are very much characterized by a general application component, such as the metadata model for describing data in image format and the respective controlled vocabularies. We have seen that the information behavior of those who use and produce images is not related to their research domain, but rather to the tools, techniques, and concepts they have about procedures for research data management and how they can include images in these procedures.

6.2 Images as data

The chapter on images as data proved to be challenging, due to the need to explore the results obtained in the previous study, in which the behavior of researchers who use and produce images in an investigation context was studied.

One of the main challenges was determining the metadata model for describing data in image format. The initial question was whether several or just one metadata model was needed. Since there are particularities between research domains, it was important to verify the importance of more than one model. The studies carried out led us to believe that only one metadata model would be necessary. What was more important than distinguishing between domains was clearly identifying which elements should be in this model, namely analyzing in detail those that allow us to describe specific characteristics of the images, such as editions, quality, and scale, for example. It was very important to determine a first version of the model, so that, together with researchers, we could arrive at the final version of the model, as close as possible to the needs identified.

Another challenge was the controlled vocabularies developed for the metadata model. In addition to all the advantages recognized by controlled vocabularies, they make description easier and easier by allowing you to select one or more values for a element. We believe that these controlled vocabularies enrich the metadata model, as not all possible values are always remembered at the time of description. However, the large number of elements in each of the controlled vocabularies meant that they had to be selected carefully, rigorously, and in detail to cover as much as possible.

Also in this chapter, we would like to highlight the importance of studies on automatic image description and data sharing and storage solutions. These have made it possible to reflect on ways of managing research data where images, as data, are essential and decisive. It is extremely important to align the topic under study with current trends, whether they relate to the automation of description or the possibilities and solutions for data sharing and storage, such as repositories.

Finally, to reflect on the guidelines for standardizing practices in the use and production of images in research. At the time this phase of the work was planned, the expectation was that these guidelines would be clearly indicative of images, making it possible to show their particularities in the RDM context. However, in the course of working on the guidelines, it became clear that not all phases of the data management lifecycle involve characteristics that are unique and specific to images. It became clear that there are times when this particularity exists, however, there are several times when the guidelines for images overlap with the guidelines for data management in general. This fact in no way detracts from the aim of this section of the work, quite the opposite. In other words, it has made it clear where we can identify the differences between image-type data and the rest.

6.2.1 RQ2. How can RDM practices be adapted to the case of images as research data?

The answer to this research question was explored in Chapter 4, where the topic of images as data was addressed.

It became clear that images can be excellent allies for researchers in the research process. We have the assumption that, on their own, images can support research, since, as data, they allow datasets to be enhanced, including more information, which allows both the data and the methodological process to be contextualized until the results are obtained. However, combined with other tools, such as metadata models and controlled vocabularies, images can further enhance research, as these tools allow the data to be properly contextualized, offering the possibility of better interpretation.

It was also very important to explore the issues related to automatic image processing, as it allowed us to understand how current tools can be an ally for researchers when it comes to describing data, when we are talking about images. This research has made it clear that the human component needs to be present, even though it is clear that automatic processing tools can support researchers by speeding up the process and helping with the description.

The analysis of data deposit and publication solutions was also useful in answering this research question. The step of publication/share is important for researchers who want to disseminate their data and see it properly cited and reused.

The guidelines presented at the end of this chapter have helped to support this research question, since the proposals therein address the issue of images as data, identifying what overlaps with other types of data and highlighting what images excel at.

6.3 Images as metadata

About the study of images as metadata, it is important to highlight once again what a great challenge this has proved to be. When we studied the literature on this subject, we realized that nothing corresponded exactly to the guiding question (images as metadata). The dynamics of metadata have been widely studied, particularly its fundamental role in the process of managing research data, especially in the description phase. The importance that metadata can assume and the multiple advantages it brings are clear. However, the values assigned to metadata are overwhelmingly textual or numerical, limiting the space for other types. The knowledge that has been acquired about images and their potential has been decisive in conceiving the new hypothesis of using images as metadata.

Based on the need to explore new ideas regarding the use of images in data description, it was essential to determine and clearly establish a new perspective that explored this scenario. This new perspective made it possible to present the standard idea and, through various pieces of evidence in the literature, to present the new perspective, always taking into account the challenges and opportunities within the field of RDM.

All the work carried out in this doctorate involved a lot of exploration and execution of tasks, always accompanied by the respective evaluation. This chapter also revealed this need for rigorous and detailed work. As this was a new perspective, it was necessary to be aware of and have in-depth knowledge of the new topic to be explored, to be prepared to align and carry out tasks that could be understood by others and properly oriented towards the intended objectives, and to be able to analyze the results obtained in detail.

6.3.1 RQ3. Can images improve the description and interpretation of research data through their use as metadata?

When this research question was drawn up, the aim was to determine a new scenario for describing data, based on a new perspective, in order to understand the phenomena of data description and interpretation in concrete terms. Aware of the characteristics and particularities of images, we wanted to explore how they can not only be an ally for researchers describing their data, but also for those who want to learn more about the data and use images to help interpret it.

The work in Chapter 5 explored exactly this. We believe that yes, images can improve the description and interpretation of research data, taking on the role of metadata. As stated in Chapter 5,

images can be excellent aids to description, complementing other descriptions (textual, for example), or even used on their own to describe (when their content is significant enough for description and subsequent interpretation). In addition, the researchers were open to the idea of using images as metadata, as they also clearly identified the potential that this type of document (image) has.

In the course of this research question, the importance of a more in-depth study of data deposit and publication platforms that do not yet have the necessary functionalities for images to be used as metadata was highlighted. As this is a new perspective, it is natural that this is not the case, but exploring this topic is very important as it will act as a driving force behind this new perspective.

6.4 Perspectives and future work

The work that has been carried out has allowed us to explore many issues, raise hypotheses, verify ideas, and envision future tasks.

First of all, we would like to highlight the issues of data sensitivity, ethics, and intellectual property. Although all resources suffer from the need for detailed attention to these issues, images have characteristics that make them vulnerable to some of these things. It is important that following on from what this work has looked at, the challenges that privacy and copyright issues bring are studied in more detail. As already mentioned, images can be a researcher's ally, however, it is necessary to ensure all legal compliance, namely the guidelines set out in the General Data Protection Regulation, as well as the legal principles of reusing third-party material. We can say that the same applies to resources other than images, however, it is necessary to explore what particularities they have that can distinguish them from other resources and therefore require more focused guidelines.

Next, we believe that the issues surrounding the automatic description of images in a research context should be explored in the future because, in addition to the notorious trends we are currently seeing, namely with the various existing tools, researchers tend to look for description aids that speed up and facilitate their work. Therefore, to take advantage of all the potential that these tools have, it is necessary to explore in more detail which are the most correct practices and procedures to apply.

Also seen as a prospect and future work is the more in-depth exploration of the various solutions for publishing and sharing data. In particular, about the new perspective on the use of images as metadata, it is extremely important to start analyzing in more detail what functionalities these solutions (as data repositories) should have to include images as values assigned to the description. In addition, it is important to study and standardize practices for using images as metadata, as we see happening with other types of documentation, such as text. In other words, clearly defining the description process that researchers must follow on these platforms, such as: the possibility of uploading images directly from the researchers' computer; uploading the image via a cloud service; directly searching for the image in a search engine; among others. In addition, it is necessary to analyze how the various aspects of images as metadata will work within these platforms, for

example: How will it work if the description is only as an image? How will it work if the image is accompanied by text?

Future work also includes raising awareness and training researchers to include images in research processes. As has already been said, images are present in day-to-day research projects, but rarely in a standardized way and following some kind of procedure or protocol. It was clear from the interaction with the various researchers in the different experiments that they are susceptible to these dynamics with images, so it is important to seize the opportunity and create mechanisms to raise awareness of formal practices for the use and production of images in science.

Finally, we would like to highlight the importance of promoting more research in the area of images in the context of RDM and enhancing communication and debate on this topic. This can happen through the development of more material for the scientific community, namely by feeding the literature with more studies on images as data and as metadata, always with a conscious alignment of RDM practices. There is also a need to participate in more discussions on the topic, such as colloquiums, conferences and forums, for example, where direct contact with third parties can encourage the creation of new ideas and raise awareness of other practices.

References

- [1] Fotografia científica: uma introdução, 2016. Available at: <https://www.fotografia-dg.com/fotografia-cientifica/>. Quoted on page 126.
- [2] Catharine Abell. The epistemic value of photographs. In Catharine Abell and Katerina Bantinaki, editors, *Philosophical Perspectives on Depiction*, pages 82–103. Oxford University Press, 2010. Available at: <https://doi.org/10.1093/acprof:oso/9780199585960.001.0001>. Quoted 2 times on pages 20 and 21.
- [3] T. Acharya and A.K. Ray. *Image Processing: Principles and Applications*. John Wiley Sons, Inc., New Jersey, 2005. ISBN: 978-0-471-71998-4. Quoted on page 123.
- [4] Tinku Acharya and Ajoy Ray. *Image Formation and Representation*. 09 2005. Available at: <https://doi.org/10.1002/0471745790.ch2>. ISBN: 9780471719984. Quoted on page 30.
- [5] 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. Available at: <https://doi.org/10.2218/ijdc.v8i2.263>. Quoted on page 82.
- [6] Mohd M. Ali N. M. Al Nasar, M. R. Factors in re-finding the personal photographs: Review and possible research directions. *Research Journal of Applied Sciences, Engineering and Technology*, 7(4):11–14, 2014. Available at: <https://doi.org/10.19026/rjaset.7.328>. Quoted on page 126.
- [7] Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha Silva, and Cristina Ribeiro. A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, 16(4):851–862, 2017. Available at: <https://doi.org/10.1007/s10209-016-0475-y>. Quoted 6 times on pages 30, 101, 102, 121, 125, and 126.
- [8] Chris Armbruster and Laurent Romary. Comparing repository types: Challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication. *SSRN Electronic Journal*, 2010. Available at: <http://dx.doi.org/10.2139/ssrn.1506905>. Quoted on page 102.
- [9] Alex Ball. Review of Data Management Lifecycle Models. Technical report, Bath, UK., 2012. Available at: <https://researchportal.bath.ac.uk/en/publications/review-of-data-management-lifecycle-models>. Quoted 5 times on pages 16, 18, 39, 51, and 65.

- [10] James Balm. The power of pictures. how we can use images to promote and communicate science. *Research in progress blog*, 2014. Available at: <https://acesse.dev/Rz5oh>. Quoted on page 123.
- [11] Marcus. Banks. *Using visual data in qualitative research*. SAGE Publications, London, 1st edition, 2007. ISBN: 9780857020260. Quoted on page 27.
- [12] R. Barthes. *Image Processing: Principles and Applications*. Edições 70, Lisbon, 2002. ISBN: 9789724413495. Quoted on page 122.
- [13] R. Barthes. *A câmara clara*. Edições 70, 2012. ISBN: 9789724413495. Quoted on page 20.
- [14] Gordon Bell. *Foreword*. Microsoft Research, 2009. Quoted on page 1.
- [15] Howard Besser. Visual access to visual images: The uc berkeley image database project. *Library Trends*, 38, 1990. Available at: <https://eric.ed.gov/?id=EJ420315>. Quoted on page 28.
- [16] National Science Board. Long-lived digital data collections: Enabling research and education in the 21st century. Technical report, National Science Boarde, 2005. Available at: <https://www.nsf.gov/pubs/2005/nsb0540/>. Quoted on page 12.
- [17] V. R. C. Boccato and M. S. L. Fujita. Discutindo a análise documental de fotografias: uma síntese bibliográfica. *Cadernos BAD*, (2), 2006. Available at: <https://www.brapci.inf.br/index.php/res/v/82351>. Quoted on page 20.
- [18] 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. Available at: <https://doi.org/10.1002/asi>. Quoted 2 times on pages 13 and 14.
- [19] Christine L. Borgman, Jillian C. Wallis, and Noel Enyedy. Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, 7(1-2):17–30, 2007. Available at: <https://doi.org/10.1007/s00799-007-0022-9>. Quoted 2 times on pages 1 and 14.
- [20] Kristin Briney. *Data Management for Researchers: Organize, Maintain and Share your Data for Research Success*. Pelagic Publishing, Exeter, UK, 2015. ISBN: 978-1-78427-030-8. Quoted 2 times on pages 1 and 23.
- [21] CASRAI. Data, March 2022. Available at: <https://casrai.org/term/data/>. Quoted 2 times on pages 12 and 21.
- [22] 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. Available at: https://doi.org/10.1007/978-3-319-24129-6_30. Quoted on page 23.
- [23] 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. Available at: <https://doi.org/10.4018/978-1-5225-2221-8.ch008>. Quoted 2 times on pages 11 and 82.

- [24] João Aguiar Castro, João Rocha da Silva, and Cristina 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*, pages 313–316, 2014. Available at: <https://ieeexplore.ieee.org/document/6970185>. Quoted 3 times on pages 24, 25, and 126.
- [25] 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. Available at: <https://doi.org/10.2218/ijdc.v15i1.705>. Quoted on page 82.
- [26] João Aguiar Castro, Joana Rodrigues, Paula Mena Matos, Célia Sales, and Cristina Ribeiro. Getting in touch with metadata: a ddi subset for fair metadata production in clinical psychology. *IASSIST Quarterly*, 47(1), 2023. Available at: <https://doi.org/10.29173/iq1008>. Quoted on page 7.
- [27] Karin Knorr Cetina. *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press, 1st edition, 1999. ISBN: 9780674258945. Quoted on page 139.
- [28] Hsin-liang Chen and Edie Rasmussen. Intellectual access to images. *Library trends*, 48, 1999. Available at: <https://www.learntechlib.org/p/88862/>. Quoted 2 times on pages 28 and 29.
- [29] Jim Collins. *Good To Great*. HarperCollins Publishers, 1st edition, 2001. ISBN: 9780712676090. Quoted on page 140.
- [30] Andrew Martin Cox and Winnie Wan Ting Tam. A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, 70(2):142–157, 2018. Available at: <https://doi.org/10.1108/AJIM-11-2017-0251>. Quoted 3 times on pages 11, 14, and 27.
- [31] Douglas Crome. Avoiding twisted pixels: Ethical guidelines for the appropriate use and manipulation of scientific digital images. *Science and engineering ethics*, 16:639–67, 12 2010. Available at: <https://doi.org/10.1007/s11948-010-9201-y>. Quoted on page 109.
- [32] Douglas Crome. Digital images are data: And should be treated as such. *Methods in molecular biology (Clifton, N.J.)*, 931:1–27, 01 2013. Available at: https://doi.org/10.1007/978-1-62703-056-4_1. Quoted on page 109.
- [33] Blaise Cronin. *The Hand of Science: Academic Writing and Its Rewards*. Scarecrow Press, 1st edition, 2005. ISBN: 0-8108-5282-9. Quoted on page 139.
- [34] M. Czerwinski and E. Horvitz. An investigation of memory for daily computing events. In *People and Computers XVI - Memorable Yet Invisible*, 2002. Available at: https://doi.org/10.1007/978-1-4471-0105-5_14. Quoted on page 126.
- [35] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. 40(2), 2008. Available at: <https://doi.org/10.1145/1348246.1348248>. Quoted on page 29.
- [36] Ritendra Datta, Jia Li, and James Wang. Content-based image retrieval—approaches and trends of the new age. volume 40, pages 253–262, 2005. Available at: <https://doi.org/10.1145/1101826.1101866>. Quoted on page 29.

- [37] M. Day. Dcc digital curation manual: Instalment on metadata. Technical report, Data Curation Center, 2005. Available at: <http://hdl.handle.net/1842/3321>. Quoted 2 times on pages 14 and 23.
- [38] Maria Dondero. The semiotics of scientific image. *The American Journal of Semiotics*, 25:1–19, 2009. Available at: <https://doi.org/10.5840/ajs2009253/411>. Quoted 2 times on pages 20 and 22.
- [39] Erik Duval and et al. Metadata principles and practicalities. *D-Lib Magazine*, 8(4), 2002. Available at: <https://www.dlib.org/dlib/april02/weibel/04weibel.html>. Quoted 2 times on pages 25 and 126.
- [40] Paul Edwards and et al. Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science*, (5):667–690, 2011. Available at: <http://www.jstor.org/stable/41301955>. Quoted on page 22.
- [41] Japan Electronics and Information Technology Industries Association. Exchangeable image file format for digital still cameras: Exif Version 2.2, 2002. Available at: <https://www.loc.gov/preservation/digital/formats/fdd/fdd000146.shtml>. Quoted 2 times on pages 26 and 27.
- [42] Amany M. Elsayed and Emad I. Saleh. Research data management and sharing among researchers in Arab universities: An exploratory study. *IFLA Journal*, 44(4):281–299, 2018. Available at: <https://doi.org/10.1177/0340035218785196>. Quoted 2 times on pages 48 and 49.
- [43] David Elswelle, I. Ruthven, and C. Jones. Dealing with fragmented recollection of context. *Department of Computer and Information Sciences, University of Strathclyde*, 2005. Available at: <https://strathprints.strath.ac.uk/2727/>. Quoted on page 126.
- [44] Peter Enser. Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. *Journal of Information Science*, 26(4):199–210, 2000. Available at: <https://doi.org/10.1177/016555150002600401>. Quoted 4 times on pages 19, 28, 29, and 39.
- [45] Peter Enser, Christine Sandom, Jonathon Hare, and Paul Lewis. Facing the reality of semantic image retrieval. *Journal of Documentation*, 63, 2007. Available at: <https://doi.org/10.1108/00220410710758977>. Quoted on page 29.
- [46] V. Van den Eynden and et al. *Managing and sharing data: A best practice guide for researchers*. 3rd edition, 2011. ISBN: 1-904059-78-3. Quoted 2 times on pages 14 and 23.
- [47] Diogo Faria. Análise e Processamento de Imagem. *Faculdade de Engenharia da Universidade do Porto*, 2010. Available at: <http://hdl.handle.net/10400.14/30624>. Quoted 3 times on pages 30, 123, and 124.
- [48] 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. Available at: https://doi.org/10.1007/978-3-030-54956-5_16. Quoted on page 8.

- [49] Miguels Fernandes, Joana Rodrigues, and Carla Teixeira Lopes. Research data management in image format - survey [dataset]. *INESC TEC research data repository*, 2020. Available at: <https://doi.org/10.25747/7ma9-9132>. Quoted 2 times on pages 9 and 41.
- [50] FORCE11. Main features of CERIF, 2019. Available at: <https://eurocris.org/services/main-features-cerif>. Quoted on page 26.
- [51] FORCE11. The FAIR Data Principles, 2019. Available at: <https://www.force11.org/group/fairgroup/fairprinciples>. Quoted 4 times on pages 13, 119, 121, and 125.
- [52] Filipe Furtado, Pedro Príncipe, and José Carvalho. Kit sobre dados de investigação. Technical report, RCAAP, 2017. Available at: <http://repositorium.sdum.uminho.pt/handle/1822/46351>. Quoted 2 times on pages 12 and 14.
- [53] F. Genova and et al. The data harvest: How sharing research data can yield knowledge, jobs and growth. Technical report, Research Data Alliance, 2014. Available at: <https://www.rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html>. Quoted 2 times on pages 1 and 14.
- [54] Rafael Gonzales and Richards Woods. *Digital Image Processing*. Prentice Hall, New Jersey, 2002. ISBN: 9781292223049. Quoted 3 times on pages 96, 123, and 124.
- [55] R. Gonzalez and R. Woods. *Digital image fundamentals*. 2 edition, 2000. ISBN: 0-201-18075-8. Quoted 3 times on pages 96, 123, and 124.
- [56] A. A. Goodrum. Image information retrieval: An overview of current research. *Information Science*, 3, 2001. Available at: <https://doi.org/10.28945/578>. Quoted on page 28.
- [57] A. S. Gordon. Browsing image collections with representations of common-sense activities. *Journal of the American Society for Information Science and Technology*, 52(11):925–929., 2001. Available at: <https://doi.org/10.1002/asi.1143>. Quoted on page 29.
- [58] Michael Gorman. Authority control in the context of bibliographic control in the electronic environment. *Cataloging Classification Quarterly*, 38:11–22, 10 2004. Available at: https://doi.org/10.1300/J104v38n03_03. Quoted on page 97.
- [59] David Guedj and Celina Ramjoué. European commission policy on open-access to scientific publications and research data in horizon 2020. *Biomedical Data Journal*, 01:11–14, 2015. Available at: <http://dx.doi.org/10.11610/bmdj.01102>. Quoted on page 102.
- [60] Patricia Harping. *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works*. Getty Publications, 2010. ISBN: 978-1-60606-026-1. Quoted on page 27.
- [61] Heather Hedden. DTaxonomies and Controlled Vocabularies Best Practices for Metadata. *Journal of Digital Asset Management*, 2010. Available at: <https://doi.org/10.1057/dam.2010.29>. Quoted 2 times on pages 23 and 27.
- [62] Rachel Heery and Manjula Patel. Application profiles: Mixing and matching metadata schemas. *Ariadne*, (25), 2000. Available at: <http://www.ariadne.ac.uk/issue/25/app-profiles/>. Quoted 2 times on pages 25 and 126.

- [63] Patrick Heidorn. Shedding light on the dark data in the long tail of science. *Library Trends*, 57:280–299, 2008. Available at: <https://doi.org/10.1353/lib.0.0036>. Quoted 3 times on pages 29, 102, and 125.
- [64] Sarah Higgins. The dcc curation lifecycle model. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2008. Available at: <https://doi.org/10.1145/1378889.1378998>. Quoted 2 times on pages 18 and 19.
- [65] Rosie Higman, Daniel Bangert, and Sarah Jones. Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights the UKSG journal*, 32, 2019. Available at: <https://insights.uksg.org/articles/10.1629/uksg.468>. Quoted 2 times on pages 1 and 13.
- [66] Dirk Huds. The impact of photography on society, 2017. Available at: <https://ourpastimes.com/the-impact-of-photography-on-society-12377030.html>. Quoted 3 times on pages 11, 30, and 123.
- [67] Ilyanaand Janis and et al. Using illustrations to make decisions on the most appropriate qualitative research methodology: The industry 4.0 scenario. *International Journal of Qualitative Methods*, 19:213–234, 2020. Available at: <https://doi.org/10.1177/1609406920907247>. Quoted on page 21.
- [68] Brigitte Jörg. Cerif: The common european research information format model. *Data Science Journal*, Sep 2010. Available at: <https://datascience.codata.org/articles/10.2481/dsj.CRIS4>. Quoted on page 26.
- [69] Hanna Kallio, Anna-Maija Pietilä, Martin Johnson, and Mari Kangasniemi. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, 72:954–2965, 2016. Available at: <https://doi.org/10.1111/jan.13031>. Quoted on page 51.
- [70] B.B. Bederson Kang, H. and B. Suh. Capture, annotate, browse, find, share: Novel interfaces for personal photo management. *International Journal of Human-Computer Interaction*, 23(2):315–337, 2007. Available at: <https://doi.org/10.1080/10447310701702618>. Quoted 2 times on pages 124 and 126.
- [71] Nair Yumiko Kobashi and Johanna Wilhelmina Smit. *Como elaborar vocabulário controlado para aplicação em arquivos*. Arquivo do Estado/Imprensa Oficial, 2003. ISBN: 85-86726-52-4. Quoted on page 27.
- [72] A.L. Lacerda. A fotografia nos arquivos: produção e sentido de documentos visuais. *História, Ciências, Saúde*, (1), 2012. Available at: <https://doi.org/10.1590/S0104-59702012000100015>. Quoted on page 124.
- [73] Carl Lagoze, Herbert Sompel, Michael Nelson, and Simeon Warner. The open archives initiative protocol for metadata harvesting. 2002. Available at: https://www.researchgate.net/publication/319395254_The_Open_Archives_Initiative_Protocol_for_Metadata_Harvesting. Quoted on page 102.
- [74] Sara Shatford Layne. Some Issues in the Indexing of Images. *Journal of the American Society for Information Science*, 45(8):583–588, 1994. Available at: [https://doi.org/10.1002/\(SICI\)1097-4571\(199409\)45:8<583::AID-ASI13>3.0.CO;2-N](https://doi.org/10.1002/(SICI)1097-4571(199409)45:8<583::AID-ASI13>3.0.CO;2-N). Quoted 3 times on pages 20, 28, and 29.

- [75] Alessandro Liberati and et al. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*, 339, 2009. Available at: <https://doi.org/10.1136/bmj.b2700>. Quoted on page 63.
- [76] Clifford A. Lynch. Institutional repositories: Essential infrastructure for scholarship in the digital age. *portal: Libraries and the Academy*, 3(2):327–336, 2003. Available at: <https://doi.org/10.1353/pla.2003.0039>. Quoted on page 102.
- [77] L. Lyon. Transparency: The emerging third dimension of open science and open data. *LIBER Quarterly*, 25(4):153–171, 2016. Available at: <https://doi.org/10.18352/lq.10113>. Quoted on page 14.
- [78] Alberto Manguel. *Lendo imagens: uma história de amor e ódio*. Companhia das Letras, 2001. ISBN: 978-85-359-0149-8. Quoted on page 98.
- [79] Krystyna Matusiak. Towards user-centered indexing in digital image collections. *OCLC Systems Services*, 22:283–298, 10 2006. Available at: <https://doi.org/10.1108/10650750610706998>. Quoted on page 29.
- [80] Liang Meiyu, Du Junping, Jia Yingmin, and Sun Zengqi. Image semantic description and automatic semantic annotation. In *ICCAS 2010*, pages 1192–1195, 2010. Available at: <https://ieeexplore.ieee.org/document/5669742>. Quoted 2 times on pages 97 and 98.
- [81] Maria Teresa Pinto Mendes and Maria da Graça Simões. Indexação por assuntos: princípios gerais e normas. *Gabinete de estudos ab*, 2002. Available at: <https://ojs.letras.up.pt/index.php/paginasab/article/view/3453>. Quoted on page 97.
- [82] Aaron Meskin and Jonathan Cohen. *Photographs as Evidence*. Blackwell, 2008. Quoted on page 20.
- [83] William K. Michener. Meta-information concepts for ecological data management. *Ecological Informatics*, 1(1):3–7, 2006. Available at: <https://doi.org/10.1016/j.ecoinf.2005.08.004>. Quoted 2 times on pages 23 and 24.
- [84] J. Mota and et al. Agrafo: A visual interface for grouping and browsing digital photos. In *Proceedings of the working conference on Advanced visual interfaces*, pages 315–337, 2008. Available at: <https://doi.org/10.1145/1385569.1385673>. Quoted on page 126.
- [85] Eadweard Muybridge. The horse in motion, 2021. Available at: <https://www.khanacademy.org/humanities/becoming-modern/early-photo/early-photo-france/a/eadweard-muybridge-the-horse-in-motion>. Quoted on page 30.
- [86] Elaine Ménard. Image indexing and retrieval: Challenges and new perspectives / indexation et repérage d’images : défis et nouvelles perspectives. *Canadian Journal of Information and Library Science*, 34:245–248, 2010. Available at: <https://doi.org/10.1353/ils.2010.0002>. Quoted on page 29.
- [87] N. Mößner. Photographic evidence and the problem of theory-ladenness. *J Gen Philos Sci*, 44:111–125, 2013. Available at: <https://doi.org/10.1007/s10838-013-9219-3>. Quoted on page 22.

- [88] Michael L. Nelson, Herbert Van de Sompel, and Simeon Warner. Advanced overview of version 2.0 of the open archives initiative protocol for metadata harvesting. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital libraries*, 2002. Available at: <https://doi.org/10.1145/544220.544367>. Quoted on page 102.
- [89] OECD. OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007. Available at: <https://11nq.com/CtUUh>. Quoted 2 times on pages 1 and 11.
- [90] OECD. Making Open Science a Reality., 2015. Available at: <http://dx.doi.org/10.1787/5jrs2f963zsl-en>. Quoted 3 times on pages 11, 12, and 13.
- [91] Journal of Biogeography. How to write a (great) Perspective article, 2021. Available at: <https://journalofbiogeographynews.org/2021/05/22/how-to-write-a-great-perspective-article/>. Quoted on page 120.
- [92] Consortium of European Social Sciences Data Archives. Adapt your data management plan: A list of data management questions based on the expert tour guide on data management, 2019. Available at: <https://dmeg.CESSDA.eu/Data-Management-Expert-Guide/1.-Plan/Adapt-your-DMP-Part-1>. Quoted on page 110.
- [93] C.L. Palmer and et al. Site-based data curation based on hot spring geobiology. *PLoS ONE*, 12(7), 2017. Available at: <https://doi.org/10.1371/journal.pone.0172090>. Quoted on page 125.
- [94] Irene V. Pasquetto, Bernadette M. Randles, and Christine L. Borgman. On the reuse of scientific data. *Data Science Journal*, Mar 2017. Quoted on page 13.
- [95] Maureen Pennock. Digital Curation: A Life-Cycle Approach to Managing and Preserving Usable Digital Information . *Library Archives*, 2007. Available at: https://www.researchgate.net/publication/228770335_Digital_curation_A_life-cycle_approach_to_managing_and_preserving_usable_digital_information. Quoted on page 14.
- [96] Molina Pinto and et al. *Indización y Resumen de Documentos Digitales y Multimedia*. 2 edition, 2002. ISBN: 84-9704-023-6. Quoted on page 97.
- [97] Graham Pryor. Why manage research data? In *Managing Research Data*, pages 1–16. Facet, 2018. Available at: <https://doi.org/10.29085/9781856048910.002>. Quoted on page 15.
- [98] Jian Qin and Kai Li. How portable are the metadata standards for scientific data? a proposal for a metadata infrastructure. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pages 25–34, 2013. Available at: <https://dcpapers.dublincore.org/pubs/article/view/3670>. Quoted on page 25.
- [99] Pauline Rafferty. Disrupting the metanarrative: A little history of image indexing and retrieval. *Knowledge Organization*, 46:4–14, 2019. Available at: <https://doi.org/10.5771/0943-7444-2019-1-4>. Quoted on page 28.
- [100] Cristina Ribeiro, Eloy Rodrigues, Eugénia Matos Fernandes, and Ricardo Saraiva. Os repositórios de dados científicos: estado da arte. Technical report, 2010. Available at: <https://repositorio-aberto.up.pt/handle/10216/23806>. Quoted 2 times on pages 1 and 14.

- [101] Fernanda Ribeiro. Indexação e Controlo de Autoridade em Arquivos. *Câmara Municipal do Porto*, 1996. Available at: <https://hdl.handle.net/10216/10721>. Quoted on page 97.
- [102] Robin Rice and J. Haywood. Research data management initiatives at university of edinburgh. *International Journal of Digital Curation*, 6(2), 2011. Available at: <https://doi.org/10.2218/ijdc.v6i2.199>. Quoted 2 times on pages 121 and 123.
- [103] Jenn Riley. Understanding Metadata: What Is Metadata, and What is it for? *NISO Primer: National Information Standards Organization (NISO)*, 2017. Available at: <https://doi.org/10.1017/S0003055403000534>. Quoted on page 23.
- [104] Joana Rodrigues. Tendências atuais e perspectivas futuras em organização do conhecimento: O controlo de autoridade de documentos fotográficos nos arquivos. *Atas do III Congresso ISKO Espanha-Portugal and XIII Congresso ISKO Espanha*, 2017. Available at: <https://dialnet.unirioja.es/servlet/articulo?codigo=6580725>. Quoted on page 97.
- [105] Joana Rodrigues, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. Hands-on data publishing with researchers: Five experiments with metadata in multiple domains. In Paolo Manghi, Leonardo Candela, and Gianmaria Silvello, editors, *Digital Libraries: Supporting Open Science*, pages 274–288, Cham, 2019. Available at: https://doi.org/10.1007/978-3-030-11226-4_22. Springer International Publishing. Quoted on page 8.
- [106] Joana Rodrigues and Carla T. Lopes. A utilização e tratamento de documentos fotográficos em projetos de investigação: o caso do generation.mobi. *Actas del IV Congreso ISKO España-Portugal 2019, XIV Congreso ISKO España*, pages 241–252, 2020. Available at: <https://dialnet.unirioja.es/servlet/articulo?codigo=7393137>. Quoted on page 8.
- [107] Joana Rodrigues and Carla Teixeira Lopes. Photography as data and metadata: description requirements for information retrieval and semantic interoperability. 15(2), 2019. Available at: <https://bulletin.jcdl.org/Bulletin/v15n2/>. Quoted on page 8.
- [108] Joana Rodrigues and Carla Teixeira Lopes. Research data management in the image lifecycle: A study of current behaviors. *RCIS: International Conference on Research Challenges in Information Science. Lecture Notes in Business Information Processing*, 446:213–234, 2022. Available at: https://doi.org/10.1007/978-3-031-05760-1_3. Quoted on page 8.
- [109] Joana Rodrigues and Carla Teixeira Lopes. Solutions for data sharing and storage: A comparative analysis of data repositories. *Linking Theory and Practice of Digital Libraries: 26th International Conference on Theory and Practice of Digital Libraries*, page 512–517, 2022. Available at: https://doi.org/10.1007/978-3-031-16802-4_55. Quoted 2 times on pages 8 and 137.
- [110] Joana Rodrigues and Carla Teixeira Lopes. Describing data in image format: Proposal of a metadata model and controlled vocabularies. *Journal of Library Metadata*, pages 213–234, 2022. Available at: <https://doi.org/10.1080/19386389.2022.2117511>. Quoted on page 7.

- [111] Joana Rodrigues and Carla Teixeira Lopes. Research data management behaviors in the image lifecycle [dataset]. *INESC TEC research data repository*, 2022. Available at: <https://doi.org/10.25747/hcrd-ht83>. Quoted 2 times on pages 8 and 51.
- [112] Joana Rodrigues and Carla Teixeira Lopes. Describing data in image format: Proposal of a metadata model and controlled vocabularies [dataset]. *INESC TEC research data repository*, 2022. Available at: <https://doi.org/10.25747/s54m-eh07>. Quoted 3 times on pages 9, 92, and 95.
- [113] Joana Rodrigues and Carla Teixeira Lopes. Images to describe data: Summary of interview transcripts to assess the role of imagery metadata in data description [dataset]. *INESC TEC research data repository*, 2023. Available at: <https://doi.org/10.25747/te4a-6544>. Quoted 2 times on pages 8 and 129.
- [114] Joana Rodrigues and Carla Teixeira Lopes. Images as metadata: A new perspective for describing research data. *Journal of Library Metadata*, 23(3-4):87–101, 2023. Available at: <https://doi.org/10.1080/19386389.2023.2252722>. Quoted on page 7.
- [115] Joana Rodrigues and Carla Teixeira Lopes. Research image management practices reported by scientific literature: Studies that explore the use and production of images in research. *Open Information Science*, 7(1), 2023. Available at: <https://doi.org/10.1515/opis-2022-0147>. Quoted on page 7.
- [116] Joana Rodrigues and Carla Teixeira Lopes. Research image management practices reported by scientific literature: Studies that explore the use and production of images in research [dataset]. *INESC TEC research data repository*, 2023. Available at: <https://doi.org/10.25747/k9f6-kt23>. Quoted 2 times on pages 8 and 66.
- [117] Joana Rodrigues and Carla Teixeira Lopes. Images annotated according to their content: a study on the description of data in image format in multiple domains [dataset]. *INESC TEC research data repository*, 2023. Available at: <https://doi.org/10.25747/szch-ve91>. Quoted 2 times on pages 8 and 98.
- [118] M. A. Sandweiss. Image and Artifact: The Photograph as Evidence in the Digital Age. *Journal of American History*, 94(1):193–202, 2007. Available at: <https://doi.org/10.2307/25094789>. Quoted 2 times on pages 27 and 126.
- [119] M. Saunders, P. Lewis, and A. Thornhill. *Research Methods for Business Students*. Pearson Education Limited, 7th edition, 2012. ISBN: 9781292016627. Quoted on page 50.
- [120] M. C. Serén. O documento fotográfico: Da mediação cultural a mediação técnica. *CEM: Cultura, Espaço e Memória*, 2011. Available at: <https://ojs.letras.up.pt/index.php/CITCEM/article/view/4855>. Quoted 3 times on pages 122, 124, and 125.
- [121] Maria do Carmo Serén. O documento fotográfico: Da mediação cultural a mediação técnica. *CEM: Cultura, Espaço e Memória*, 3, 2012. Available at: <https://api.semanticscholar.org/CorpusID:172586201>. Quoted on page 21.
- [122] James Roberto Silva. Fotografia e ciência: A utopia da imagem objetiva e seus usos nas ciências e na medicina. *Boletim do Museu Paraense Emílio Goeldi: Ciências Humanas*, 9(2):343–360, 2014. Available at: <https://doi.org/10.1590/1981-81222014000200006>. Quoted 2 times on pages 125 and 126.

- [123] P. Sinha and R. Jain. Classification and annotation of digital photos using optical context data. page 309–318. Proceedings of the 2008 international conference on Content-based image and video retrieval, 2008. Available at: <https://doi.org/10.1145/1386352.1386394>. Quoted on page 126.
- [124] The Royal Society. Final report - science as an open enterprise. Technical report, The Royal Society, 2012. Available at: <https://royalsociety.org/topics-policy/projects/science-public-enterprise/Report/>. Quoted on page 13.
- [125] Structural Reform Group. DDI Version 3.0 Conceptual Model. 2004. Available at: <https://ddialliance.org/DDI/Lifecycle/3.0/XMLSchema>. Quoted 4 times on pages 16, 39, 51, and 65.
- [126] Katiusa Stumpf. Ensino da leitura de imagem: expressão semântica e conteúdo sintático. *Revista Científica Multidisciplinar Núcleo do Conhecimento*, 7, 2020. Available at: <https://www.nucleodoconhecimento.com.br/comunicacao/expressao-semantica>. Quoted 2 times on pages 97 and 98.
- [127] B. Suh and B.B. Bederson. Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition. *International Journal of Human Computer Interaction*, 19(4):524–544, 2007. Available at: <https://doi.org/10.1016/j.intcom.2007.02.002>. Quoted on page 126.
- [128] 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. Available at: <https://doi.org/10.1371/journal.pone.0021101>. Quoted 3 times on pages 13, 23, and 49.
- [129] J. Tešić. Metadata practices for consumer photos. *IEEE MultiMedia*, 12(3):86–92, 2005. Available at: <https://doi.org/10.1109/MMUL.2005.50>. Quoted on page 26.
- [130] Andrew Treloar and Ross Wilkinson. Rethinking metadata creation and management in a data-driven research world. In *2008 IEEE Fourth International Conference on eScience*, pages 782–789, 2008. Available at: <https://ieeexplore.ieee.org/document/4736899>. Quoted on page 24.
- [131] UK Data Archive. Research data lifecycle, 2019. Available at: <https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx>. Quoted 3 times on pages 39, 51, and 65.
- [132] Heus P. Thomas W. Vardigan, M. Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, 3(1):107–113, 2008. Quoted on page 26.
- [133] Berenica Vejvoda, K. Jane Burpee, and Paula Lackie. Image Management as a Data Service. *IASSIST Quarterly*, 40(27-34), 2016. Available at: <https://doi.org/10.29173/iq399>. Quoted 2 times on pages 17 and 39.
- [134] Hollie C. White. Considering personal organization: Metadata practices of science. *Journal of Library Metadata*, 10:156–172, 2010. Available at: <https://doi.org/10.1080/19386389.2010.506396>. Quoted 4 times on pages 15, 121, 122, and 124.

- [135] Kelley E. Wilder. Photography and the art of science. *Visual Studies*, 24(2):163–168, 2009. Available at: <https://doi.org/10.1080/14725860903106161>. Quoted on page 22.
- [136] Mark D. Wilkinson and et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 2016. Available at: <https://doi.org/10.1038/sdata.2016.18>. Quoted on page 13.
- [137] 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. Available at: <https://doi.org/10.1002/asi>. Quoted 3 times on pages 23, 24, and 25.
- [138] R. K. Yin. *Case Study Research: Design and Methods*. SAGE Publications, 2013. ISBN: 978-1-4522-4256-9. Quoted on page 128.
- [139] W. Zhou, H. Li, and Q. Tian. Recent advance in content-based image retrieval: A literature survey. 2017. Available at: <https://arxiv.org/abs/1706.06064>. Quoted 2 times on pages 28 and 29.

Appendix A

Semi-structured interview guide: Exploring image management practices through interviews

Interviewee name:

Research domain:

Date:

1. Would you like to see research data from other researchers described with images? What advantages and disadvantages do you identify? (show example)
2. Would you use images in the description of your data if publishing platforms allowed it? What advantages and disadvantages do you identify?
3. Are you able to identify any scenario in which the description of the data fell short of expectations because you cannot use values other than text for the description? If yes, which one?
4. Remembering the dataset you deposited in the INESC TEC data repository (show deposit), can you identify:
 - 4.1 existing elements where it would be possible to complement the textual description with images? If yes, which ones?
 - 4.2 existing elements where the textual description would be replaced by images? If yes, which ones?
 - 4.3 new elements where it was possible to include images? If yes, which ones?
5. Have you thought about the possibility of a description with images?

6. Do you consider that the use of this description model would involve a greater or lesser expenditure of time compared to the use of textual description only?

Appendix B

Semi-structured interview guide: Images to describe research data: a case study on the use of imagery metadata

Name of interviewee:

Research domain:

Data:

1. Would you like to see other researchers' research data described with images? What advantages and disadvantages can you identify? (show example)
2. Would you use images to describe your data if the publishing platforms allowed it? What advantages and disadvantages would you identify?
3. Can you identify any scenarios in which the data description fell short of expectations because you couldn't use values for the description other than text? If so, which one?
4. Looking back to the dataset you deposited in INESC TEC's data repository (show deposit), can you identify:
 - 4.1 existing elements where it would be possible to complement the textual description with images? If so, which ones?
 - 4.2 existing elements where you could replace the textual description with images? If so, which ones?
 - 4.3 new elements where images could be included? If so, which ones?
5. Have you ever thought about the possibility of describing with images?

6. Would using this description model require more or less time than using only textual descriptions?