

Miguel Martins Fernandes

A utilização e produção de imagens em contexto de
investigação: contribuições para boas práticas de gestão
de dados em formato de imagem

Dissertação realizada no âmbito do Mestrado em Ciência da Informação,
orientada pela Professora Doutora Carla Teixeira Lopes e coorientada pela
Dra. Joana Rodrigues

Faculdade de Engenharia e Faculdade de Letras
Universidade do Porto

Julho de 2020

A utilização e produção de imagens em contexto de investigação: contribuições para boas práticas de gestão de dados em formato de imagem

Miguel Martins Fernandes

Dissertação realizada no âmbito do Mestrado em Ciência da Informação, orientada pela Professora Doutora Carla Teixeira Lopes e coorientada pela Dra. Joana Rodrigues

Membros do Júri

Presidente: Professor Doutor António Lucas Soares

Faculdade de Engenharia - Universidade do Porto

Orientadora: Professora Doutora Carla Teixeira Lopes

Faculdade de Engenharia - Universidade do Porto

Arguente: Professora Doutora Maria Cristina Guardado

Escola Superior da Tecnologia de Águeda - Universidade de Aveiro

Agradecimentos

Na elaboração deste trabalho foram vários os que contribuíram para que fosse possível atingir o fim desta dissertação.

À Professora Carla Teixeira Lopes, pela disponibilidade, apoio, espírito crítico, conhecimento e conselhos dados ao longo do desenvolvimento do trabalho.

À Joana Rodrigues, igualmente pelo apoio, acompanhamento, conhecimento, conselhos transmitidos e por me introduzir na gestão de dados de investigação em formato de imagem.

Aos investigadores que aceitaram disponibilizar um pouco do seu tempo para participar nas entrevistas: Dr^a. Susana Barbosa, Dr^a. Carolina Catorze, Dr. Pedro Ramos, Dr^a. Ângela Lomba, Dr^a. Cristiana Alves, Dr^a. Marisa Cardoso, Dr. Romão Santos e Dr. Nuno Franco.

Ao Professora Luís Teixeira pela ajuda na validação do questionário.

Aos membros do laboratório I123, especialmente ao Professor João Rocha da Silva e à Yulia Karimova pela disponibilidade e contributo dado na importação para o Dendro e depósito dos dados respetivamente.

Aos docentes do Mestrado em Ciência da Informação pela partilha do conhecimento e experiência.

À minha família, por acreditarem sempre nas minhas capacidades e pelo apoio, sacrifício e incentivo. Sem eles não conseguiria alcançar esta etapa.

À minha namorada, Ana, pela atenção, carinho, encorajamento e apoio dado durante todo o trabalho. As palavras de agradecimento serão sempre poucas.

Por fim, aos meus colegas e amigos, André e Miguel, por todos os bons momentos passados durante estes cinco anos e, também, pelo apoio e opinião dada durante esta fase.

A todos um muito obrigado!

Resumo

Nas últimas décadas temos assistido à produção de grandes quantidades de dados na ciência. Esta produção mais intensiva deve-se, em parte, aos desenvolvimentos tecnológicos e científicos de meados do século XX que fizeram emergir novos métodos, ferramentas e instrumentos produtores de dados. Neste contexto, torna-se necessário criar mecanismos que garantam a utilização eficiente dos dados durante o maior período possível. A gestão de dados de investigação assume um papel essencial para assegurar que os dados são facilmente localizáveis, acessíveis, com menor probabilidade de serem perdidos e mais utilizáveis, tanto durante a investigação, como no futuro. As imagens são também utilizadas como dados de investigação, constituindo um elemento informacional fundamental em determinados projetos de investigação. No âmbito da gestão de dados de investigação, a problemática da gestão de imagens utilizadas como dados de investigação tem sido pouco abordada. A presente dissertação descreve o estudo das práticas e hábitos de investigadores na gestão das imagens e recomenda também boas práticas para essa gestão. A abordagem metodológica envolveu a criação e utilização de um questionário e realização de entrevistas a investigadores para conhecer as suas práticas na produção e utilização de imagens. A informação recolhida foi analisada e usada no desenvolvimento das recomendações. Estas recomendações assumem a forma de um guia de orientações e um modelo de metadados para auxiliar a descrição das imagens. O seu objetivo primário consiste em fomentar boas práticas na gestão de dados de investigação em formato de imagem. O modelo de metadados passou ainda por um momento de avaliação por parte de investigadores. Os resultados obtidos permitiram demonstrar que as imagens são efetivamente utilizadas como dados de investigação, embora os investigadores não estejam sensibilizados para a sua gestão. Verificou-se também a existência de hábitos similares e diferentes nos diferentes domínios e frequências de uso da imagem. A avaliação do modelo de metadados foi positiva e este considerado útil pelos investigadores.

Palavras-chave: Gestão de dados de investigação; Gestão de imagens; Modelo de metadados

Abstract

In the past few decades, we have seen the production of large amounts of data in science. This data intensive science is due in part to the technological and scientific developments of the mid-20th century, which produced new methods, tools and instruments that produce data. In this context, it is necessary to create mechanisms that guarantee the efficient use of data for the longest possible period. Research data management plays an essential role in ensuring that data is easily located, accessible, less likely to be lost and more usable, both during an investigation and in the future. Images are also used as research data, constituting a fundamental informative element in research projects. In the scope of research data management, the problem of image management has been little addressed. This dissertation describes the study of the practices and habits of researchers in image management and also recommends practices for this management. A methodological approach involves creating and using a questionnaire and conducting interviews with researchers to learn about practices in the production and use of images. The information collected was analyzed and used in the development of the recommendations. These recommendations take the form of a guide to manage images in research and a metadata model to help researchers describe their images. Its primary objective is to foster good research data management practices in image format. The metadata model underwent by an evaluation session performed by researchers. The results obtained allow us to demonstrate that images are effectively used as research data, although the researchers are not aware of their management. Similar and different habits were also found in different domains and frequencies of image use. The evaluation of the metadata model was positive and it was considered useful by the researchers.

Keywords: Research Data Management; Image Management; Metadata Model

Lista de ilustrações

Figura 1 - Árvore de objetivos.	2
Figura 2 - Ciclo da Investigação.	11
Figura 3 - Ciclo de gestão da informação de Choo.	12
Figura 4 - Ciclo de vida da informação de Pinto e Malheiro.	13
Figura 5 - Ciclo de vida dos dados do Data Documentation Initiative.	14
Figura 6 - Ciclo de vida dos dados do DataONE.	15
Figura 7 - Ciclo de vida dos dados do UK Data Archive.	16
Figura 8 - Modelo de gestão de dados de investigação do Data Curation Centre.	18
Figura 9 - Organização geral do trabalho.	33
Figura 10 - Processo de desenvolvimento do questionário.	33
Figura 11 - Metodologia das entrevistas.	35
Figura 12 - Frequência de utilização de imagens por área científica.	37
Figura 13 - Elaboração de um guia de orientação por domínio científico e frequência de uso.	38
Figura 14 - Utilização de instrumentos de captura por área científica.	39
Figura 15 - Utilização de instrumentos de captura por frequência de uso.	39
Figura 16 - Edição de imagem por área científica.	41
Figura 17 - Edição de imagem por frequência de uso.	41
Figura 18 - Descritores relevantes por área científica.	46
Figura 19 - Descritores utilizados por área científica.	46
Figura 20 - Formato de armazenamento por área científica.	47
Figura 21 - Formato de armazenamento por frequência de uso.	48
Figura 22 - Promotor da partilha de imagens por área científica.	49
Figura 23 - Processo de desenvolvimento do modelo de metadados.	59
Figura 24 - Data properties e annotation properties do descritor "tipo".	60
Figura 25 - Ontologia no Dendro.	61

Lista de tabelas

Tabela 1 - Comparação dos ciclos de vida da informação.	13
Tabela 2 - Comparação dos ciclos de vida dos dados de investigação.	17
Tabela 3 - Comparação dos modelos de metadados.	29
Tabela 4 - Resultados significativos no planeamento por área científica.	38
Tabela 5 - Resultados significativos no planeamento por frequência de uso da imagem.	39
Tabela 6 - Resultados significativos na criação/compilação por área científica.	40
Tabela 7 - Resultados significativos na criação/compilação por frequência de uso.	40
Tabela 8 - Resultados significativos na asseguaração da qualidade por área científica.	42
Tabela 9 - Resultados significativos na asseguaração da qualidade por frequência de uso.	42
Tabela 10 - Resultados significativos no processamento/análise por área científica.	43
Tabela 11 - Resultados significativos no processamento/análise por frequência de uso.	43
Tabela 12 - Resultados significativos na descrição por frequência de uso.	44
Tabela 13 - Resultados significativos na descrição por área científica.	45
Tabela 14 - Resultados significativos no armazenamento por área científica.	47
Tabela 15 - Resultados significativos no armazenamento por frequência de uso.	48
Tabela 16 - Resultados significativos na partilha por área científica.	49
Tabela 17 - Resultados significativos na partilha por frequência de uso.	50
Tabela 18 - Descritores mencionados pelos entrevistados.	55
Tabela 19 - Descritores da ontologia.	62
Tabela 20 - Descrição de um conjunto de imagens através da ontologia.	64
Tabela 21 - Tipos de licenças de utilização.	75

Lista de abreviaturas e siglas

AAT - Art and Architecture Thesaurus
ABCD - Access to Biological Collection Data
CASRAI - Consortia Advancing Standards in Research Administration Information
CEE - Ciências Exatas e Engenharia
CERIF - Common European Research Information Format
CNA - Ciências Naturais e do Ambiente
CSH - Ciências Sociais e Humanidades
CVS - Ciências da Vida e da Saúde
DDI - Data Documentation Initiative
DCC - Digital Curation Centre
DIF - Directory Interchange Format
DOIs - Digital Object Identifiers
EML - Ecological Metadata Language
EXIF - Exchangeable Image File Format
FAIR - Findable, Accessible, Interoperable and Re-usable
FEUP - Faculdade de Engenharia da Universidade do Porto
FITS - Flexible Image Transport System
FLUP - Faculdade de Letras da Universidade do Porto
INESC-TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência
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 - Organização para a Cooperação e Desenvolvimento Económico
OWL - Web Ontology Language
QUDV - Quantities, Units, Dimensions and Values
UP – Universidade de Porto
URI - Uniform Resource Identifiers
RDA - Research Data Alliance
RDF - Resource Description Framework
XMP - Extensible Metadata Platform
XML - Extensible Markup Language

Índice de conteúdo

1. Introdução	1
1.1. Contexto e Motivação	1
1.2. Objetivos e resultados esperados	1
1.3. Contribuições	2
1.4. Estrutura da dissertação	3
2. Revisão da literatura	5
2.1. Gestão de dados de investigação	5
2.1.1. Princípios FAIR e dados abertos	6
2.1.2. Benefícios e desafios na partilha de dados de investigação	7
2.1.3. Contexto e definição de dados de investigação	8
2.1.4. Ciclo de gestão de dados de investigação	9
2.1.5. Ciclo de vida da informação	11
2.1.6. Ciclos de vida dos dados de investigação	13
2.1.7. Modelo de gestão de dados de investigação	17
2.2. Imagem como dado de investigação	19
2.2.1. Natureza dos recursos imagéticos	19
2.2.2. A imagem como evidência científica	19
2.2.3. O uso da imagem na ciência	20
2.3. Descrição de dados de investigação	22
2.3.1. Metadados	23
2.3.2. Modelos de Metadados	25
2.3.2.1. Dublin Core	25
2.3.2.2. Common European Research Information Format	26
2.3.2.3. Exchangeable Image File Format	26
2.3.2.4. Modelos de metadados específicos	27
2.3.3. Desafios na descrição da imagem	29
3. Metodologia	32
3.1. Inquérito por Questionário	33
3.2. Inquérito por Entrevista	35
3.3. Recomendação de boas práticas	36
4. Utilização e produção de imagens em investigação - estudo via inquérito por questionário	37
4.1. Planeamento	37
4.2. Criação/Compilação	39
4.3. Assegurar a qualidade	41

4.4. Processamento/Análise	42
4.5. Descrição	44
4.6. Armazenamento	46
4.7. Partilha	48
5. Utilização e produção de imagens em investigação - estudo via inquérito por entrevista	51
5.1. Planeamento	52
5.2. Criação, recolha e compilação	53
5.3. Assegurar a qualidade	53
5.4. Processamento/Análise	54
5.5. Descrição	54
5.6. Armazenamento	55
5.7. Partilha	56
6. Proposta de modelo de metadados para a descrição de imagens	58
6.1. Processo de desenvolvimento da ontologia	58
6.2. Descritores do modelo	61
6.3. Exemplo prático de utilização do modelo	63
6.4. Avaliação do modelo de metadados	66
7. Proposta de boas práticas para a gestão de imagens em investigação	68
7.1. Planeamento	68
7.2. Documentação e Metadados	69
7.3. Armazenamento e Backup	71
7.4. Partilha	73
8. Conclusões e perspetivas de trabalho futuro	77
Referências bibliográficas	80
ANEXOS	87
ANEXO I - Questionário	87
ANEXO II - Guião da entrevista	91
ANEXO III - Documento para importação no Dendro.	93
ANEXO IV - Avaliação do modelo de descrição de dados em formato de imagem	96

1. Introdução

1.1. Contexto e Motivação

Atualmente existe pouco conhecimento relativo aos processos de produção, utilização e gestão de imagens em contexto de investigação. Assim o seu estudo torna-se importante pois, por um lado, as imagens constituem um elemento informacional valioso para a investigação. Por outro lado, ao serem utilizadas como um dado de investigação torna-se necessário a sua inclusão nos processos de gestão de dados de investigação.

Com o desenvolvimento tecnológico que se tem vindo a verificar e a imposição do digital na sociedade, colocam-se novos desafios e tendências à forma como se lida com a informação. As imagens não ficaram de fora deste desenvolvimento, verificou-se o aparecimento de novas formas de capturar e produzir imagens e, o aumento da capacidade de armazenamento e processamento. Estes feitos traduzem-se num aumento progressivo da utilização de imagens. Com o maior uso de imagens, surgem também novos desafios na descrição, recuperação, interpretação e uso sistemático das mesmas. O advento da análise automática de imagens e a introdução de novas práticas na descrição de imagens trazem algumas soluções para os desafios introduzidos pelo crescente uso da imagem.

No âmbito da Ciência da Informação, particularmente na gestão de dados de investigação, a imagem tem sido utilizada para registar os resultados do processo investigativo e servir como evidência da investigação. Assim, é necessário incluir as imagens nos processos de gestão de dados de investigação para as tornar acessíveis e reutilizáveis em investigações futuras. Porém, as imagens possuem algumas características que tornam a gestão deste tipo de dados de investigação especial. Geralmente as imagens ocupam um volume considerável e têm de ter um tratamento especial nas fases de armazenamento e preservação. Nestas fases tem de se garantir a existência de espaço suficiente para as armazenar e que não existe perda de informação nem de qualidade nas imagens. Na partilha, as imagens requerem técnicas específicas para a proteger a privacidade e práticas de preservação que garantam a reutilização de imagens tal como elas foram produzidas. O pouco foco que se tem dado à gestão da imagem em contexto de investigação faz com que este tema seja pouco falado na literatura. Por conseguinte, é urgente o seu desenvolvimento devido ao valor que estas possuem no contexto de investigação.

1.2. Objetivos e resultados esperados

O principal objetivo desta dissertação é contribuir para o estudo e fomento de boas práticas na utilização e produção de imagens de apoio à investigação.

Na Figura 1 apresenta-se a árvore de objetivos para esta dissertação. Como podemos observar, no primeiro nível de objetivos temos o estudo do uso e produção de imagens em contexto de investigação e a recomendação de boas práticas para o uso e produção de imagens em projetos de investigação. Para estudar o uso e produção de imagens em contexto de investigação torna-se necessário, por um lado, analisar quais os procedimentos de gestão de dados de investigação em formato de imagem existentes em diferentes grupos de investigação e, por outro lado, sistematizar os procedimentos relativos ao ciclo da imagem em diferentes grupos de investigação. Para contribuir para a recomendação de boas práticas para o uso e produção de imagens em projetos de investigação é necessário comparar e desenvolver um modelo de metadados, de modo a auxiliar os investigadores a descrever as imagens, e desenvolver um documento que oriente a gestão de dados imagéticos durante a investigação.

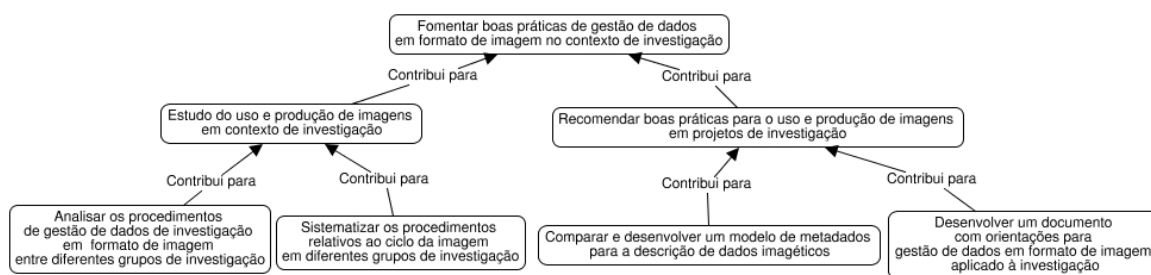


Figura 1 - Árvore de objetivos.

Desta dissertação resulta um guia de boas práticas para a gestão de dados em formato de imagem no contexto de investigação e um modelo de metadados para a descrição das imagens utilizadas em contexto de investigação.

1.3. Contribuições

O guia de boas práticas e o modelo de metadados desenvolvido no âmbito desta dissertação estão disponíveis *online*¹. É importante que estas ferramentas estejam acessíveis, de modo a constituírem uma ajuda para os investigadores na gestão de imagens.

No período em que se desenvolveu a dissertação, foi escrito e aceite um artigo científico para a conferência internacional *24th International Conference on Theory and Practice of Digital Libraries*. O artigo apresenta e discute os resultados obtidos no questionário. O objetivo do artigo foi contribuir para a exploração e conhecimento dos hábitos e processos realizados pelos investigadores, em relação à gestão de dados de investigação em formato de imagem.

¹ <https://drive.google.com/drive/folders/16BEDlkiAlb2kmTGZFzR44ekpjlESTUZ?usp=sharing>

1.4. Estrutura da dissertação

A dissertação encontra-se dividida em oito secções: “Introdução”, “Revisão da literatura”, “Metodologia”, “Utilização e produção de imagens em investigação - estudo via inquérito por questionário”, “Utilização e produção de imagens em investigação - estudo via inquérito por entrevista”, “Proposta de modelo de metadados para a descrição de imagens”, “Proposta de boas práticas para a gestão de imagens em investigação” e “Conclusões e perspectivas de trabalho futuro”.

A presente secção “Introdução” contextualiza a pertinência do tema da dissertação e apresenta os objetivos, resultados esperados e a estrutura da dissertação.

A “Revisão da literatura” enquadra teoricamente o trabalho a ser desenvolvido na presente dissertação. No ponto intitulado “Gestão de dados de investigação” descreve-se o seu aparecimento, importância, benefícios e desafios da gestão de dados de investigação. E, ainda, o contexto dos dados utilizados no processo de investigação e os modelos e ciclos de vida de gestão de dados de investigação. Na subsecção “Imagem” aborda-se a natureza dos recursos imagéticos, com especial foco na capacidade de transmitirem informação, o uso de imagens na ciência e o fundamento para servirem como evidência científica. Na última subsecção, “Descrição de dados de investigação”, contextualiza-se a utilização de metadados e de modelos de metadados na descrição de dados de investigação. Apresentam-se também alguns modelos de descrição frequentemente usados em diversos domínios. Ainda neste ponto, abordam-se os desafios encontrados na descrição de imagens e as abordagens existentes na descrição deste formato.

A secção “Metodologia” explica a metodologia empregue para atingir os objetivos da dissertação, nomeadamente as técnicas de recolha de dados e os métodos de análise de dados. Nesta secção estará fundamentado o desenho de investigação seguido para alcançar os objetivos propostos tal como o enquadramento e descrição detalhada da abordagem metodológica.

Em “Utilização e produção de imagens em investigação - estudo via inquérito por questionário” apresentam-se os resultados obtidos após a análise das respostas do questionário. De forma igual, em “Utilização e produção de imagens em investigação - estudo via inquérito por entrevista” encontram-se os resultados obtidos com a realização das entrevistas.

As secções “Proposta de modelo de metadados para a descrição de imagens” e “Proposta de boas práticas para a gestão de imagens em investigação” apresentam o modelo de metadados desenvolvido e as boas práticas na gestão de dados em formato de imagem.

Na secção “Conclusões e perspectivas de trabalho futuro” apresentam-se as principais conclusões retiradas do desenvolvimento do trabalho e as perspectivas de trabalho futuro identificadas.

2. Revisão da literatura

2.1. Gestão de dados de investigação

A revolução tecnológica e científica que se tem vindo a sentir desde meados do século XX, deu origem ao aparecimento de novos métodos, instrumentos e ferramentas de investigação. Estas mudanças conduziram ao aumento do volume, complexidade e importância dos dados de investigação. Aliado ao aumento da capacidade computacional e de armazenamento digital, a recolha de dados é cada vez mais intensiva, causando alterações e desafios à forma como se armazena, preserva, acede e partilha dados inseridos no contexto da atividade científica (Ribeiro, Rodrigues, Fernandes & Saraiva, 2010). A utilização generalizada por todas as áreas de investigação de tecnologias de informação potenciou a abertura e livre troca de ideias, informação e conhecimento, assim como o aparecimento de novas formas de colaborar. Desta forma, o acesso a dados de investigação de forma responsável e eficiente é algo que deve ser retirado das novas oportunidades e benefícios trazidos pelas tecnologias de informação (OECD, 2007).

A ciência moderna é caracterizada pela criação, disseminação e análise de grandes quantidades de dados e pela variedade de recursos de criação e recolha de dados. Esta nova fase da ciência originou o aparecimento do termo *data-intensive science*, onde a recolha, curadoria e análise de dados são atividades essenciais para os dados terem valor no futuro (Bell, 2009; Borgman, Wallis, & Enyedy, 2007). A criação e desenvolvimento de *software* e *hardware*, de instituições, infraestruturas e de pessoas especializadas, assim como de diretrizes e políticas direcionadas para a gestão de dados de investigação têm como objetivo mitigar os desafios à sua gestão. Os desafios englobam as questões de segurança, privacidade, troca e direitos de autoria dos dados de investigação e aumento da capacidade de extração de conhecimento dos dados de investigação através da sua partilha e reutilização (Genova et al., 2014).

A gestão de dados de investigação envolve um conjunto de práticas direcionadas para a informação recolhida e produzida durante o processo de investigação. É um processo estruturado e alinhado com o contexto e área científica da investigação (Higman, Bangert, & Jones, 2019), é algo que se realiza antes, durante e depois do processo de investigação terminar. Abrange o planeamento da gestão de dados, documentação e organização dos dados, melhoria da análise, assim como a garantia da segurança e armazenamento adequado dos dados produzidos na investigação. A gestão de dados combina diferentes práticas com o objetivo de tornar os dados facilmente localizáveis, menos prováveis de serem perdidos e mais utilizáveis, tanto durante a investigação, como no futuro (Briney, 2015).

2.1.1. Princípios FAIR e dados abertos

Estritamente ligado à gestão de dados de investigação encontram-se os princípios *Findable, Accessible, Interoperable and Re-usable* (FAIR) e de dados abertos. Embora sejam conceitos diferentes, quando aplicados em conjunto fortalecem, enfatizam e enriquecem os objetivos de cada conceito. Segundo Higman, Bangert e Jones (2019), a gestão de dados é a base que possibilita os princípios FAIR e a ciência aberta, se os dados não forem criados e geridos de maneira correta desde o início de investigação, a sua partilha não vai ser realizada de forma eficiente, dificultando a aplicação dos conceitos FAIR e de dados abertos.

O conceito de dados FAIR surgiu em 2014, impulsionado pelo desafio que é a descoberta de conhecimento na *data-intensive science*. Os Princípios FAIR vêm auxiliar humanos e máquinas a descobrir, aceder, integrar e analisar dados apropriados para as tarefas da investigação tal como para os fluxos de trabalho e algoritmos associados (FORCE11, 2019). Os Princípios FAIR definem as propriedades que os recursos de dados contemporâneos, ferramentas, vocabulários e infraestruturas devem ter para apoiar a descoberta e acesso aos dados. Não são uma especificação que sugere a tecnologia, normas e modo de implementação. O alto nível dos Princípios FAIR precedem a implementação, funcionando como um guia que assiste os curadores de dados a escolher a solução que mais se adequa aos seus dados de investigação (Wilkinson et al., 2016).

O conceito de dados abertos traduz-se em dados passíveis de serem acedidos e reutilizados sem qualquer tipo de restrição legal ou técnica, embora a utilização dos dados se encontre sujeita às restrições legais que lhes são aplicáveis (OECD, 2015). Os benefícios dos dados abertos refletem-se na criação de novos postos de emprego, atração de investimento, desenvolvimento de novos produtos e tecnologias e, a nível governamental, maior transparência na tomada de decisão e nas políticas públicas (Royal Society, 2012). Aplicado à investigação científica, o objetivo é tornar dados de investigação publicamente disponíveis para acesso e reutilização com as menores restrições possíveis (Higman et al., 2019).

A partilha de dados consiste na divulgação e reutilização dos dados por terceiros, esta pode ser realizada através de canais privados ou públicos, como o depósito dos dados em coleções de dados públicas. O seu público alvo pode ir desde investigadores de um ramo científico específico ao público em geral. O grau de utilidade, credibilidade e valor dos dados de investigação pode variar muito, os dados podem-se encontrar altamente estruturados e descritos, como podem ser não estruturados e acompanhados de pouca documentação descritiva (Borgman, 2012).

Num estudo realizado a nível internacional sobre práticas de gestão de dados (Tenopir

et al., 2011), verificou-se que os investigadores encaram positivamente a possibilidade de partilharem os seus dados na totalidade ou apenas parte com certas condições e restrições de acesso e uso. Algumas formas de fomentar a partilha nos investigadores passam pela citação formal dos dados, obtenção de cópias dos artigos e aprendizagem através dos produtos e publicações que usam os seus dados. Os investigadores apontaram ainda que a sua capacidade de responder a determinadas questões científicas, como o aquecimento global e as alterações climáticas, eram limitadas pela falta de acesso a dados criados por outros investigadores.

2.1.2. Benefícios e desafios na partilha de dados de investigação

A partilha e acesso a dados de investigação evita a repetição de investigações e os custos inerentes à sua realização. Neste sentido, diversas entidades financiadoras da atividade científica têm produzido orientações e regulamentos para a divulgação dos dados produzidos na investigação. A reutilização de dados de investigação permite a verificação, validação e deteção de erros, assim como a experimentação de novas hipóteses não testadas na investigação original, funcionando como a base para a realização de novos projetos. Por outro lado, a avaliação e revisão de candidaturas a bolsas de investigação tornam-se processos mais simples, devido ao acesso facilitado aos dados produzidos pelos investigadores e ao seu impacto no desenvolvimento científico, ou seja, à qualidade e eficiência da investigação. A divulgação dos dados produzidos no decorrer da investigação pode ser benéfica para o aumento ou reforço do reconhecimento e credibilidade dos investigadores, instituições envolvidas e dos repositórios que ajudam a preservar e a disseminar os dados (Ribeiro et al., 2010; Eynden, Corti, Woollard, Bishop, & Horton, 2011).

A possibilidade de acesso aos dados que servem de evidência da investigação aumenta a transparência, a identificação de fraudes e a dissuasão de plágio (Ribeiro et al., 2010). A transparência é conseguida através da combinação de comportamentos, práticas e processos de investigação associados à sua reprodutibilidade. Serve para demonstrar o rigor das metodologias e protocolos de investigação e, consequentemente, para reforçar a opinião pública sobre a qualidade, confiança e integridade dos resultados, conclusões e teorias retiradas da investigação (Lyon, 2016).

Porém, a partilha de dados de investigação e a sua reutilização, assim como interpretação por terceiros, é uma tarefa difícil devido a várias camadas de complexidade relacionadas com a natureza dos dados e da investigação, inovação, bolsas de investigação, privacidade e proteção de dados, motivação e incentivos aos investigadores, políticas públicas, propriedade intelectual e retorno económico (Borgman, 2012). Outra faceta negativa da partilha de dados prende-se com o facto dos dados serem utilizados ou

compreendidos de forma errada. O que afeta a reprodução e o ritmo da investigação, e por outro lado, deixa disponível dados que podem conduzir a erros científicos (Borgman et al., 2007; Genova et al., 2014).

A partilha e reutilização de grandes quantidades de dados depende da existência de infraestruturas de partilha de dados, como bases de dados, repositórios e centros de dados, que possibilitem o armazenamento, curadoria e acesso continuado. Por sua vez, os dados para serem recuperados corretamente necessitam de formatos padronizados e de metadados (Day, 2005).

2.1.3. Contexto e definição de dados de investigação

Segundo o modelo de referência *Open Archival Information System* (OAIS) citado pelo *Digital Curation Centre* (DCC) (Digital Curation Center, 2019b), dados são representações de informação reinterpretáveis, formalizadas de maneira a serem compreendidas, processadas ou comunicadas, como por exemplo, sequências de bits, uma tabela com números, os caracteres numa página, o registo de som ou um espécime de rocha lunar.

O documento “*Principles and guidelines for access to research data from public funding*” da Organização para a Cooperação e Desenvolvimento Económico (OECD) (OECD, 2007) define dados de investigação como os registos factuais utilizados como fontes primárias na investigação científica, aceites pela comunidade científica e indispensáveis para validar os resultados da investigação. Estes registos podem ser registos textuais, numéricos, imagens ou som. O relatório “*Making open science a reality*” (OECD, 2015) acrescenta que os dados de investigação são recolhidos e produzidos numa grande variedade de formatos desde folhas de cálculo digitais e bases de dados até compilações de questionários, imagens e objetos. Refere ainda que a sua consulta e utilização pode requerer, em muitos casos, o uso de programas de computador específicos, entre outros.

Na mesma linha de pensamento o *Consortia Advancing Standards in Research Administration Information* (CASRAI) (2015) refere que dados de investigação dizem respeito aos dados, observações ou registos numéricos, textuais, imagens e vídeos, nos mais diversos formatos digitais. Os dados servem de evidência do processo de investigação. Estes são utilizados como fontes primárias para suportar a investigação e/ou aceites pela comunidade científica como necessários para validar os resultados da investigação. Relativamente à sua natureza, os dados de investigação podem ser dados experimentais, observacionais, operacionais, fornecidos por terceiros, dados públicos, dados processados, dados de monitorização ou dados reaproveitados.

Os dados de investigação podem-se distinguir em dados primários e secundários,

dependendo do grau de processamento a que são submetidos. Dados primários são dados que não foram alvo de qualquer tipo de transformação ou processamento, por exemplo, dados recolhidos através de um instrumento sem qualquer processamento. Dados secundários resultam da interpretação, processamento ou transformação dos dados primários, como por exemplo, aplicação de modelos estatísticos aos dados recolhidos por um instrumento (Furtado, Príncipe, & Carvalho, 2017).

De acordo com a sua origem, os dados de investigação podem ser classificados da seguinte forma (OECD, 2015) (National Science Board, 2005):

- Dados observacionais – dados resultantes do registo de informação histórica e de fenómenos que só acontecem uma vez, através de telescópios, satélites, sensores, questionários e demais instrumentos de recolha de dados. Em muitos casos, os dados não podem ser replicados pelo que a sua preservação é aconselhada. Alguns exemplos deste tipo de dados são fotografias de uma supernova e observação direta da temperatura do oceano numa data específica;
- Dados experimentais – dados recolhidos através de experiências controladas, como ensaios clínicos e biomédicos, teste farmacêuticos, entre outros. Particularmente importante preservar os dados experimentais cujas condições de replicação não são fáceis de reproduzir, dispendiosas monetariamente ou por motivos éticos. Identificação de padrões na expressão genética constitui um exemplo destes dados;
- Dados computacionais – dados criados a partir de simulações computacionais de grande escala. É importante preservar estes dados pois apesar de serem recriados através da repetição da simulação, o tempo disponível para utilizar o computador pode ser limitado e os dados podem ser utilizados como base para análise, visualização ou *data mining*. Por vezes, é mais relevante preservar o modelo que deu origem aos dados, quando existem recursos que permitam reproduzir a simulação facilmente. Os resultados de execução de um modelo computacional são exemplo de dados computacionais;
- Dados de referência – dados altamente curados e pretendidos por diversas áreas científicas, estes dados são criados para diversos propósitos, podendo ir desde o mapeamento do genoma humano e documentação de proteínas até dados longitudinais sobre aspetos económicos e sociais.

2.1.4. Ciclo de gestão de dados de investigação

Os modelos do ciclo de gestão de dados de investigação representam abstrações explícitas e simplificações dos elementos e processos que constituem a gestão de dados.

Consiste numa forma simples, compreensível e orgânica de visualizar as diferentes fases da gestão de dados de investigação através de um modelo descritivo. Os modelos podem variar desde abstrações do processo de investigação a modelos detalhados de criação, reutilização ou curadoria de dados aplicados num contexto específico, como uma universidade ou instituição, ou generalizados a qualquer investigação (Cox & Tam, 2018). O ciclo procura definir os conceitos e momentos chave no planeamento e implementação da gestão de dados de investigação. Estes conceitos e fases-chave podem sofrer alterações com base na área científica, do tipo de dados, entre outros fatores (Furtado et al., 2017).

Esta abordagem torna-se necessária pois os recursos digitais são frágeis e suscetíveis de mudar devido ao desenvolvimento tecnológico durante o seu ciclo de vida. Assim, a existência ou a falta de atividades durante o seu ciclo influencia diretamente a capacidade de gerir e preservar os recursos. A reutilização dos materiais digitais só é possível se existirem etapas de curadoria que salvaguardem a sua integridade e autenticidade (Pennock, 2007).

A gestão de dados de investigação implica a organização de dados desde o momento em que entram no ciclo de investigação até ao arquivo e disseminação dos resultados. Tem como finalidade assegurar a qualidade dos resultados obtidos bem como contribuir para um processo investigativo novo e inovador a partir da informação existente. A curadoria e preservação de dados de investigação são dois conceitos estritamente ligados à gestão de dados de investigação. A curadoria trata de assegurar que os resultados do projeto são elegíveis para arquivar e que constituem ativos importantes da investigação passíveis de serem reutilizados. Por outro lado, a preservação foca a acessibilidade e reutilização dos recursos a longo termo (Whyte & Tedds, 2011).

O ciclo de gestão de dados de investigação deve encontrar-se sincronizado com o processo investigativo, pelo que se torna importante conhecer o seu funcionamento. O processo de investigação começa geralmente com o aparecimento de uma ideia ou hipótese. A ideia vai ser desenvolvida e trabalhada para dar origem a um produto, na maioria dos casos, um artigo científico. Ao longo do processo são utilizados equipamentos, metodologias e passos que permitem avançar no processo de investigação. Cada um destes passos é suportado por dados que ajudam a desenvolver a investigação e apoiam a tomada de decisão. Pryor (2018) apresenta um modelo de investigação centrado nos dados (Figura 2), onde a recolha de dados começa na formulação das hipóteses para analisar o que já foi feito e selecionar dados para a investigação. De seguida, a investigação vai necessitar de recolher dados publicados para reanalisar os dados em novos contextos ou para comparar com os dados produzidos no

decorrer da investigação, esta fase é a principal fonte de novos dados de investigação. Todas as fases do ciclo usam ou geram dados a determinado ponto. A publicação e reutilização depois dos dados serem interpretados, analisados e sintetizados apresentam-se como essenciais para o desenvolvimento do conhecimento e de novas oportunidades de investigação.

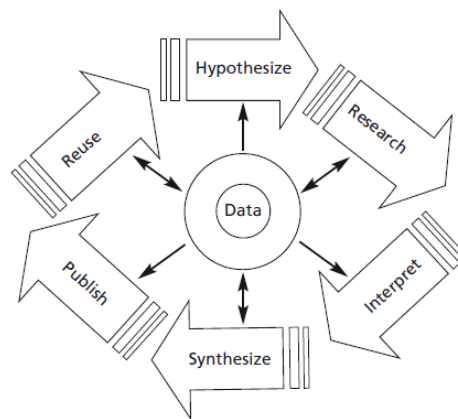


Figura 2 - Ciclo da Investigação.

Fonte: (Pryor, 2018)

2.1.5. Ciclo de vida da informação

O ciclo de vida da informação tenta captar as fases pelas quais a informação passa num determinado contexto. De seguida apresentam-se alguns ciclos de vida da informação.

2.1.5.1. Modelo de Choo

Segundo Choo (2002) (Figura 3), a informação em contexto organizacional passa pelas seguintes fases ao longo do seu ciclo de vida: identificação das necessidades de informação, aquisição da informação, organização e armazenamento, serviços e produtos de informação, distribuição, uso e comportamento adaptativo. O ciclo inicia-se com o comportamento adaptativo, este refere-se à capacidade da organização iniciar um novo ciclo de aprendizagem baseado na informação produzida através das ações e decisões. Na segunda etapa é identificado o contexto da organização e os requisitos informacionais necessários para resolver problemas e apoiar a tomada de decisão. Estritamente ligado à identificação das necessidades encontra-se aquisição e seleção das fontes de informação. De seguida a informação necessita de ser estruturada de acordo com as necessidades e modo de uso da informação, o seu armazenamento contribui para a formação de um repositório ativo da informação e conhecimento da organização. A informação adquirida em conjunto com a armazenada podem dar origem a diferentes produtos e serviços de informação, que vão realçar a qualidade da informação e responder às necessidades e uso da informação. A utilização da informação consiste na criação e aplicação do conhecimento através de processos interpretativos e de tomada de

decisão.

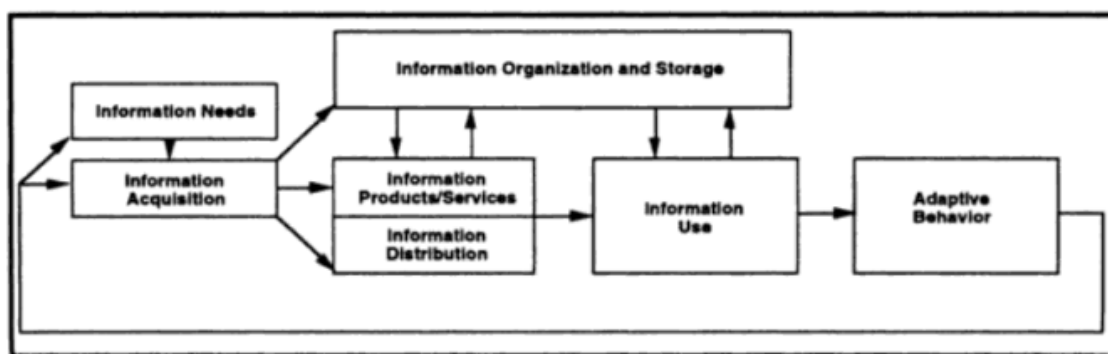


Figura 3 - Ciclo de gestão da informação de Choo.

Fonte: (Choo, 2002)

2.1.5.2. Modelo de Pinto e Malheiro

Pinto e Malheiro (2005) referem que uma “Organização precisa de uma abordagem que congregue, desde a fase de conceção da plataforma tecnológica (hardware e software), até à produção, circulação, avaliação, armazenamento, disponibilização e preservação da informação, toda a Organização e os seus processos de negócio”. Apresentam uma visão holística e sistémica que apoiará a estratégia da organização e “permitirá à organização transformar-se desde logo numa organização <<aprendente>>, a que se seguirá a <<organização inteligente>>”. Para gerir o ciclo de vida da informação (Figura 4) não basta apenas possuir os meios de captura/produção, processamento, armazenamento e disponibilização de informação. É essencial planear a tecnologia que vai suportar a gestão da informação assim como a administração do sistema, a sua segurança e acessos multinível. Tal como conhecer e avaliar os processos e atores produtores de informação e ainda planear a produção de informação, desde as normas de formatos de dados, metodologias de importação/exportação, organização, classificação, descrição e indexação até às políticas de acesso e preservação (Pinto, 2016).

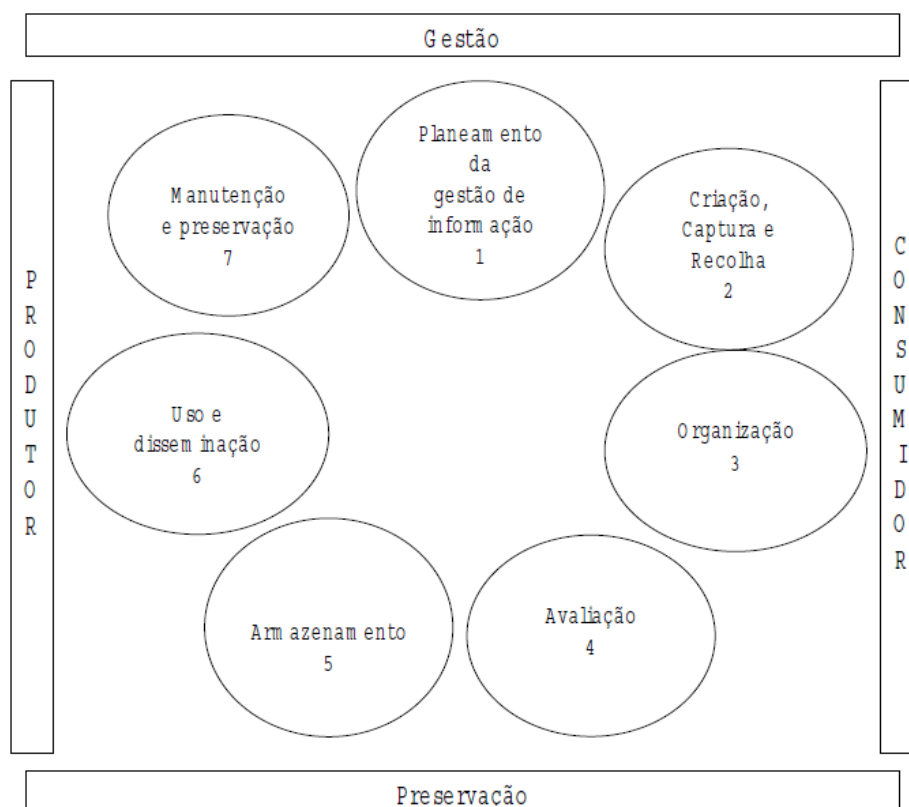


Figura 4 - Ciclo de vida da informação de Pinto e Malheiro.

Fonte: (Pinto & Silva, 2005)

Na Tabela 1 encontra-se uma comparação entre os dois modelos apresentados.

Tabela 1 - Comparação dos ciclos de vida da informação.

	Aborda o ciclo de investigação no modelo proposto	Foca as atividades de gestão	Foca as atividades de curadoria e preservação	Visa a reutilização de dados	Visa melhorar a capacidade de aprendizagem organizacional
(Choo, 2002)		X			X
(Pinto & Silva, 2005)		X	X		X

2.1.6. Ciclos de vida dos dados de investigação

De forma similar ao ciclo da informação, o ciclo de vida dos dados de investigação tenta captar as fases pelas quais os dados de investigação passam num determinado contexto. De seguida apresentam-se alguns ciclos de vida dos dados de investigação.

2.1.6.1. Modelo do Data Documentation Initiative

No contexto do processo de investigação, os dados passam igualmente por um ciclo de vida que reflete as diferentes fases pelas quais os dados de investigação atravessam. O *Data Documentation Initiative* (DDI) apresenta um ciclo de vida (Figura 5) pensado para suportar as atividades de programação e arquivo, típicas da aplicação do modelo de metadados por si desenvolvido. Assim, o modelo foca os dados desde o seu desenvolvimento, codificação, preservação, disseminação até à análise. O modelo tem como objetivo capturar informação sobre a criação dos dados em determinado contexto e possuir informação sobre as estruturas e relações que permitam a descoberta, extração e manipulação dos dados (Structural Reform Group, 2004). O ciclo começa com o planeamento da investigação, envolve planear quais os dados necessários, os métodos de recolha e como vão ser processados, analisados e utilizados. A recolha de dados e o processamento consistem na captura e/ou criação de dados, seguido do seu processamento e análise para dar resposta às questões da investigação. A etapa de arquivo acrescenta valor aos dados ao longo do tempo através da atribuição de metadados, da preservação e acesso a longo termo. Os dados são depois distribuídos de forma direta ou através de uma biblioteca ou arquivo de dados, e passíveis de serem publicados em livros, revistas, páginas web ou em outros serviços. A fase de análise consiste no uso dos dados para fins diferentes do originalmente conceptualizado. No mesmo seguimento, o reaproveitamento envolve a utilização dos dados num contexto diferente, como por exemplo, reestruturação dos dados, integração com dados de conjuntos de dados similares e produção de materiais pedagógicos (Ball, 2012).

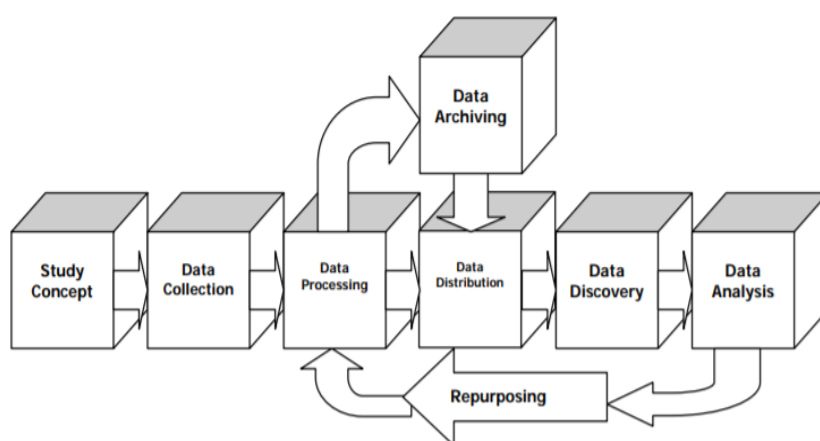


Figura 5 - Ciclo de vida dos dados do Data Documentation Initiative.

Fonte: (Structural Reform Group, 2004)

2.1.6.2. Modelo do DataONE

O ciclo de vida proposto pelo *DataONE* (Figura 6) aborda o ciclo de vida dos dados no contexto da investigação, neste modelo os dados passam pelas fases de planeamento, asseguração, descrição, preservação, descoberta, integração e análise. O planeamento envolve a criação de um plano de gestão de dados, onde se define como os dados vão ser tratados. A coleta de dados implica a recolha de dados através de diversos equipamentos e fontes, bem como a sua organização em documentos ou bases de dados. Os dados passam depois pela revisão ao abrigo de normas, regulamentos e por processos de controlo de qualidade, como por exemplo, verificação de dados duplicados. A fase de descrição implica a agregação de metadados aos dados para facilitar a compreensão e reutilização dos mesmos. A preservação refere-se ao depósito dos dados em centros de dados ou repositórios e a sua proteção contra a degradação e obsolescência dos formatos. Através dos metadados agregados anteriormente e do depósito em repositórios, os dados podem ser descobertos e acedidos. Os dados podem depois ser integrados com outros conjuntos de dados, tendo em conta as suas diferenças semânticas e metodológicas sem perder a sua origem. Por fim, os dados são analisados para responder às necessidades da investigação (Ball, 2012).

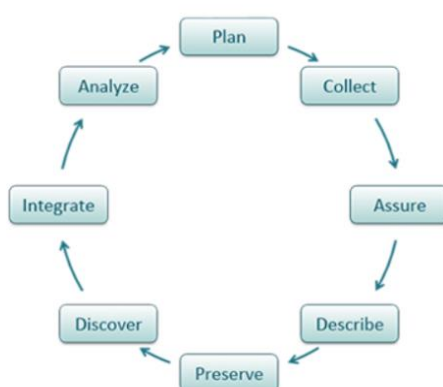


Figura 6 - Ciclo de vida dos dados do DataONE.

Fonte: (DataONE, 2019)

2.1.6.3. Modelo do UK Data Archive

O ciclo de vida proposto pelo *UK Data Archive* (2019) (Figura 7) é constituído pelo planeamento, coleta, processamento e análise, publicação e partilha, preservação e reutilização de dados de investigação. Na fase de planeamento reflete-se sobre a gestão de dados, os métodos de recolha, protocolos e normas, o consentimento para a partilha e as fontes de dados a utilizar. A coleta envolve a recolha de dados e dos seus metadados se possível. O processamento e análise englobam as ações que são realizadas para

processar e analisar os dados durante a investigação. Esta fase incorpora as seguintes ações: introdução, digitalização, transcrição e tradução de dados; verificação, validação; limpeza e anonimização; derivação de dados; descrição e documentação de dados; gestão e armazenamento de dados; análise e interpretação; produção de resultados e citação das fontes de dados. Na publicação e partilha de dados são estabelecidos os direitos autorais dos dados, produzida documentação para os utilizadores e metadados de descoberta, definido o controlo de acesso e a publicação, partilha e promoção dos dados. As ações de preservação englobam a migração dos dados para o formato mais adequado, o armazenamento e realização de backups, criação de documentação de preservação e a preservação e curadoria dos dados. A reutilização dos dados compreende a condução de análises secundárias, o escrutínio e revisão da investigação e o uso dos dados em recursos de aprendizagem.



Figura 7 - Ciclo de vida dos dados do UK Data Archive.

Fonte: (UK Data Archive, 2019)

Este modelo foi adaptado por Vejvoda, Burpee e Lackie (2016) ao ciclo de vida das imagens na investigação. O modelo tem como objetivo fornecer uma conceptualização da gestão de imagens ao longo do ciclo de vida de investigação e recomendar algumas práticas para a sua gestão.

Na Tabela 2 encontra-se representado um resumo das principais características dos ciclos de vida apresentados.

Tabela 2 - Comparação dos ciclos de vida dos dados de investigação.

	Aborda o ciclo de investigação no modelo proposto	Foca as atividades de gestão	Foca as atividades de curadoria e preservação	Visa a reutilização de dados	Visa melhorar a capacidade de aprendizagem organizacional
(Structural Reform Group, 2004)				X	
(DataONE, 2019)	X	X		X	
(UK Data Archive, 2019)	X	X		X	

2.1.7. Modelo de gestão de dados de investigação

O modelo de gestão proposto pelo DCC (Figura 8) tem como objetivo disponibilizar uma abordagem para gerir materiais digitais, para a sua curadoria e preservação desde a conceptualização até à reutilização e preservação a longo termo. O modelo identifica as ações de curadoria, aplicáveis a diferentes níveis de granularidade, que ocorrem ao longo do ciclo, aquelas que têm que ocorrer sequencialmente e as que apenas são realizadas ocasionalmente (Higgins, 2008).

As ações realizadas ao longo de todo o ciclo englobam a descrição e representação da informação, planeamento da preservação, participação e observação da comunidade e curadoria e preservação. A descrição e representação da informação compreende a utilização de normas para assegurar a descrição adequada e controlo ao longo do tempo. Inclui ainda a representação da informação necessária para perceber e apresentar os recursos digitais e metadados associados. O planeamento da preservação inclui os planos de gestão e administração de todas as ações do ciclo. Participação e observação da comunidade diz respeito à monitorização das atividades da comunidade e participação na criação de normas, ferramentas e programas. A curadoria e preservação envolve a conscientização e realização das ações planeadas de curadoria e preservação (Higgins, 2008).

As atividades realizadas sequencialmente abrangem a conceptualização, criação e receção, disposição e seleção, ingestão, ações de preservação, armazenamento, acesso, uso e reutilização e transformação. A conceptualização aborda a conceção e planeamento

da criação dos dados, incluindo os métodos de captura e formas de armazenamento. A criação e receção envolve a criação de dados de acordo com as políticas definidas assim como a captura de metadados administrativos, descritivos, estruturais e técnicos. A avaliação e seleção dos dados para curadoria e preservação a longo termo insere-se na fase de disposição e seleção. Sendo depois transferidos e armazenados de forma segura num arquivo, repositório, centro de dados ou outro tipo custodial, conforme as políticas e requisitos legais existentes. As ações de preservação incluem limpeza de dados, validação, alocação de metadados de preservação, representação da informação e verificação da adequação da estrutura e do tipo de ficheiro dos dados. Assegurar que os dados se encontram acessíveis a utilizadores e reutilizadores, faz parte da fase de acesso, uso e reutilização, onde são definidos o controlo de acesso e os procedimentos de autenticação. A fase de transformação envolve a criação de novos dados a partir dos originais (Higgins, 2008).

As atividades ocasionais englobam a destruição ou transferência para outro tipo de armazenamento dos dados que não foram selecionados para preservação e curadoria a longo termo. As atividades de recuperação, reavaliação e seleção dos dados que anteriormente falharam os procedimentos de avaliação. E por último, a migração dos dados para formatos diferentes, conforme as necessidades de armazenamento ou obsolescência da tecnologia usada (Higgins, 2008).

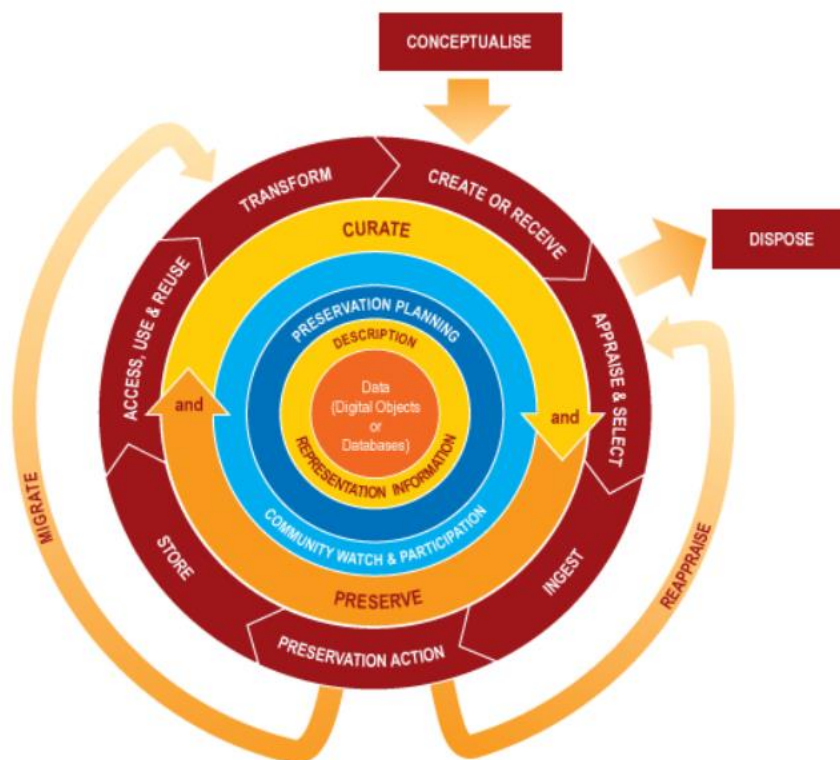


Figura 8 - Modelo de gestão de dados de investigação do Data Curation Centre.

2.2. Imagem como dado de investigação

O desenvolvimento tecnológico, especialmente das tecnologias de informação e comunicação, começou uma nova era na comunicação visual. O rápido desenvolvimento tecnológico e a facilidade de replicação e disseminação, aumentaram o acesso e o valor das coleções de imagens (Enser, 2000).

Esta secção tem como objetivo enquadrar a utilização da imagem como dado em contexto de investigação, começa-se por abordar a natureza dos recursos imagéticos, passando depois para o uso da imagem como evidência científica e o seu papel na ciência.

2.2.1. Natureza dos recursos imagéticos

A imagem pode ser entendida como algo que retrata, consiste numa forma de representação da qual a fotografia, imagens em movimento, desenho e pintura fazem parte, entre muitas outras. A grande diferença que distingue as imagens fotográficas das não fotográficas (desenho e pintura) é o seu método de produção, a fotografia é obtida através de um processo mecânico enquanto o desenho e a pintura tradicionalmente não (Abell, 2010).

A imagem contém informação ou significado de forma diferente do texto. Enquanto o texto comunica através de símbolos convencionados e arbitrários, a imagem comunica através da representação e semelhança dos objetos tais como ele são. O facto da imagem ser utilizada em conjunto com texto sugere que a imagem possui em si informação diferente do texto, conseguindo transmitir coisas que o texto não consegue (Layne, 1994).

Em relação às imagens usadas em contexto de investigação, Dondero (2009) revisita a semiótica de Charles Sanders Pierce para referir que as imagens científicas contém sempre três dimensões, dimensão de indexação, dimensão icónica e dimensão simbólica. A dimensão de indexação perpetua o seu contexto e traços de produção. A dimensão icónica refere-se à estabilização das formas da imagem de objetos científicos, consiste na transformação dos vestígios de fenómenos e dados em objetos científicos concretos através da visualização. Por último, a dimensão simbólica permite que a imagem seja considerada como o produto negociado de uma prática de comunicação negociada, diz respeito ao facto de não podermos estudar e compreender a imagem só com base nela. Tem-se que ter em conta os seus objetivos de produção, as práticas disciplinares da área científica e as relações entre a instituição, políticas e a disciplina.

2.2.2. A imagem como evidência científica

A imagem fornece evidência sobre a existência de determinado objeto e das suas

caraterísticas, apenas quando é produzida por um processo fiável, isto é, o processo que liga a imagem e o objeto deve ser capaz de criar perceções verdadeiras sobre o objeto. O processo deve reproduzir corretamente a informação sobre o objeto e fazê-lo de forma representativa. Para além de ser necessário que o processo de produção de imagens seja fiável, as imagens devem também ser ricas do ponto de vista informacional. A imagem apenas fornece informação sobre as caraterísticas de um objeto quando existe informação necessária para perceber essas mesmas caraterísticas (Abell, 2010).

O processo fotográfico revolucionou as observações realizadas pelos cientistas com o aumento do que podia ser observado, assim como pelo aumento da qualidade do registo das observações. Antes da fotografia, as observações eram registadas manualmente (pintura e desenho) pelo cientista, o que levava anos de treino para produzir registos adequados (Wilder, 2009).

Os instrumentos no decorrer da investigação são utilizados como ferramentas para detetar e registar certas caraterísticas do que se está a investigar. As câmaras fotográficas e outros instrumentos de produção de imagens, possuem a capacidade de detetar e registar as propriedades de objetos. Sendo neste sentido, instrumentos possíveis de serem empregues no processo de investigação. A existência de uma relação de causalidade entre os dados finais e o objeto investigado, faz com que a produção de imagens, como ferramenta de deteção e registo, se equipare aos outros instrumentos. No entanto, a mera deteção não revela de imediato a sua relevância para a investigação, a imagem produzida, por exemplo pela câmara fotográfica, tem que ser analisada pelo investigador à luz da teoria e métodos que guiam a investigação e ser demonstrado que a imagem e o seu processo de obtenção são logicamente independentes do que se pretende comprovar (Möbner, 2013).

Para a imagem ser considerada válida e bem sucedida deve constituir um resultado comensurável do seu processo produtivo, tem em conta as hipóteses, expetativas e a sua necessidade para a investigação (Dondero, 2009).

2.2.3. O uso da imagem na ciência

A primeira área científica a ser vista como a ideal para a utilização de imagens, nesta caso fotografias, foi a Astronomia, embora rapidamente tenha sido adotada por outras áreas científicas como a Medicina (Silva, 2014).

A utilização de fotografias, desenho e outros recursos similares, em áreas como a Antropologia, Geografia ou História, é prática aceite pela comunidade científica (Roger, 2017). Nas Ciências Sociais, as imagens podem ser produzidas pelo investigador com o objetivo de documentar e analisar determinados aspetos sociais e/ou interações sociais. Outro método de investigação utilizado é a recolha e estudo de imagens produzidas ou

usadas pelo objeto de estudo da investigação, ou seja, estudar a forma como o objeto de estudo se relaciona com as imagens (Banks, 2007).

Na História, a utilização de imagens serve, frequentemente, como base para o desenvolvimento do processo de análise e interpretação. Com a digitalização de imagens e a criação de coleções digitais facilmente acessíveis via Web, o historiador passou a ter acesso a imagens que não poderiam ser acedidas no seu formato físico bem como a ferramentas que possibilitam a descoberta de novos detalhes nas imagens (Sandweiss, 2007).

Nas Ciências da Saúde, raios-x e ultrassons, são imagens com alto valor sendo utilizadas como dados clínicos e diagramas anatómicos para comunicar informação crítica (Roger, 2017). A produção de imagens, nesta área, tem um papel importante na clínica, cirurgia, histologia, fisiologia e medicina legal (Silva, 2014).

A imagem no contexto científico pode desempenhar várias funções (Roger, 2017):

- Recrutar participantes – pessoas relutantes em participar numa investigação tendem a entrar mais, quando percebem que as suas imagens ajudam os investigadores a perceber melhor determinado objeto de estudo. Elas sentem que participam, contribuem e interagem com o processo de investigação numa forma que tem valor para elas. Roger refere que segundo a sua experiência, as imagens atuam como uma ponte favorável no recrutamento de vários tipos de participantes, como, homens rurais, indígenas e doentes crónicos;
- Enriquecer a coleção de dados – as imagens ajudam a perceber certos aspetos que não se encontram facilmente perceptíveis. Uma entrevista pode ser enriquecida através do registo em imagem de detalhes que as palavras nem sempre conseguem captar e que são importantes para aprofundar a compreensão e conhecimento de um tema, como emoções ou experiências vividas que poderiam ser perdidas só com texto;
- Produzir conhecimento – as imagens podem funcionar como o elemento base para a criação de novas teorias e conceitos;
- Envolver os gestores, decisores e a comunidade – a publicação das imagens utilizadas durante o processo de investigação, por exemplo através de exposições, pode ser um elemento importante para aumentar a consciência, educação e mudar perceções sobre determinados temas.

As imagens na investigação são dados e consequentemente não devem ser escolhidas e manipuladas de forma inapropriada para servir o propósito da investigação, o que resulta

numa investigação sem valor, desperdício de tempo e dinheiro. As imagens devem ser recolhidas em número suficiente para garantir que o erro na análise da imagem e do seu conteúdo é pequeno. Na criação das imagens é de alta importância verificar se o instrumento de captura de imagem está devidamente calibrado e alinhado com o objeto (Cromey, 2013). Na utilização e manipulação de imagens, especialmente imagens digitais, existem algumas considerações éticas que devem orientar o seu uso (Cromey, 2010):

- As imagens científicas são dados que podem ser comprometidos por manipulações indevidas;
- As manipulações das imagens digitais devem ser sempre realizadas numa cópia da imagem original;
- Ajustes simples em toda a imagem são aceites;
- O recorte da imagem é normalmente aceitável, desde que não altere o conteúdo da imagem;
- As imagens que vão ser comparadas entre si, devem ser capturadas em condições idênticas e submetidas a processos de processamento idênticos;
- Manipulações realizadas apenas em áreas específicas da imagem é questionável;
- O uso de filtros para melhorar a qualidade de imagem não é recomendado;
- Clonar ou copiar objetos de outras imagens e compor uma imagem nova é muito questionável;
- Evitar o uso de formatos de imagem com compressões que afetem a qualidade da imagem;
- As medições da intensidade devem ser realizadas numa imagem uniformemente processada e calibrada segundo normas reconhecidas;
- A resolução e magnificação são importantes;
- Necessário ser cauteloso ao mudar o tamanho, em píxeis, da imagem digital.

2.3. Descrição de dados de investigação

A grande quantidade de dados disponíveis, particularmente em formato digital, em conjunto com o aumento da dependência dos investigadores da produção e análise de grandes quantidades de dados e das políticas a encorajar o livre acesso aos dados de investigação, tornaram os metadados cruciais no processo de curadoria e preservação digital. A capacidade de descoberta e compreensão dos dados por parte dos

investigadores depende da existência de metadados de qualidade, para que possam ser reutilizados, reanalisados e integrados com outras fontes de dados (Day, 2005).

A descrição de dados de investigação através de metadados altamente estruturados é por vezes vista como algo difícil de alcançar, no entanto é necessário para haver reutilização e partilha de dados. Os metadados compõem um conjunto de informações que permitem descrever tanto objetos físicos como objetos digitais (Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011).

A presente secção tem como objetivo contextualizar a importância da descrição de dados de investigação através de metadados, aborda as questões relacionadas com metadados, modelos de metadados no contexto científico e os desafios que a descrição de imagens enfrenta.

2.3.1. Metadados

Os principais benefícios da descrição dos dados de investigação para a ciência passam pela maior longevidade dos dados. Ao serem anexados metadados aos dados de investigação é prolongado o tempo em que os dados têm valor. Outros benefício são o aumento da reutilização e a partilha de dados (Michener, 2006).

A documentação de dados é uma boa prática na gestão de dados de investigação porque permite a contextualização dos dados, através da explicação de como os dados foram criados ou recolhidos, o que significam, qual o seu conteúdo e estrutura, assim como a que processos de manipulação foram sujeitos (Eynden et al., 2011). A atribuição de metadados consiste numa forma altamente estruturada de documentar informação sobre dados. A sua contextualização é essencial para a compreensão e utilização futura. Dados de investigação sem documentação de contextualização não possuem valor, pois não é possível retirar qualquer significado deles (Briney, 2015). Os metadados evoluíram desde inicialmente promoverem a descoberta, aquisição, compreensão e utilização dos dados pelos investigadores até, mais recentemente, possibilitarem a automação da descoberta, ingestão, processamento e análise (Michener, 2006).

Os metadados podem ser caracterizados de acordo com os objetos que caracterizam. Metadados descritivos, identificam e descrevem o conteúdo do objeto, como por exemplo, título, autor, descrição, língua e palavras-chave. Metadados técnicos, referem-se ao modo de criação ou recolha dos dados e capturam as características técnicas dos formatos, alguns exemplos são tipo de ficheiro, tamanho do ficheiro e instrumento de captura ou criação. Metadados estruturais, codificam a estrutura interna dos documentos ou as relações estruturais com outros objetos, por exemplo, número de páginas, secções, documentos relacionados. Metadados de preservação, recolhem informações que são necessárias para as atividades de preservação e curadoria.

Metadados de proveniência, descrevem modificações que foram realizadas nos dados sendo importantes em processos de auditoria. Um exemplo deste tipo de metadado é a versão do ficheiro. Por fim, metadados de direitos descrevem as restrições/licenças ou a ausência delas a que a utilização dos dados se encontra sujeita, neste tipo incluem-se a licença de utilização, direitos de autor e detentor dos direitos (Treloar & Wilkinson, 2008).

Willis, Greenberg e White (2012) através dos estudo de modelos de metadados específicos para diferentes domínios, elaboraram uma lista de requisitos fundamentais dos esquemas de metadados para documentar dados científicos:

- Abstração – promover através da abstração a captura de novas necessidades e atualização do esquema de metadados;
- Extensibilidade, flexibilidade e modularidade – salvaguardar a longevidade do modelo, facilitar a sua adoção, modificação e extensão a necessidades que ainda não foram identificadas;
- Compreensão e suficiência – definir o conjunto de elementos de forma a que sejam compreensíveis e capturem os elementos essenciais para documentar o domínio;
- Simplicidade – ter em conta os diferentes níveis de especialização e oferecer suporte através de ferramentas e recursos que ajudem à sua utilização;
- Troca de dados – possuir a capacidade de fomentar a troca, partilha e comunicação de dados entre a comunidade científica;
- Recuperação – ser possível descobrir e aceder aos dados através dos metadados, tendo em conta, os padrões de acesso de domínios específicos;
- Arquivo – possibilitar a preservação a longo prazo da informação contida nos dados;
- Publicação – acautelar a associação dos dados a publicações científicas.

A descrição dos recursos que suportam a investigação é um processo complexo devido à diversidade da tipologia dos dados de investigação, os objetivos com que são recolhidos e a sua utilização. A existência de diversos domínios científicos e de conceitos que a eles são específicos, contribuem também para o aumento da complexidade da descrição. Muitas vezes é necessário a colaboração entre investigadores e gestores de dados de investigação para aferir os requisitos necessários para a descrição das coleções de dados de determinado domínio científico (Castro, Silva, & Ribeiro, 2014; Willis et al., 2012).

Outro aspeto que demonstra a complexidade do processo de descrição é o custo de produção dos metadados. O momento de captura dos metadados implica uma análise do seu valor e benefícios para os seus utilizadores. Quanto maior for o detalhe e a estruturação dos metadados, maior é o custo, especialmente em relação ao consumo de tempo (Michener, 2006). A ideia de capturar a maior quantidade de metadados possível e depois retirar os benefícios da grande quantidade de metadados recolhidos tem que ser ponderada. A recolha de metadados deve começar pelo reconhecimento da sua utilidade e pela análise custo/benefício (Treloar & Wilkinson, 2008).

2.3.2. Modelos de Metadados

Os modelos de metadados para coleções de dados de investigação contêm informação essencial sobre os autores, contexto, informação temporal, localização e qualidade dos dados. Estes elementos formam uma rede interrelacionada de nós que apoiam a descoberta, seleção e acesso para os utilizadores dos dados e a organização, gestão e preservação para os gestores dos dados de investigação. Os modelos em conjunto com componentes tecnológicos, políticas, regras e boas práticas de utilização de metadados constituem a infraestrutura base para aplicação de metadados (Qin & Li, 2013).

Os modelos de metadados podem ser de descrição genérica, ou seja, podem ser aplicados a diferentes domínios, ou modelos específicos de aplicação apenas numa área científica. Estes últimos fornecem melhor documentação e facilitam o acesso a dados de determinados domínios científicos, através da sua incorporação em repositórios temáticos e em outras ferramentas (Willis et al., 2012).

A aplicação de um único modelo de metadados pode não corresponder totalmente às necessidades de descrição num contexto específico. Nestes casos realiza-se a recolha e seleção de descritores de um ou mais modelos para a construção de um esquema combinado, sem perder a interoperabilidade dos esquemas base. O resultado final é a aplicação de um perfil caracterizado por permitir especificar o vocabulário controlado, formato e relações dos elementos, refinar definições das normas utilizadas e combinar diferentes modelos para especificar o seu uso e assim suportar uma aplicação, função, comunidade e ambiente particular (Castro et al., 2014; Heery & Patel, 2000; Duval, Hodgins, Sutton, & Weibel, 2002).

De seguida, são apresentados alguns modelos de metadados potencialmente relevantes para a descrição de dados de investigação em formato de imagem.

2.3.2.1. Dublin Core

O *Dublin Core Metadata Initiative* (2019) é um modelo de metadados de aplicação genérica, surgiu em 1995 num encontro em Dublin, Ohio, inicialmente desenhado para facilitar a descoberta de informação na web, através da introdução de metadados nas

páginas. Os desenvolvimentos levaram-no a caminhar tendo em vista a interoperabilidade, característica dos princípios da web semântica e dados ligados. Utiliza *Uniform Resource Identifiers* (URI) para identificar os objetos descritos pelos metadados e os termos usados nos vocabulários. A versão mais recente dos descritores que constituem o Dublin Core data de Julho de 2012, alterando a especificação de 2010 usada até então.

O modelo é formado por um conjunto de quinze elementos genéricos, “*Creator*”, “*Contributor*”, “*Publisher*”, “*Title*”, “*Date*”, “*Language*”, “*Format*”, “*Subject*”, “*Description*”, “*Identifier*”, “*Relation*”, “*Source*”, “*Type*”, “*Coverage*” e “*Rights*”, que podem ser combinados com elementos de outros esquemas através de *Resource Description Framework* (RDF) e assim formar vocabulários diversos.

2.3.2.2. Common European Research Information Format

O *Common European Research Information Format* (CERIF) é a norma de metadados que a União Europeia recomenda aos seus membros para documentar a investigação. Permite a descrição das entidades envolvidas na investigação, atividades, contexto e resultados bem como grande flexibilidade com relações semânticas formais. Isto fomenta o controlo de qualidade, arquivo, acesso, troca de informação e conhecimento entre os decisores, investigadores, avaliadores, gestores de investigação, editores e, por fim, o público em geral.

O CERIF é representado através de um modelo de base de dados relacional tendo como suporte o modelo entidade-relação. Atualizações recentes ao modelo adicionaram uma camada semântica formal e formatação em *Extensible Markup Language* (XML). O modelo é composto pelas entidades principais (“*Person*”, “*OrganisationUnit*” e “*Project*”), entidades resultado (“*ResultPublication*”, “*ResultPatent*” e “*ResultProduct*”), entidades de segundo nível (“*Funding*”, “*Facility*”, “*Equipment*”, “*Prize*”, “*CV*”, “*Expertise*”, “*Qualification*”, “*Citation*”, “*Metrics*”, “*Event*”, “*PostalAddress*” e “*ElectronicAddress*”), entidades de ligação, especificações multilingues e especificações semânticas (Jörg, 2010; euroCRIS, 2019).

2.3.2.3. Exchangeable Image File Format

O *Exchangeable Image File Format* (EXIF) é um formato normativo para ficheiros de imagens digitais capturadas por câmara digitais que usa vocabulário controlado para representar metadados administrativos da imagem. Foi desenvolvido pela *Japan Electronics and Information Technology Industries Association* (JEITA) e atualmente encontra-se na versão 2.2. Tem como objetivo capturar e fornecer informação útil sobre os atributos da imagem, processamento e informação de localização (Tešić, 2005). Permite ainda a organização das imagens e possibilita a pesquisa de atributos da

imagem. A norma apenas suporta os formatos TIFF e JPEG e especifica a estrutura dos dados da imagem, os campos utilizados pela norma bem como a definição e gestão das versões dos formatos (Japan Electronics and Information Technology Industries Association, 2002).

O sistema da imagem pressuposto na norma engloba equipamentos, aplicações comerciais e aplicações especializadas. Os equipamentos incluem aparelhos de captura de imagem como câmaras digitais. As aplicações comerciais referem-se a programas genéricos que permitem o processamento de imagem. As aplicações especializadas englobam programas especializados em processamento de imagem (Japan Electronics and Information Technology Industries Association, 2002).

Esta norma permite recolher informação sobre o ficheiro de imagem, lente utilizada para capturar, dados sobre a imagem, dados pessoais e direitos da imagem.

Por vezes, o EXIF é utilizado em conjunto com o *Extensible Metadata Platform* (XMP), uma tecnologia que permite a criar, processar e embeber metadados em ficheiros digitais de vídeo, áudio, imagem, entre outros.

2.3.2.4. Modelos de metadados específicos

Apresentam-se agora alguns modelos de metadados utilizados em áreas científicas específicas. Estes modelos foram selecionados pois são amplamente conhecidos e utilizados em repositórios de referência desses domínios.

O DDI é uma norma de descrição de dados, com base na linguagem de anotação XML aplicado nas ciências sociais. O DDI surgiu da necessidade existente de documentar e comunicar melhor as características dos dados empíricos produzidos, recolhidos ou compilados pelos investigadores. A sua especificação faz com que seja compatível e complementar a outros modelos de metadados como o Dublin Core, MARC, ISO 11179, SDMX e ISO 19115. A norma encontra-se atualmente na versão 3.2, desde Março de 2014, sendo constituído por 17 módulos com vários termos: “*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*” e “*Study Unit*” (DDI Alliance, 2019; Vardigan, Heus, & Thomas, 2008).

O *Directory Interchange Format* (DIF) é uma norma descritiva de dados científicos do domínio das ciências da Terra, desenvolvido em finais de 1980 com a finalidade de documentar, trocar informação e possibilitar a descoberta, acesso e uso de dados do domínio referido e outros relacionados. O DIF10, versão atualmente recomendada, faz

parte das normas base do *Unified Metadata Model* (UMM) da NASA. Permite descrever os instrumentos de captura de dados, características espaciais e temporais bem como aspetos relacionados com o projeto com o qual os dados se encontram associados (Directory Interchange Format, 2018; NASA, 2019).

O *Access to Biological Collection Data* (ABCD) é um modelo de descrição, baseado em XML, para o acesso e troca de dados primários de biodiversidade (observações e espécimes), tenta ter uma estrutura compreensível, estruturada e compatível com outros modelos para o suporte da descrição de dados em várias bases de dados. A atual especificação do modelo foi revista em 2007, sendo que os seus termos dividem-se nos seguintes grandes grupo: associação, biótipo, unidade de jardim botânico, unidade de coleção cultural, conjunto de dados, recolha, unidade de herbário, identificação, objeto multimédia, unidade paleontológica, unidade de recursos genéticos das plantas, sequência, sinecologia, unidade e contato do conteúdo (Access to Biological Collection Data task group, 2007).

O *Ecological Metadata Language*, mais conhecido por EML, consiste numa especificação de metadados com o objetivo de descrever dados de investigação para que os investigadores possam documentar, preservar e partilhar os seus dados e resultados. O seu principal foco são os dados obtidos através de observações na área da ecologia, ciências do ambiente e da Terra. Define um vocabulário compreensível e passível de ser lido através da linguagem de anotação XML, contém módulos para identificar e citar pacotes de dados vistos como essenciais, descrever a extensão temporal, espacial, taxonómica e temática, documentar os métodos e protocolos de investigação, descrever estruturas de dados e conteúdos complexos e anotar os dados de investigação de forma precisa com vocabulário semântico. (Jones et al., 2019; Feigraus, Andelman, Jones, & Schildhauer, 2005).

Na Tabela 3 encontra-se representado um resumo dos modelos de metadados.

Tabela 3 - Comparação dos modelos de metadados.

Modelo de Metadados	Abrangência	Aplicação
Access to Biological Collection Data	Específico (Biodiversidade)	Dados de investigação
Common European Research Information Format	Genérico	Dados de investigação
Data Documentation Initiative	Específico (Ciências Sociais)	Dados de investigação
Directory Interchange Format	Específico (Ciências da Terra)	Dados de investigação
Dublin Core	Genérico	Web
Ecological Metadata Language	Específico (Ecologia, Ciências do Ambiente e da Terra)	Dados de investigação
Exchangeable Image File Format	Específico	Imagem digital

2.3.3. Desafios na descrição da imagem

O rápido desenvolvimento tecnológico, em particular no armazenamento e telecomunicações, assim como na facilidade com que as imagens são incorporadas em aplicações que permitem a exibição, publicação e disseminação de imagens, levou ao aumento das necessidades informacionais dos utilizadores em relação às imagens (Chen & Rasmussen, 1999).

A indexação de imagens evoluiu também com o desenvolvimento tecnológico. Anteriormente, a indexação estava dependente da descrição realizada pelo humano através de descritores textuais e classificadores, apoiados em vocabulários controlados. Os esquemas de indexação eram criados, na grande maioria, localmente ou para refletir as características de coleções de imagens específicas. Algumas tentativas de criar esquemas genéricos de descrição de imagens foram criados como, por exemplo, o *Art and Architecture Thesaurus* (AAT) e o *Library of Congress Thesaurus of Graphic Materials* (LCTGM). Atualmente, tem-se verificado a automação do processo de recuperação e indexação, devido ao crescente aumento de soluções baseadas no conteúdo da imagem (Goodrum, 2000; Zhou, Li, & Tian, 2017).

Os objetivos da indexação de imagens passam por fornecer acesso às imagens tendo em conta os seus atributos e fornecer acesso a grupos de imagens, não só a imagens individuais (Layne, 1994). A descrição de imagens e a sua recuperação traz consigo o conhecimento de que interpretar signos visuais e imagens não é igual à interpretação de texto e linguagem, especialmente quando as imagens não são acompanhadas por um

texto que aponte o significado da imagem (Rafferty, 2019). As principais abordagens à indexação e recuperação de imagens são a baseada em conceitos e a baseada em conteúdo (Chen & Rasmussen, 1999; Enser, 2000).

A recuperação baseada em conceitos usa descrição textual para recuperação de imagens. Pode utilizar interrogações textuais, vocabulários controlados, linguagem natural e legendas. Pode ser puramente descritiva, subjetiva ou abstrata (Chen & Rasmussen, 1999). A recuperação baseada em conteúdo utiliza análise computacional da imagem para extrair automaticamente características como a cor, textura e forma, localmente ou globalmente, em imagens digitais (Chen & Rasmussen, 1999; Goodrum, 2000). A indexação baseada em conceitos fornece uma imagem de alto nível do seu conteúdo, mas a sua implementação é muito dispendiosa, especialmente em tempo, e sofre com a falta de consistência na indexação devido à subjetividade na interpretação das imagens (Chen & Rasmussen, 1999). A indexação baseada no conteúdo é menos dispendiosa a implementar, mas possui alguns desafios técnicos relativos à representação e organização da imagem bem como à medição de similaridade da imagem (Zhou et al., 2017).

A descrição de imagens realizada por humanos implica o conhecimento de alguns pressupostos nomeadamente:

- As imagens são ricas, podem conter informação útil para investigadores de diferentes áreas (Besser, 1990);
- As imagens são, por vezes, utilizadas para finalidades diferentes das antecipadas pelo criador (Besser, 1990);
- Capturar em palavras o conteúdo ou significado de uma imagem é um desafio intelectual significativo (Enser, 2000);
- A mesma imagem pode ter significado distinto para diferentes pessoas (Chen & Rasmussen, 1999);
- A imagem pode ser genérica ou específica, o que esta representa pode ser útil ao nível de identidade genérico ou específico (Layne, 1994);
- Pode ser difícil chegar a um consenso sobre o conteúdo e significado da imagem, assim como sobre os atributos apropriados para descrever a imagem (Chen & Rasmussen, 1999);
- O indexador tem de seleccionar aspetos abstratos partilhados por ele e pelo utilizador para fornecer descritores adequados. Os modelos mentais do indexador ou do utilizador podem ser descritivamente ricos ou pobres

dependendo da situação (Heidorn, 1999);

- Dificuldades em fazer corresponder os termos empregues pelo utilizador para descrever a sua necessidade de informação com os existentes no vocabulário controlado (Gordon, 2001).

Igualmente, a descrição de imagens realizada automaticamente implica também conhecimento de alguns aspetos-chave nomeadamente:

- A lacuna semântica, ou seja, a dificuldade em traduzir características visuais de baixo nível em conceitos semânticos de alto nível (Zhou et al., 2017);
- Dificuldades em identificar elementos que não estão diretamente visíveis na imagem nomeadamente, tempo, espaço, eventos e significado, conceitos abstratos e emotivos bem como características não desejadas (Enser, Sandom, Hare, & Lewis, 2007);
- A lacuna de intenção, dificuldade que o utilizador tem em expressar de forma precisa o conteúdo visual que pretende recuperar (Zhou et al., 2017);
- Os humanos segmentam as imagens melhor que as máquinas devido à capacidade de associar elementos, adquirida ao longo dos anos (Datta, Joshi, Li, & Wang, 2008);
- As técnicas de compreensão da imagem enfrentam ainda alguns desafios relativos à velocidade de execução, confiabilidade da segmentação e robustez na classificação de imagens semelhantes (Datta, Li, & Wang, 2005);
- As técnicas que permitem a automação podem ser computacionalmente complexas, o que torna a sua aplicação difícil em situações reais (Datta et al., 2005).

Para mitigar alguns desafios e melhorar a recuperação de imagens tem-se verificado a introdução de folksonomias e *tags* na descrição das imagens, onde os utilizadores atribuem livremente palavras-chave às imagens. Esta abordagem permite a adição de um vocabulário particular, por vezes muito diferente daquele que o indexador atribui no mesmo contexto (Ménard, 2010). A classificação social apesar de não constituir uma solução para os desafios existentes na descrição da imagem, oferece uma nova oportunidade de melhorar a indexação de imagens e a interação com os utilizadores. A abordagem tradicional à indexação é complementada por esta. A classificação tradicional oferece mais consistência, nível de especificação similar, interoperabilidade e acesso uniformizado através de normas de metadados e vocabulários controlados. A classificação social traz a possibilidade de os utilizadores descreverem na sua língua

materna, perspectivas novas, especialização, interação com os utilizadores e uma indexação orientada para o utilizador (Matusiak, 2006). Tem-se também verificado a emergência de sistemas de recuperação de imagens híbridos que envolvem características e funcionalidades da recuperação baseada em conceitos e em conteúdos (Enser, 2000).

Na descrição de imagens torna-se importante encontrar um equilíbrio entre quais as imagens que devem ser indexadas individualmente ou coletivamente, qual o nível e de que forma os atributos das imagens devem ser diferenciados e qual a profundidade de indexação desejável (Layne, 1994).

3. Metodologia

A presente dissertação teve como objetivos, estudar o uso e produção de imagem em contexto de investigação bem como recomendar boas práticas para o uso e produção de imagens em projetos de investigação. Para concretizar o primeiro objetivo foi realizado um estudo exploratório para conhecer os processos relativos aos uso e produção de imagens na investigação. Para realizar este estudo recorreu-se a duas ferramentas distintas, porém complementares, o inquérito por questionário e por entrevista. O estudo exploratório ajuda a perceber melhor um problema quando não existe informação disponível sobre o mesmo. Assim, esta abordagem propicia a realização de uma investigação preliminar para auxiliar a reflexão sobre os processos de gestão de dados de investigação em formato de imagem.

Numa primeira fase elaborou-se o inquérito por questionário e, posteriormente, realizaram-se entrevistas com alguns investigadores. A aplicação do questionário é adequada quando se quer conhecer e analisar os comportamentos de uma população, neste caso, investigadores. Os dados recolhidos através deste instrumento possibilitam a realização de quantificações. O inquérito por entrevista permite analisar com uma maior profundidade os elementos de análise recolhidos e conhecer o sentido que os entrevistados dão às suas práticas (Quivy & Campenhoudt, 1995).

O segundo objetivo foi alcançado através da elaboração de um guia de orientação para a gestão de imagens na investigação e de um modelo de metadados para auxiliar os investigadores a descreverem as imagens utilizadas. O guia de orientação e o modelo foram suportados pela análise dos dados recolhidos através do questionário e das entrevistas, em conjunto com as leituras realizadas.

Depois de serem elaboradas as recomendações, o modelo de metadados inserido neste âmbito foi avaliado por investigadores. Esta avaliação teve como finalidade recolher opiniões sobre a utilização do modelo e adequação dos descritores. Na Figura 9 podemos ver a organização geral de todo o trabalho com uma visão de alto nível.

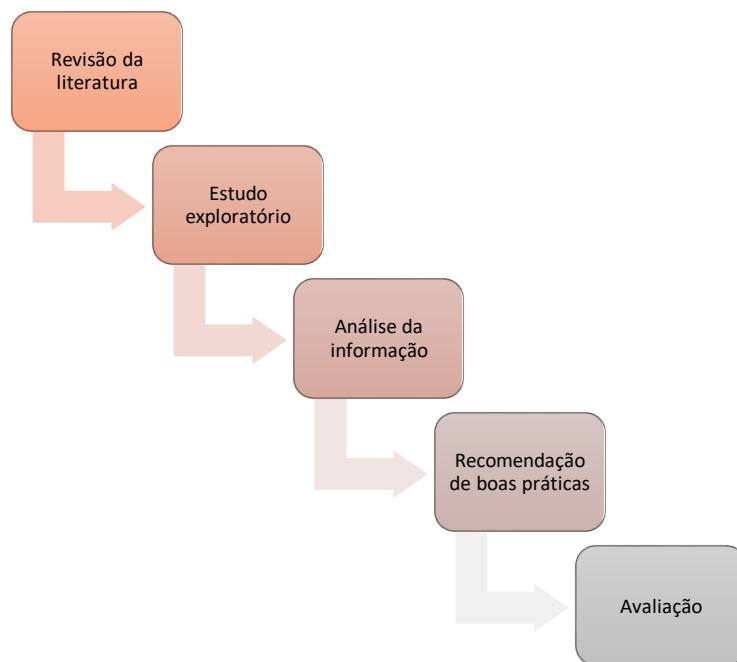


Figura 9 - Organização geral do trabalho.

3.1. Inquérito por Questionário

A realização do questionário teve como objetivo estudar os procedimentos e hábitos na utilização e produção de imagens em contexto de investigação. A Figura 10 mostra a abordagem metodológica do questionário. Esta consistiu em três fases principais: elaboração, disseminação e análise. A primeira compreendeu a estruturação e elaboração do questionário. A segunda diz respeito à divulgação do questionário. Por fim, a terceira etapa envolveu a organização dos dados recolhidos e a respetiva análise.

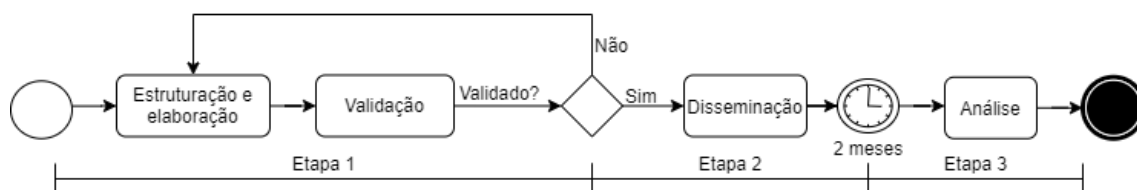


Figura 10 - Processo de desenvolvimento do questionário.

A estrutura do questionário e a elaboração das questões (Anexo I) foram baseadas pelo ciclo de vida da informação e dos dados de investigação apresentados na revisão da literatura. As perguntas foram agrupadas por etapas: planeamento, criação/compilação, asseguarção da qualidade, processamento/análise, descrição, armazenamento e partilha.

O questionário passou por várias revisões, onde a adequação e formulação das questões foram discutidas e refletidas. Antes da divulgação, o questionário foi testado com um investigador para analisar como era interpretada cada pergunta. O objetivo foi garantir que as questões eram bem compreendidas por investigadores não especializados no

assunto.

O questionário começou a ser difundido a 20 de fevereiro de 2020 e aceitou respostas até 26 de março de 2020. O questionário foi distribuído por e-mail na comunidade de investigação da Universidade do Porto², nomeadamente o i3S e Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC-TEC), e em unidades de investigação externas à Universidade do Porto (UP) que obtiveram financiamento em 2019 da Fundação para a Ciência e Tecnologia³. Estas unidades foram o Centro de Matemática e Aplicações da Universidade Nova de Lisboa, Centro Cardiovascular da Universidade de Lisboa, Centro de Informática e Sistemas da Universidade de Coimbra e o Centro de Estudos Filosóficos e Humanísticos da Universidade Católica. Para as unidades externas, foi enviado um *email* aos coordenadores de cada grupo a solicitar a divulgação do questionário entre os seus membros.

Os inquiridos foram pessoas com experiência em investigação, nomeadamente estudantes de doutoramento, bolsiros e professores universitários. As respostas foram tratadas anonimamente e as perguntas não permitiam a identificação dos participantes.

Na análise, métodos quantitativos foram utilizados para analisar as perguntas fechadas e a análise de conteúdo usada para as questões de resposta aberta. O questionário foi composto por dois tipos de resposta fechada. Um recolhia respostas numa escala de Likert de cinco pontos (Nunca, Raramente, Ocasionalmente, Frequentemente e Sempre), o que dava origem a variáveis ordinais. O outro tipo, envolvia a seleção, ou não, de opções de resposta fornecidas, o que criava variáveis nominais.

Para cada questão, foi analisada a tendência geral do conjunto de respostas e comparadas entre quatro áreas científicas (Ciências da Vida e da Saúde – CVS; Ciências Exata e Engenharia – CEE; Ciências Naturais e do Ambiente – CNA; Ciências Sociais e Humanidades – CSH) e entre três frequências de uso da imagem na investigação (Baixo (B) – menos de 25% dos projetos de investigação; Moderado (M) – em cerca de 50% dos projetos de investigação; Elevado (E) – em mais de 75% dos projetos de investigação). Para as variáveis ordinais, foi utilizado o teste de hipóteses *Kruskal-Wallis* para analisar diferenças entre os grupos. No caso de se detetar diferenças, era aplicado o *Pairwise Test Mann-Whitney with Bonferroni correction* para identificar os locais onde existiam diferenças. Para as variáveis nominais, foi usado o teste de *Chi-square* para testar a igualdade de proporções.

² <https://www.i3s.up.pt/>; <https://www.inesctec.pt/en>

³ <https://www.cisuc.uc.pt/>; <http://ccul.pt/>; <https://cefh.braga.ucp.pt/>; <https://www.cma.fct.unl.pt/>

3.2. Inquérito por Entrevista

As entrevistas tiveram como objetivo aprofundar as questões sobre a utilização e produção de imagens e complementar os requisitos para o modelo de metadados. Decidiu-se fazer entrevistas semi-estruturadas pela maior flexibilidade e adaptação da entrevista mediante as respostas obtidas. Esta atividade dividiu-se em quatro etapas principais (Figura 11): estruturação, realização das entrevistas, transcrição e análise.

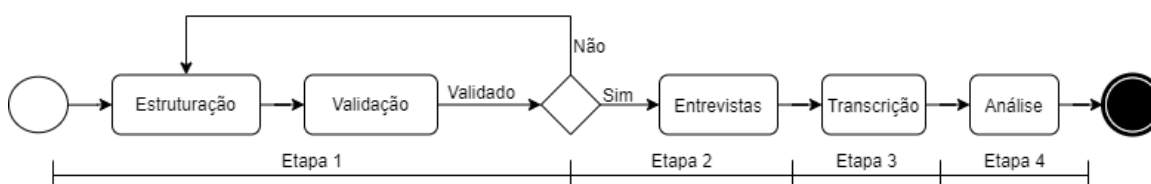


Figura 11 - Metodologia das entrevistas.

O primeiro passo foi a estruturação do guião das entrevistas. Tal como o questionário, as entrevistas foram estruturadas com base nos ciclos de vida apresentados na revisão da literatura. Neste guião (Anexo II) evitou-se realizar questões já existentes no questionário, o foco foi questionar os investigadores sobre aspetos da gestão das imagens que não tinham sido abordados previamente. Como estas questões necessitavam de uma resposta mais longa, não foram inseridas no questionário.

No total foram realizadas oito entrevistas a investigadores de diferentes domínios de investigação da comunidade da UP. Dos entrevistados, quatro pertencem às Ciências Exatas e Engenharia, dois às Ciências da Vida e da Saúde e outros dois às Ciências Naturais e do Ambiente. Escolheram-se diferentes domínios de investigação pois a utilização de imagens não é igual em todas as áreas científicas, permitindo assim perceber as diferenças e semelhanças entre elas.

O número inicial de entrevista planeadas era de dezasseis entrevistas, quatro investigadores por cada domínio de investigação (CVS, CEE, CNA e CSH). Porém, este número de entrevistas não foi atingido devido à fraca adesão e disponibilidade dos investigadores, devido à pandemia de covid-19. Outro motivo foi o facto de se não conhecer um grande número de investigadores que utilizassem imagens. O domínio das CSH não se encontra representado pois nenhum investigador contactado mostrou disponibilidade em participar.

Apesar de o número de entrevistas não ser inicialmente planeado, as entrevistas realizadas permitiram obter informação sobre possíveis diferenças na utilização de imagens entre investigadores da mesma e de diferentes áreas científicas.

As entrevistas foram realizadas através de uma plataforma digital e gravadas com a devida permissão dos investigadores para serem analisadas posteriormente. Dois

investigadores não tiveram disponibilidade para responder às questões na plataforma e, por isso, responderam por escrito. Antes de se proceder à análise foi necessário transcrever o conteúdo das entrevistas. Na análise dos dados recolhidos utilizou-se o método qualitativo, mais concretamente análise de conteúdo. A análise começou por uma leitura inicial das entrevistas, seguida por uma segunda leitura e sinalização das partes mais importantes. Para cada pergunta foi elaborada uma tabela com o essencial das respostas de cada investigador.

3.3 Recomendação de boas práticas

As recomendações aqui desenvolvidas tiveram como objetivo contribuir para o fomento de boas práticas no âmbito da gestão de imagens em contexto de investigação. Estas recomendações assumiram a forma de um modelo de metadados e de um guia de orientações para a gestão de imagens. Estas recomendações constituem duas ferramentas importantes para acompanhar e auxiliarem os investigadores nas diferentes etapas da gestão das imagens.

A metodologia de desenvolvimento das recomendações começou com a análise dos dados recolhidos através do questionário e das entrevistas. Estes dados permitiram perceber a familiarização dos investigadores com a gestão de imagens bem como os seus processos e hábitos. Foram ainda realizadas algumas pesquisas e leituras, que também serviram como suporte às recomendações.

De seguida, procedeu-se ao desenvolvimento do modelo de metadados e, posteriormente, ao guia de orientação. No capítulo dedicado a cada recomendação, pode-se encontrar mais informação relativa ao seu desenvolvimento.

Depois de finalizadas as recomendações, elaborou-se uma pequena tarefa de descrição para os investigadores avaliarem o modelo de metadados desenvolvido. Esta avaliação recolheu informação sobre a usabilidade e utilidade do modelo. Devido a limitações relacionadas com o tempo, apenas foi possível realizar a avaliação do modelo de metadados.

4. Utilização e produção de imagens em investigação - estudo via inquérito por questionário

Na presente secção são apresentados os resultados obtidos através do questionário. Os resultados são apresentados de acordo com as fases do ciclo de vida dos dados de investigação. É utilizado o * para indicar resultados significantes com $\alpha=0,05$ e ** para indicar resultados significantes com $\alpha=0,01$.

O questionário recebeu respostas de 107 investigadores. Destes, 41 (38.3%) são das Ciências da Vida e da Saúde, 30 (28%) das Ciências Exatas e Engenharia, 12 (11.2%) das Ciências Naturais e do Ambiente e 24 (22.4%) das Ciências Sociais e Humanidades. A Figura 12 relaciona o uso de imagens como dado de investigação com o domínio de investigação. Como se pode observar existe maior tendência para o uso de imagens como dado de investigação nas Ciências da Vida e da Saúde.

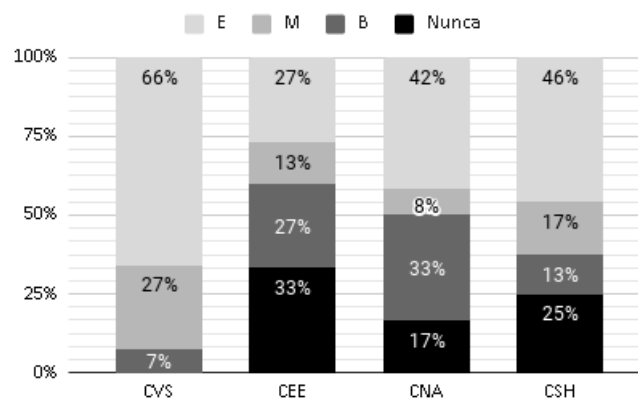


Figura 12 - Frequência de utilização de imagens por área científica.

4.1. Planeamento

Os investigadores, quando questionados sobre a existência de um documento para orientar o uso e a produção de imagens durante a investigação, 69,7% responderam nunca ou raramente, 16,9% ocasionalmente e apenas 13,5% responderam frequentemente ou sempre. Mesmo os investigadores mais acostumados a usar imagens (grupo elevado) raramente elaboraram este documento, como pode ser visto na Figura 13. Não foram encontradas diferenças significativas entre as áreas científicas ou a frequência de uso da imagem.

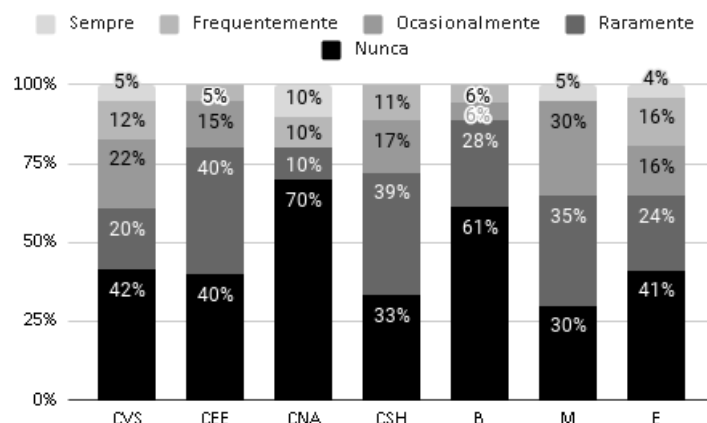


Figura 13 - Elaboração de um guia de orientação por domínio científico e frequência de uso.

Quanto à maneira e à frequência com que os investigadores adquirem e produzem imagens, a maioria produz imagens sempre ou frequentemente (73%), embora cerca de 59,5% dos investigadores também considere que ocasional ou frequentemente usam imagens de projetos anteriores. De realçar que 89,9% dos entrevistados afirmaram que nunca compraram imagens a terceiros e 61,8% deles afirmaram que raramente ou nunca adquirem imagens a terceiros, mesmo imagens sem custo associado.

O domínio das Ciências da Vida e Saúde está associado a uma produção significativamente maior de imagens no contexto de investigação, quando comparado com as Ciências Naturais e do Ambiente e as Ciências Sociais e Humanidades. Isso está visível na Tabela 4, que mostra as diferenças significativas entre as áreas científicas na fase do planeamento.

Tabela 4 - Resultados significativos no planeamento por área científica.

Variável	Descrição	CVS>	CEE>	CNA>	CSH>
Ordinal	Imagem produzida no contexto da investigação	CNA* CSH*			

Da mesma forma, aqueles com baixa frequência de uso de imagens, produzem menos imagens que os investigadores com uso moderado e elevado de imagens na investigação. Isso é visível na Tabela 5, que mostra as diferenças significativas entre a frequência de uso da imagem.

Tabela 5 - Resultados significativos no planeamento por frequência de uso da imagem.

Variável	Descrição	E>	M>	B>
Ordinal	Imagem produzida no contexto da investigação	B**	B**	
Ordinal	Imagem advém de investigações anteriores	B**		

4.2. Criação/Compilação

As imagens são, sempre ou frequentemente, produzidas por computador (51,7%), seguido pelo microscópio (42,7%), câmara do telemóvel (31,5%) e câmara fotográfica tradicional (21,3%). Nas Figuras 14 e 15, podemos ver a percentagem de utilização destes instrumentos por área científica e frequência de uso.

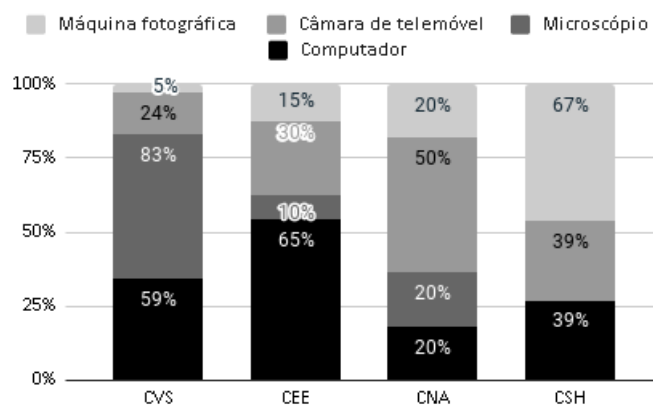


Figura 14 - Utilização de instrumentos de captura por área científica.

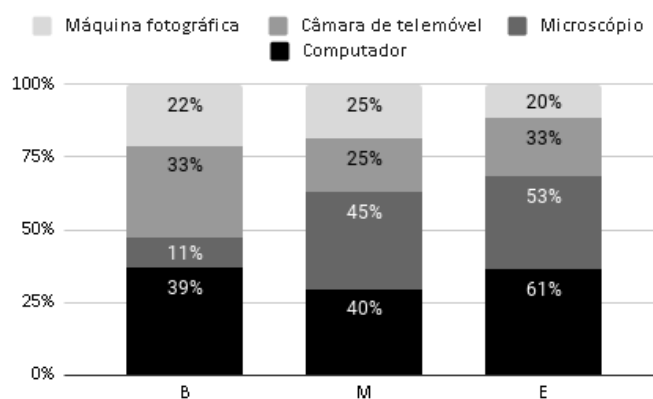


Figura 15 - Utilização de instrumentos de captura por frequência de uso.

Em relação a diferenças significativas entre as áreas científicas (Tabela 6), os investigadores das Ciências Sociais e Humanidade usam significativamente mais a câmara fotográfica tradicional do que os das Ciências Exatas e Engenharias e Ciências da

Vida e da Saúde. Estes últimos são quem menos usa este instrumento em comparação com os investigadores das Ciências Naturais e do Ambiente. De forma semelhante, os investigadores das Ciências da Vida e da Saúde são quem mais usa o microscópio em comparação com as restantes áreas científicas.

Tabela 6 - Resultados significativos na criação/compilação por área científica.

Variável	Descrição	CVS>	CEE>	CNA>	CSH>
Ordinal	Imagem capturada via câmara tradicional			CVS*	CVS** CEE**
Ordinal	Imagem capturada via microscópio	CEE** CSH** CNA**			
Ordinal	Realização de edições simples na imagem	CNA*			

Na Tabela 7, podemos ver que quem está menos habituado a usar imagens na investigação, usa menos o microscópio para produzir imagens.

Tabela 7 - Resultados significativos na criação/compilação por frequência de uso.

Variável	Descrição	E>	M>	B>
Ordinal	Imagem capturada via microscópio	B**	B*	
Ordinal	Realização de edições simples na imagem	B**	B**	

Em relação à edição/manipulação das imagens, 65,2% dos investigadores responderam que recortam a imagem sempre ou frequentemente, 47,2% fazem sempre ou frequentemente ajustes simples em propriedades como o contraste, brilho e saturação e 49,4% alteram as dimensões da imagem, sempre ou frequentemente. Cerca de 59,6% dos investigadores raramente ou nunca combinam objetos de diferentes imagens para criar uma nova e 53,9% raramente ou nunca usam filtros para melhorar a qualidade da imagem.

A Figura 16 relaciona a edição de imagens com o domínio de investigação. Em relação a diferenças significativas, os investigadores da Ciências Naturais e Ambientais realizam significativamente menos ajustes simples nas imagens, comparativamente aos investigadores das Ciências da Vida e Saúde (Tabela 6).

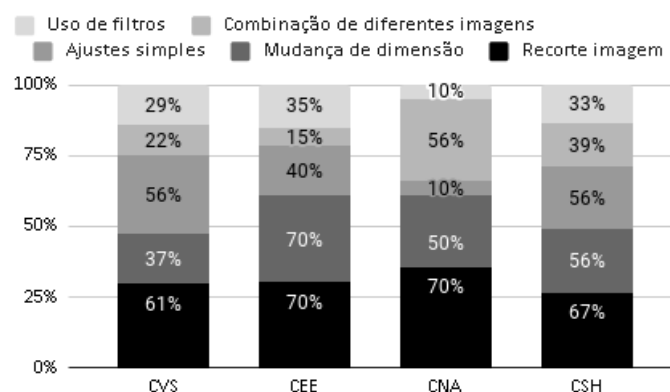


Figura 16 - Edição de imagem por área científica.

A Figura 17 relaciona a edição de imagens com a frequência de uso. Investigadores que usam imagens com menos frequência, realizam significativamente menos edições simples em comparação com os restantes investigadores (Tabela 6).

Dos inquiridos que afirmaram fazer edições nas imagens, 57,6% indicaram que preservam sempre a imagem original e 23,5% preservam frequentemente. Apenas 2,4% dos investigadores nunca preservam a imagem original. Não foram encontradas diferenças significativas entre as áreas científicas e a frequência de uso da imagem.

Por sua vez, a documentação das edições é feita sempre ou frequentemente por 34,1% dos inquiridos, sendo que 47,1% raramente ou nunca documenta. Também aqui não foram observadas diferenças significativas na relação entre as áreas científicas e a frequência de uso da imagem.

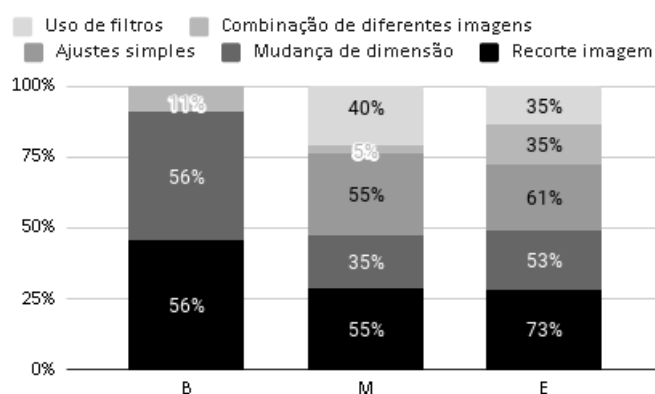


Figura 17 - Edição de imagem por frequência de uso.

4.3. Assegurar a qualidade

Em relação à qualidade das imagens, 53,9% dos inquiridos responderam que verificam sempre ou frequentemente a qualidade das suas imagens. Apenas 7,9% dos investigadores disseram que nunca o fazia e 14,6% apontaram que raramente verificam

a qualidade.

Os investigadores das Ciências Exatas e Engenharia verificam a qualidade das imagens significativamente menos do que os investigadores das Ciências da Vida e Saúde (Tabela 8).

Tabela 8 - Resultados significativos na asseguuração da qualidade por área científica.

Variável	Descrição	CVS>	CEE>	CNA>	CSH>
Ordinal	Verificação da qualidade das imagens	CEE*			

O mesmo acontece nos investigadores que usam imagens com menos frequência. Estes verificam a qualidade das imagens significativamente menos do que os investigadores mais acostumados a usar imagens (grupo de elevado uso) (Tabela 9).

Tabela 9 - Resultados significativos na asseguuração da qualidade por frequência de uso.

Variável	Descrição	E>	M>	B>
Ordinal	Verificação da qualidade das imagens	B*		

Os processos mencionados pelos investigadores para garantir a qualidade das imagens são variados. No entanto, destacam-se: revisão de imagens e suas propriedades (36%), calibração de instrumentos (23,3%) e descarte de inadequadas (9%). Não encontramos diferenças significativas em relação à área científica e à frequência de uso da imagem.

4.4. Processamento/Análise

Quando questionados sobre os programas de computador mais utilizados no processamento e análise de imagens, os investigadores mencionaram principalmente o *ImageJ* (33,7%), *Photoshop* (19,7%), *Paint* (10,1%), *Cell Profiler* (9%) e linguagens de programação (7,9%). Cerca de um quarto (24,7%) dos investigadores não usa nenhum programa de computador. Os investigadores das Ciências da Vida e da Saúde são quem mais usa de forma significativa o programa *ImageJ* para analisar as suas imagens em comparação com as restantes áreas científicas (Tabela 10). Os investigadores com baixa frequência de utilização de imagens usam de forma significativa o programa *Paint* quando comparado com quem usa de forma moderada e elevada as imagens na investigação (Tabela 11).

Pode-se dizer que as imagens são objeto de diferentes tipos de análise. Aqueles que mais se destacam são a análise de conteúdo (37,1%), cálculos matemáticos (18%), quantificações (11,2%) e medições (5,6%). Cerca de 20,2% dos inquiridos não analisa as

imagens. Os investigadores das Ciências Sociais e Humanidades são quem menos analisa a imagem quando equiparados às Ciências da Vida e Saúde e às Ciências Exatas e Engenharia (Tabela 10).

Por fim, foi observado que os investigadores tendem a combinar análise manual e automática. Há um equilíbrio entre o modo manual (38,2%, sempre ou frequentemente) e o modo automático (42,7%, sempre ou frequentemente). Na Tabela 10, podemos ver que a análise automática é significativamente mais alta nos domínios das Ciências da Vida e da Saúde e das Ciências Exatas e da Engenharia. Por outro lado, investigadores das Ciências Sociais e Humanidades são os que utilizam menos a análise manual em comparação com as Ciências da Vida e da Saúde.

Tabela 10 - Resultados significativos no processamento/análise por área científica.

Variável	Descrição	CVS>	CEE>	CNA>	CSH>
Nominal	Utilização do ImageJ	CSH** CNA** CEE**			
Nominal	Não analisa a imagem				CVS** CEE*
Ordinal	Análise manual da imagem	CSH*			
Ordinal	Análise automática da imagem	CSH** CNA**	CSH** CNA*		
Ordinal	Documentação da análise da imagem	CNA* CSH** CEE*			

Os investigadores que raramente usam imagens, usam significativamente menos modos automáticos e manuais para analisar imagens em comparação com aqueles que usam muito as imagens (Tabela 11).

Tabela 11 - Resultados significativos no processamento/análise por frequência de uso.

Variável	Descrição	E>	M>	B>
Nominal	Utilização do Paint			E* M*
Ordinal	Análise manual da imagem	B*		
Ordinal	Análise automática da imagem	B*		
Ordinal	Documentação da análise da imagem	B**	B**	

Cerca de 46% dos investigadores apontaram que nunca ou raramente documentam a análise das imagens. Os investigadores nas Ciências da Vida e Saúde são os que documentam com maior frequência os passos dados na análise das imagens em comparação com as outras áreas científicas (Tabela 10). Os investigadores com baixa frequência de uso das imagens, documentam a análise significativamente menos do que os investigadores com uso mais intensivo (grupos de uso moderado e elevado), como pode ser visto na Tabela 11.

4.5. Descrição

Existe uma ligeira tendência para associar anotações a imagens individuais (44,9% sempre ou frequentemente), em vez de as associar a um conjunto de imagens (37,1% sempre ou frequentemente). Não foram verificadas diferenças significativas em relação à área científica. No entanto, investigadores com uma baixa frequência de uso da imagem, não anotam conjuntos de imagens com a mesma frequência que aqueles que usam muito imagens (Tabela 12).

Tabela 12 - Resultados significativos na descrição por frequência de uso.

Variável	Descrição	E>	M>	B>
Ordinal	Anotação de conjunto de imagens	B*		

No que diz respeito ao suporte onde as anotações são feitas, destaca-se o uso de outros documentos que não a imagem. Um total de 32 (35,9%) inquiridos disseram que escrevem sempre ou frequentemente num documento em papel, cerca de 36 (40,5%) investigadores anotam num documento digital. Apenas 16 (18%) investigadores anotam na imagem, sempre ou frequentemente. Investigadores das Ciências Exatas e Engenharia são os menos propensos a tomar notas num documento em papel que não a imagem em comparação principalmente com Ciências da Vida e Saúde, mas também com Ciências Sociais e Humanidades. Estes dois últimos têm a maior percentagem de investigadores que anotam num documento em papel, 56,1% e 27,8% respetivamente (Tabela 13). Não há uma clara preferência no suporte de anotação em relação à frequência de uso de imagens em contexto de investigação.

Tabela 13 - Resultados significativos na descrição por área científica.

Variável	Descrição	CVS>	CEE>	CNA>	CSH>
Ordinal	Anotação das imagens num documento em papel	CEE**			CEE*
Nominal	Uso do autor como descritor	CEE*		CEE*	
Nominal	Uso da descrição como descritor	CEE*			
Nominal	Uso do instrumento de captura como descritor	CEE** CSH**			
Nominal	Uso da metodologia como descritor	CSH**			
Nominal	Uso da amostra como descritor	CEE** CSH**			

De realçar que 65 (83,1%) dos entrevistados afirmaram que nunca ou raramente usam uma aplicação que os ajude a descrever as imagens. Apenas 6 investigadores (6,7%) disseram que usam, sempre ou frequentemente, uma aplicação para os auxiliar nesta tarefa. Não foram verificadas diferenças significativas em relação à área científica e à frequência do uso da imagem.

Em relação às normas de metadados utilizadas na descrição da imagem, fica claro que poucos investigadores usam e conhecem o tópico. Das três opções fornecidas no questionário (Dublin Core, CERIF e EXIF), nenhuma apresentou percentagem de uso acima de 1,1% para “sempre” e acima de 6,7% para “frequentemente”. Por sua vez, a percentagem média para “nunca” é de 85,7%. Quando posta a possibilidade de apresentarem normas diferentes das do questionário, apenas 5 investigadores apresentaram uma resposta. O que mostra uma clara tendência para a não utilização de normas de metadados para a descrição das imagens. Não foram encontradas diferenças significativas em relação à área científica e à frequência de uso da imagem.

Quando questionados sobre o vocabulário/elementos de descrição que os investigadores usam e consideram relevantes, as respostas variam. Os mais relevantes são os elementos “título” (68,5%), “autor” (53,9%), “data” (57,3%) e “descrição” (51,7%). Os descritores que se seguem são os menos vistos como relevantes, “direitos de utilização” (39,3%), “formato” (40,4%) e “instrumento de captura” (41,6%). Entre os descritores mais utilizados, estão “título” (83,1%), “autor” (62,9%), “data” (62,9%) e “descrição” (64%). Com as menores percentagens de utilização, estão os descritores “direitos de utilização” (24,7%), “formato” (39,3%) e “instrumento de captura” (39,3%).

As Figuras 18 e 19 mostram os descritores utilizados e considerados relevantes por área científica.

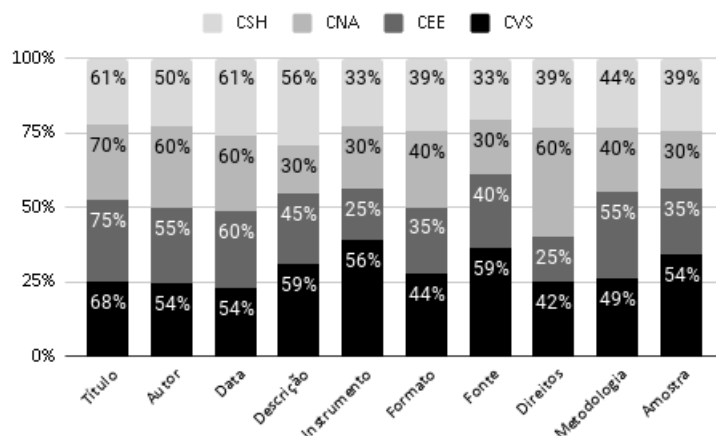


Figura 18 - Descritores relevantes por área científica.

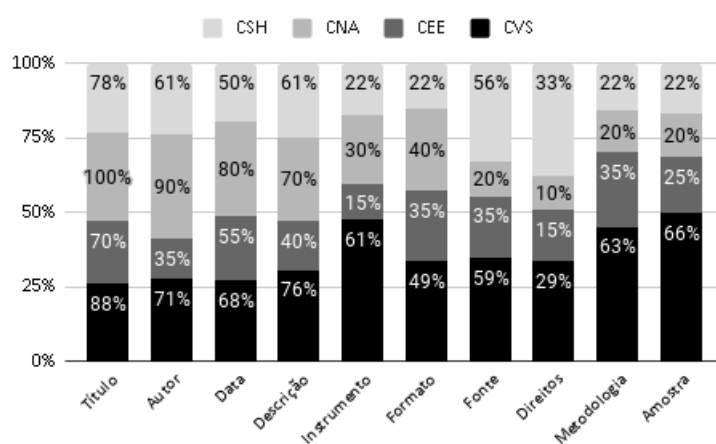


Figura 19 - Descritores utilizados por área científica.

Investigadores das Ciências Exatas e Engenharia são os menos propensos a usar o “autor” em comparação com as Ciências da Vida e Saúde e Ciências Naturais e do Ambiente, bem como o descritor “descrição” em comparação com os das Ciências da Vida e Saúde. Na direção oposta, é mais provável que os investigadores das Ciências da Vida e da Saúde usem o descritor “instrumento de captura” e “amostra” em comparação com os domínios das Ciências Exatas e Engenharia e Ciências Sociais e Humanidades. As Ciências da Vida e da Saúde também têm maior probabilidade de usar o descritor “metodologia” do que as Ciências Sociais e Humanas (Tabela 13).

Em relação à relevância atribuída aos descritores, não foram encontradas diferenças significativas entre a área científica e a frequência de uso de imagens como dado de investigação.

4.6. Armazenamento

Em relação ao local de armazenamento das imagens, destaca-se o computador (95,5% sempre ou frequentemente), seguido pelo disco externo (73%), *pen drive* (41,5% sempre

ou frequentemente) e *cloud* (40,4% sempre ou frequentemente). Não se verificaram diferenças significativas entre a área científica e a frequência de uso da imagem.

Os formatos mais usados para armazenar são TIFF (74,2%), JPEG (83,1%) e PNG (61,8%). Embora com percentagens de uso muito baixas, os formatos RAW (13,5%), BMP (7,9%), SVG (2,2%) e PDF (3,4%) também são usados. A Figura 20 mostra a distribuição do formato de armazenamento por domínio científico.

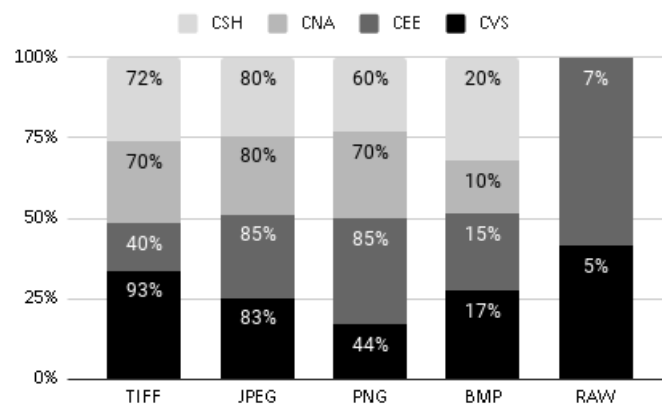


Figura 20 - Formato de armazenamento por área científica.

Em relação às diferenças significativas, os investigadores das Ciências Exatas e da Engenharia são os que menos usam o formato TIFF para armazenar as suas imagens quando comparados com os das Ciências da Vida e da Saúde. Similarmente, investigadores das Ciências da Vida e Saúde são menos propensos a usar o formato PNG do que os das Ciências Exatas e de Engenharia (Tabela 14).

Tabela 14 - Resultados significativos no armazenamento por área científica.

Variável	Descrição	CVS>	CEE>	CNA>	CSH>
Nominal	Armazenamento das imagens em TIFF	CEE**			
Nominal	Armazenamento das imagens em PNG		CVS*		

A Figura 21 mostra a preferência do formato de armazenamento por frequência de uso.

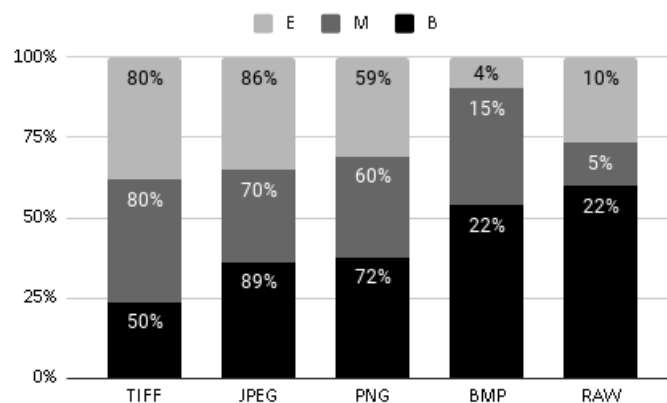


Figura 21 - Formato de armazenamento por frequência de uso.

Em relação às diferenças significativas, os investigadores que raramente usam imagens têm menos probabilidade de usar o formato TIFF para armazenar imagens em comparação com aqueles que usam muito frequentemente imagens (Tabela 15).

Tabela 15 - Resultados significativos no armazenamento por frequência de uso.

Variável	Descrição	E>	M>	B>
Ordinal	Armazenamento das imagens em TIFF	B*		

Em relação ao volume de imagens armazenadas durante um projeto de investigação, cerca de 58,4% dos investigadores disseram que não eram capazes de quantificá-lo. Não se verificaram diferenças significativas em relação à área científica e à frequência de uso da imagem.

Quando questionados sobre a frequência com que realizam cópias de segurança, 40,4% responderam mensalmente, 15,8% semanalmente e 19,7% diariamente. Cerca de 18% dos investigadores não fazem cópias de segurança regularmente e 19,1% não fazem cópias de segurança. Quando os investigadores realizam cópias de segurança, estas costumam ser feitas de forma manual (65,2%), apenas 13,5% responderam que eram feitas automaticamente.

Nestas duas perguntas, não foram encontradas diferenças significativas em relação à área científica e à frequência do uso da imagem.

4.7. Partilha

O momento de partilha ocorre principalmente no final da investigação (61,8%, sempre ou com frequência) ou durante a investigação (59,5%, sempre ou com frequência). Cerca de 41,6% dos investigadores responderam que partilham, frequentemente, as imagens

algum tempo após o projeto de investigação acabar. Nesta questão, não encontramos diferenças significativas em relação à área científica e à frequência de uso de imagem.

Para a resposta sobre quem promove a partilha de imagens, a resposta mais frequente foi o investigador principal (71,9%), seguido pelo investigador produtor da imagem (44,9%), instituição (31,5%) e entidade financiadora (11,2%). Na Figura 22 pode-se observar a distribuição do promotor da partilha de imagens por área científica.

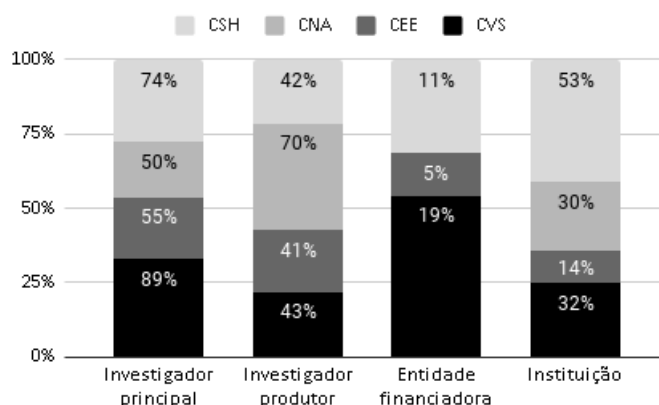


Figura 22 - Promotor da partilha de imagens por área científica.

O investigador principal é quem promove a partilha de imagens de forma mais significativa nas Ciências da Vida e da Saúde em comparação com as Ciências Exatas e Engenharia (Tabela 16). Não foram encontradas diferenças significativas em relação à frequência de uso da imagem.

Tabela 16 - Resultados significativos na partilha por área científica.

Variável	Descrição	CVS>	CEE>	CNA>	CSH>
Nominal	Partilha é promovida pelo investigador principal	CEE**			
Ordinal	Inclusão do local de armazenamento nos artigos científicos		CVS**		CVS*

Em relação aos direitos de utilização, os investigadores responderam com mais frequência que deixam as imagens “restritas no acesso e uso” (35,4% sempre ou frequentemente), seguidas pelo “acesso e uso livre” (27,7% sempre ou frequentemente) e “acesso livre, mas uso restrito” (20,7 % sempre ou com frequência).

Quando questionados sobre se partilhavam as imagens em arquivos de imagens, repositórios institucionais ou temáticos, a resposta dos investigadores foi nunca ou raramente (85%). Cerca de 4,5% responderam que partilham as imagens em artigos científicos. Nestas duas questões, não foram encontradas diferenças estatisticamente

significativas entre áreas científicas e frequências de uso.

A maioria dos investigadores (77,5% nunca ou raramente) não menciona o local onde as imagens são armazenadas nos artigos científicos. Apenas 14,6% responderam que mencionam, sempre ou frequentemente, o local de armazenamento nos artigos científicos. Os investigadores das Ciências da Vida e Saúde são menos propensos a mencionar o local em que as imagens são armazenadas nos artigos científicos, quando comparados principalmente com os das Ciências e Engenharia Exatas, mas também com os das Ciências Sociais e Humanas (Tabela 15). Surpreendentemente, os investigadores que utilizam imagens em mais de 75% de seus projetos de investigação são os menos propensos a mencionar o local onde as imagens são armazenadas (Tabela 17).

Tabela 17 - Resultados significativos na partilha por frequência de uso.

Variável	Descrição	E>	M>	B>
Ordinal	Inclusão do local de armazenamento nos artigos científicos			E*

5. Utilização e produção de imagens em investigação - estudo via inquérito por entrevista

Nas entrevistas participaram oito investigadores de diferentes domínios científicos. Os investigadores encontram-se identificados pela área científica a que pertencem mais um algarismo. Estas entrevistas tiveram como objetivo perceber melhor a forma como os investigadores utilizam as imagens nos seus projetos de investigação. Todos os investigadores entrevistados pertencem à comunidade de investigadores da UP.

O investigador CEE1 pertence às Ciências Exatas e da Engenharia, a sua investigação foca principalmente a ciência de dados, mais concretamente a análise temporal de dados ambientais. É especialista em análise de registos climáticos e dados de satélite para quantificar alterações climáticas. Possui também experiência na monitorização de radioatividade ambiental e na análise e interpretação da variação da radiação entre o Sol e a Terra, entre outras interações.

O investigador CEE2 encontra-se ligado às Ciências Exatas e da Engenharia. A sua investigação versa sobre o desenvolvimento e estudo de aplicações relacionadas com a engenharia de sistemas de empresariais.

O investigador CVS1 é da área das Ciências da Vida e da Saúde. Os seus projetos de investigação estão relacionados com a descoberta de genes passíveis de estarem envolvidos com a resistência do cancro da mama e implicações éticas na edição do genoma.

O investigador CVS2 é também das Ciências da Vida e da Saúde. Presentemente encontra-se a estudar aplicações termográficas na experimentação animal. Outros tópicos das suas investigações incluem bem-estar animal, ética animal e aspetos legais, sociais e científicos na investigação animal.

O trabalho do investigador CEE3 enquadra-se nas Ciências Exatas e da Engenharia. A sua investigação está relacionada com sistemas de informação e computação gráfica, mais concretamente com a área de *User Interface/User Experience Design*.

O investigador CNA1 pertence ao domínio das Ciências Naturais e do Ambiente. A sua investigação relaciona-se com a biodiversidade e ecologia. Atualmente encontra-se a estudar redes ecológicas entre comunidades de plantas no contexto das alterações no uso da terra.

O investigador CEE4, investigador das Ciências Exatas e da Engenharia, encontra-se ligado à engenharia de sistemas empresariais. A sua investigação versa sobre a aplicação de tecnologia para a valorização eficiente, mobilização e aproveitamento de recursos

florestais.

A investigação do investigador CNA2 encontra-se inserida no âmbito das Ciências Naturais e do Ambiente. O tópico geral dos seus projetos de investigação visa a compreensão de como as alterações ambientais afetam a biodiversidade, em paisagens onde a floresta e a agricultura são dominantes no uso da terra.

Como já foi mencionado no capítulo relativo à metodologia, nenhum investigador das Ciências Sociais e Humanidades mostrou disponibilidade para participar nas entrevistas.

5.1. Planeamento

A primeira pergunta colocada aos investigadores pedia que descrevessem o ciclo de vida das imagens durante o processo investigativo.

Os investigadores mencionaram que o ciclo de vida das imagens na investigação depende do tipo de imagem que se está a utilizar e do seu contexto de produção. As etapas que constituem o ciclo de vida das imagens são diferentes de investigador para investigador, mesmo em investigadores do mesmo domínio científico. Os processos mais comuns são a produção/recolha, processamento/análise e armazenamento. Poucos investigadores mencionaram a realização de tarefas de planeamento, verificação da qualidade e descrição das imagens.

Quando questionados sobre se costumavam produzir um documento que orientasse a produção e utilização de imagens ao longo da investigação, sete investigadores responderam que não e um investigador que sim.

O investigador CEE1 respondeu nunca que tinha sentido necessidade de produzir tal documento e o investigador CVS2 apontou motivos culturais, não é costume nos projetos de investigação realizar este documento. O investigador CVS1 conhece algumas recomendações de revistas e ferramentas de acesso aberto, focadas principalmente na criação de metadados. O investigador CNA1 e CNA2 produzem um documento de metadados, onde gerem todos os dados utilizados durante a investigação.

O investigador CEE4 referiu que é costume produzir um documento para assegurar que a equipa sabe o que está delineado, o que está aprovado e que, em suma, serve como um guia para todos os elementos da equipa saberem o que fazer em relação às imagens.

Estes resultados estão em linha com os obtidos no questionário, mostrando uma tendência para a não realização do planeamento da utilização de imagens nem o seguimento de orientações.

Dos oito investigadores, seis responderam que gostariam de ter recomendações para a utilização e produção de imagens durante a investigação. Um respondeu talvez e outro

que não.

Os investigadores mencionaram que era útil ter um documento que orientasse de forma clara os investigadores a proactivamente prepararem os dados para publicar. Referiram também que teriam um impacto positivo no planeamento, produção e utilização das imagens. As recomendações seriam importantes para salvaguardar as características de utilização das imagens, pois assim seria mais fácil e rápido avaliar se uma imagem possui as características que necessita. Para facilitar a compreensão das recomendações, é importante que estas sejam acompanhadas por exemplos.

5.2. Criação, recolha e compilação

Em relação ao processo de criação, recolha e compilação de imagens, as respostas obtidas foram variadas. Os processos de produção de imagens são heterogéneos mesmo entre investigadores do mesmo domínio científico. Normalmente, esta tarefa é constituída pela questão de investigação e produção/recolha das imagens, para posteriormente serem analisadas. Apenas dois investigadores mencionaram que nesta fase costumam realizar procedimentos para retificar possíveis defeitos das imagens. De forma semelhante, só um investigador costuma documentar o processo de criação das imagens.

Quando comparados com os resultados obtidos no questionário, os resultados aqui obtidos acompanham a tendência da não documentação do processo de produção das imagens. Também aqui, os investigadores costumam produzir as imagens no âmbito de projeto de investigação.

5.3. Assegurar a qualidade

Quando questionados sobre se costumavam realizar atividades para assegurar a qualidade das imagens, quatro investigadores responderam sim e quatro investigadores responderam que não costumam realizar nenhuma atividade para assegurar a qualidade das imagens. O investigador CNA1 referiu que normalmente não realiza, apenas realiza quando não consegue perceber a imagem original. Para o investigador CEE4 não existe verificação, se a imagem está visualmente apelativa e reflete o que se quer transmitir, é aprovada.

O investigador CVS1 referiu que costuma tratar o ruído existente nas imagens, fazer ajustes simples como, por exemplo, corte da imagem e atenção ao contraste. Estes processos são principalmente realizados em imagens resultantes do método *western blot*, um método utilizado para identificar proteínas.

O investigador CVS2 para assegurar a qualidade das suas imagens costuma verificar e, por vezes, explicar as escalas. A nível estético, verifica a fonte e tamanho das imagens.

O investigador CNA2 verifica sempre as resoluções espacial e temporal das imagens, por

vezes, quando é necessário georreferencia a imagem.

Os resultados obtidos nesta secção encontram-se em linha com os do questionário, na medida em que se verificou que os processos de verificação da qualidade são variados.

5.4. Processamento/Análise

Relativamente à pergunta sobre o processo de extração de informação das imagens, quatro investigadores responderam que não analisavam a imagem.

O investigador CEE costuma analisar de forma visual as imagens e, por vezes, mede distâncias entre determinados pontos da imagem.

O investigador CVS1 referiu que os processos de processamento e análise dependem do investigador, do que este considera importante e do tipo de imagem que é produzida.

O investigador CVS2 começa por visualizar as imagens, depois analisa-as de acordo com os métodos escolhidos para a investigação. No fim reporta a análise que efetuou.

O investigador CNA2 afirmou que o processamento e análise da imagem está dependente do que se pretende fazer e atingir com a imagem. Alguns dos métodos que aplica são classificação semiautomática das imagens e processos geoestatísticos. Pode ocorrer extração de informação direta ou indireta da imagem.

Tal como no questionário, as respostas dos investigadores permitiram perceber que os métodos de processamento/análise são diversos. Com as entrevistas conseguiu-se perceber algumas razões para que assim seja. O processamento/análise é influenciado pelo investigador, tipo de investigação e resultado que se espera atingir.

5.5. Descrição

Em relação à descrição das imagens, três investigadores não costumam descrever. No decorrer das entrevistas, observou-se que a maioria dos investigadores não se encontra sensibilizada para a utilização de modelos de metadados na descrição das imagens. Apenas três investigadores assinalaram a utilização de metadados, no entanto não mencionaram nenhum modelo de metadados. O mesmo foi apurado na análise das respostas do questionário.

Cerca de três investigadores mencionaram que a descrição da imagem era influenciada pela imagem e seu contexto de produção, pelo investigador e público-alvo. Foi mencionado que quando a imagem é apresentada a pessoas com conhecimento do contexto da imagem, esta é feita com mais detalhes técnicos. Quando a descrição é direcionada a pessoas sem conhecimento do contexto da imagem, a descrição foca detalhes genéricos.

Na Tabela 18 encontram-se os descritores utilizados ou considerados importantes pelos

investigadores na descrição das imagens. Nesta tabela, sobressaem os descritores “autor”, “título” e “fonte”. Estes resultados são similares aos obtidos no questionário, o “autor” e “título” são também dois dos descritores mais utilizados.

Tabela 18 - Descritores mencionados pelos entrevistados.

Investigador	Resposta
CEE1	Fonte; Direitos de utilização
CEE2	Autor; Local de origem
CVS1	Características de obtenção e processo de tratamento da imagem; Legenda; Escala; Recursos relacionados
CVS2	Legenda; contexto (metodologia de produção)
CEE3	Título; Autor; Fonte.
CNA1	Fonte; Autor; Descrição; Data; Método de produção; Processos de tratamento e análise
CEE4	Contexto
CNA2	Como foi gerada; métodos utilizados; resolução espacial e temporal

Todos os investigadores, excetuando o investigador CEE3, responderam que um modelo de metadados para a descrição das imagens traria benefícios e teriam interesse em utilizar.

Os investigadores mencionaram que o modelo os iria ajudar na descrição e trazia a capacidade de guardar informação sobre a imagem de forma sintetizada e padronizada. A existência de um modelo transversal a grupos de investigação multidisciplinares foi considerada benéfica, pois contribui para harmonizar e criar pontos de comunicação comuns entre diferentes domínios de investigação.

Outro aspeto abordado pelos investigadores, foi o facto de ser importante o modelo poder ser adaptado mediante o tipo de investigação e a área científica. Por um lado, o modelo devia ser genérico e ter aplicabilidade geral. Por outro lado, devia ter em atenção detalhes específicos relacionados com o tipo de imagem, contexto, área científica que podem ser importantes para a descrição da imagem.

5.6. Armazenamento

Em relação à pergunta sobre se era costume existirem orientações para a organização das imagens, todos os investigadores responderam que tal prática não era costume. Os investigadores apontaram que tinham um método próprio de organização dos dados de

investigação, mas que não seguiam recomendações formalizadas. Normalmente as imagens costumam estar armazenada no computador e organizadas por pastas de acordo com o projeto de investigação correspondente. Os inquiridos do questionário também elegeram o computador como o local de armazenamento mais habitual.

Quando questionados sobre se nos projetos de investigação era usual existir um local central para armazenar as imagens, quatro investigadores responderam que não era usual.

O investigador CVS2 respondeu que dependia da dimensão do projeto, se o projeto for grande é costume as imagens encontram-se numa plataforma *cloud* para todos acederem. Em projetos de pequena dimensão, as imagens encontram-se todas no computador.

O investigador CEE3 mencionou que se encontravam armazenadas num computador que era acessível pela rede interna da instituição. O investigador CNA1 também referiu o computador e acrescentou que eram armazenadas num disco externo, por vezes as imagens estão nestes dois dispositivos ao mesmo tempo.

O investigador CNA2 referiu que, no seu caso concreto, tinha começado a depositar todos os seus dados numa *network attached storage*, um sistema de armazenamento de dados em rede.

5.7. Partilha

Em relação à partilha das imagens, apenas o investigador CEE1 respondeu que normalmente não partilha, apenas quando necessita de mostrar algo nos artigos ou os colegas pedem. O investigador CNA1 referiu que de momento não estava a partilhar, mas que iria dar esse passo para outros investigadores reutilizarem as suas imagens.

As imagens costumam ser partilhados com colegas do grupo/instituição onde desenvolvem investigação, quando os colegas pedem ou o investigador sabe que podem ser úteis para o trabalho que estes se encontram a desenvolver. A motivação é a partilha de informação, realização de colaborações e obtenção de opinião sobre a imagem.

Dos oito investigadores que participaram nas entrevistas, sete investigadores responderam que nunca tinham depositado as suas imagens num repositório ou arquivo de imagens. O investigador CEE2 foi o único a responder positivamente, afirmou que colocou a imagem na subpasta do projeto com um nome de fácil entendimento sobre o que se tratava.

Os motivos apontados pelos investigadores para não depositarem as imagens foram a inexistência de necessidade, o facto de ser pouco usual e de não se falar muito do tema. Foram também mencionados condicionamentos no depósito devido à política de gestão

de dados.

Os investigadores CVS2 e CNA2 referiram que já tinham depositado dados em repositórios, mas nunca tinham depositado imagens. O primeiro não vê sentido em depositar as imagens despromovidas de contexto, depois de passar muitas horas a pensar onde as colocar nos artigos. O segundo apontou que era mais útil partilhar a informação que deu origem à imagem. Embora tenha referido que nunca tinha pensado em depositar determinados tipos de imagens que produz, por exemplo, fotografias de espécies raras.

As entrevistas revelaram que normalmente a partilha não é efetuada através do depósito das imagens em repositórios. O mesmo também foi verificado através da análise dos dados do questionário. Embora a partilha em repositórios não seja frequente, os investigadores mostraram abertura para o fazer no futuro.

6. Proposta de modelo de metadados para a descrição de imagens

O modelo de metadados proposto neste trabalho assume a forma de uma ontologia, focada em representar metadados para a descrição de dados de investigação em formato de imagem. Uma ontologia consiste num modelo para representar informação desenvolvida em linguagem compreendida por humanos e máquinas como, por exemplo, RDF, *Web Ontology Language* (OWL) e XML. Dependendo da sua extensibilidade, pode ser considerada “leve” ou “pesada”. A ontologia leve apenas contém as classes e atributos essenciais para a descrição, enquanto a ontologia pesada fornece uma descrição detalhada, possuindo um grande número de axiomas e restrições (Corcho, 2006). Optou-se por construir uma ontologia leve porque queria-se criar uma ontologia simples sem muitos axiomas e restrições que tornariam a sua utilização complexa.

O objetivo deste modelo de metadados é fornecer aos investigadores um conjunto de elementos que os ajudem a descrever as suas imagens. A sua aplicação transversal a todos os domínios de investigação contribui para a padronização da descrição e criação de pontos de contactos entre os vários domínios.

A sua simplicidade faz com que seja facilmente aplicada em sistemas de gestão de dados de investigação, como a plataforma Dendro⁴. Uma plataforma colaborativa de gestão de dados de investigação que permite organizar, descrever e promover o depósito dos dados em repositórios.

A ontologia é flexível pois permite que sejam adicionados elementos e restrições mais complexas caso surja essa necessidade em determinado contexto de investigação. Esta prática existe em repositórios de dados de investigação como os repositórios do INESC-TEC e da UP. Nestes repositórios existem descritores obrigatórios e descritores optativos que podem ser selecionados pelo investigador se este achar que são importantes. Os investigadores podem ainda adicionar descritores que não estejam na lista caso necessitem. No caso específico do repositório do INESC-TEC, alguns descritores são adicionados automaticamente como, por exemplo, o volume ocupado pelo conjunto de dados e a data da última modificação.

6.1. Processo de desenvolvimento da ontologia

O procedimento de desenvolvimento (Figura 23) foi o seguinte: levantamento de requisitos, comparação de modelos de metadados, desenvolvimento da ontologia e

⁴ <https://github.com/feup-infolab/dendro>

importação para o Dendro.

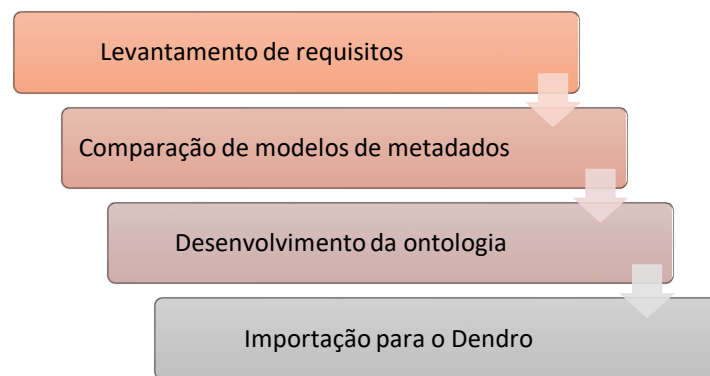


Figura 23 - Processo de desenvolvimento do modelo de metadados.

Na primeira fase realizou-se o levantamento dos descritores que deveriam compor o modelo. Para isso foram utilizados os resultados obtidos no questionário e nas entrevistas, cuja análise permitiu identificar quais os descritores relevantes e usados pelos investigadores. O levantamento foi completado com uma pesquisa em vários documentos publicados que ajudaram a estabelecer o modelo.

Na fase seguinte fez-se o mapeamento de alguns modelos de metadados para averiguar se os descritores selecionados já existiam em outros modelos. Os descritores que já existiam foram adaptados para o modelo, caso não existissem os descritores eram criados de raiz. Para além da análise dos modelos de metadados presentes na revisão da literatura, foram realizadas pesquisas no repositório *Linked Open Vocabularies*⁵ para verificar a pré-existência dos descritores propostos. Este repositório reúne vocabulários para a descrição geral ou específica de diversos domínios e setores de atividade.

De seguida procedeu-se ao desenvolvimento da ontologia no *Protegé*, uma aplicação gratuita e de código-aberto que permite a construção de modelos e aplicações baseadas em conhecimento com ontologias. Os descritores foram formalizados através de 17 *data properties*, aos quais foram adicionadas *annotation properties* para:

- Explicar o significado de cada descritor (*rdfs:comment*);
- Mencionar se é optativo (*rdfs:comment*);
- Representar o descritor em linguagem natural (*rdfs:label*);
- Mencionar o modelo de metadados de onde foi retirado (*rdfs:isDefinedBy*);
- Mencionar um possível valor com que o descritor pode ser preenchido (*hasAlternative*).

⁵ <https://lov.linkeddata.es/dataset/lov>

Os significados dos descritores foram retirados dos modelos de metadados correspondentes. Na Figura 24 podem observar-se as *data properties* da ontologia no *Protegé* e as *annotation properties* para o descritor “tipo”.

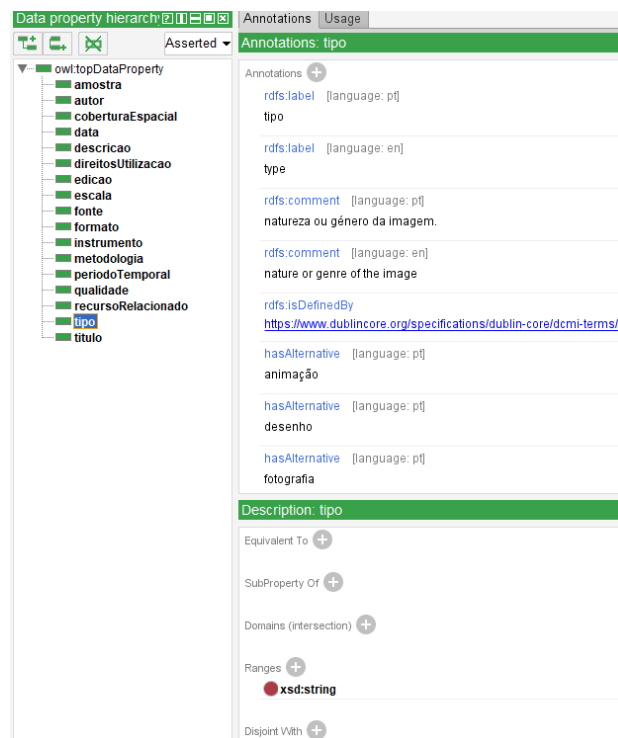


Figura 24 - *Data properties* e *annotation properties* do descritor "tipo".

Para a importação da ontologia para o Dendro foi necessário o desenvolvimento de um documento de texto para identificar o nome da ontologia, quais os seus descritores, qual a tipologia de dados (*string* para todos os casos) e para definir qual a forma da caixa de texto a apresentar para cada descritor. Este documento está disponível no Anexo III. Na Figura 25, pode-se observar a ontologia no Dendro depois de ser importada.

Figura 25 - Ontologia no Dendro.

6.2. Descritores do modelo

O modelo é composto por 17 descritores dos quais 13 são descritores obrigatórios e 4 descritores optativos (instrumento, cobertura espacial, amostra e escala), que os investigadores podem utilizar se acharem importante para descrever o seu conjunto de imagens. Destes descritores, 11 foram retirados do Dublin Core, 2 do DDI, 2 do DIF, 1 do *Quantities, Units, Dimensions and Values* (QUDV) e 1 descritor não tinha correspondência em nenhum modelo.

A Tabela 19 apresenta os descritores que compõem a ontologia, uma breve descrição de cada um destes descritores e o modelo de metadados de onde foram retirados. Estes descritores foram selecionados com base nas respostas dos investigadores no questionário e nas entrevistas, mas também na capacidade destes contribuírem para a contextualização, avaliação e reutilização dos conjuntos de imagens por outros investigadores.

Tabela 19 - Descritores da ontologia.

Descritor	Descrição	Modelo de Metadados
Amostra	Descreve a amostra utilizada na experiência	DDI
Autor	Entidade responsável pela autoria do recurso	Dublin Core
Cobertura espacial	Características espaciais relevantes para o conjunto de imagens. Pode ser um local ou as coordenadas geográficas	Dublin Core
Data	Um ponto ou período de tempo associado a um evento no ciclo de vida da imagem	Dublin Core
Descrição	Uma descrição sobre a imagem	Dublin Core
Direitos de utilização	Uma declaração sobre os direitos de propriedade associados à imagem	Dublin Core
Edição	Uma declaração sobre as edições/manipulações realizadas na imagem	-
Escala	Especifica os valores numéricos ou textuais definidos na escala de medida	QUDV
Fonte	Um recurso de onde a imagem deriva	Dublin Core
Formato	Formato do ficheiro ou meio físico em que se encontra a imagem	Dublin Core
Instrumento	O nome do instrumento usado para produzir a imagem	DIF
Metodologia	Metodologia e o processamento envolvidos na produção da imagem	DDI
Período temporal	Um período temporal relevante para o conjunto de imagens. Pode ser uma data ou período de tempo	Dublin Core
Qualidade	Informação sobre a qualidade dos dados ou quaisquer procedimentos de garantia de qualidade seguidos na produção da imagem	DIF
Recurso relacionado	Link ou apontamento para um recurso estritamente relacionado com imagem	Dublin Core
Tipo	Natureza ou género da imagem	Dublin Core
Título	Nome dado ao conjunto de imagens	Dublin Core

A utilização correta do modelo implica o conhecimento das seguintes considerações:

- A ontologia foca essencialmente a descrição de conjuntos de imagens, pelo que todas as imagens partilham os mesmos elementos de descrição. Contudo, o modelo também pode ser aplicado a uma só imagem;

- Quando um descritor possui dois valores, deve-se colocar separadamente o elemento de descrição, ou seja, repetir o elemento para cada valor diferente. Esta prática é também recomendada em modelos de metadados como o Dublin Core e o DDI;
- No descritor “fonte”, caso as imagens sejam todas produzidas pelo autor deve ser preenchido com “Autoria própria”. Se as imagens forem compiladas de um ou mais sítios deve-se colocar o respetivo local de onde foram retiradas. O Dublin Core recomenda que seja utilizado um sistema de identificação formal;
- O descritor “recursos relacionados” pode ser usado para mencionar um artigo científico realizado com base nas imagens, mas também para mencionar locais onde esteja documentação ou ficheiros relacionados com as imagens. Por exemplo, o local de depósito do código utilizado para analisar as imagens. O Dublin Core recomenda que esta ligação seja feita através de um URI, se não for possível a utilização de um sistema de identificação formal;
- Para “tipo” o Dublin Core recomenda a utilização de um vocabulário controlado. Assim, o descritor possui uma lista de termos pré-definidos para facilitar o seu preenchimento, mas outros valores podem ser adicionados se necessário. Os termos são fotografia, gráficos, ilustração, desenho, gravura, retrato, pintura, planta, vídeo e animação;
- O descritor “formato” deve ser utilizado para descrever a manifestação digital ou física da imagem. O Dublin Core recomenda também a utilização de um vocabulário controlado quando possível. De igual forma ao descritor “tipo”, este descritor possui uma lista de termos pré-definidos. Estes termos são .bmp, .dicom, .jpeg, .pdf, .png, .raw, .svg e .tiff.
- A “data” pode ser mencionada a diferente granularidade. O Dublin Core recomenda a expressão da data segundo a norma ISO 8601 ou um perfil da norma ISO (YYYY-MM-DDThh:mm:ssTZD). Recomenda também se a data inteira não for conhecida, usar-se o mês e o ano ou apenas o ano.

6.3. Exemplo prático de utilização do modelo

De seguida apresenta-se um exemplo de aplicação do modelo de descrição a um conjunto de imagens (Tabela 20). Esta descrição é realizada do ponto de vista do investigador, ou seja, assume-se o cenário em que o investigador está a descrever o conjunto de imagens que produziu.

O conjunto de imagens descrito está depositado no repositório *Cell Image Library*⁶ em domínio público. O conjunto é referente a fotografias de células humanas resultantes de uma experiência realizada segundo um protocolo de pintura celular (Bray et al., 2017).

Tabela 20 - Descrição de um conjunto de imagens através da ontologia.

Descritor	
Título	Human U2OS cells - compound cell-painting experiment
Data	2017
Período temporal	-
Autor	Sigrun Gustafsdottir
Autor	Vebjorn Ljosa
Autor	Katherine Sokolnicki
Autor	Joshua Bittker
Autor	Nicole Bodycombe
Autor	Mark Bray
Autor	Vlado Dancik
Autor	Tom Hasaka
Autor	Suk Hon
Autor	Melissa Kemp
Autor	Kejie Li
Autor	Deepika Walpita
Autor	Mathias Wawer
Autor	Todd Golub
Autor	Stuart Schreiber

⁶ <http://www.cellimagelibrary.org/home>

Autor	Paul Clemons
Autor	Anne Carpenter
Autor	Alykhan Shamji
Fonte	Autoria própria
Direitos de utilização	As imagens estão em domínio público, isentas de restrições de direitos autorais. Por cortesia, os utilizadores devem citar os autores do conjunto de imagens.
Descrição	O conjunto de imagens fornece uma base para testar métodos de criação de perfil baseados em imagem assim como a capacidade de distinguir os efeitos de pequenas moléculas. As imagens são de células U2OS tratadas cada uma com mais de 30.000 compostos bioativos conhecidos e rotuladas com seis rótulos que caracterizam sete organelos (ensaio de pintura celular).
Recurso relacionado	https://github.com/gigascience/paper-bray2017
Recurso relacionado	https://academic.oup.com/gigascience/article/6/12/giw014/2865213
Metodologia	Cinco canais fluorescentes, adquiridos em 6 campos de vista por poço a $\times 20$ de magnificação ($0.656 \mu\text{m}/\text{pixel}$). Cada canal foi armazenado em separado, num ficheiro de imagem TIFF 16-bit em escala cinza.
Edição	As imagens não foram editadas
Formato	.tiff (formato digital)
Tipo	Fotografia
Qualidade	Cada campo de vista foi avaliado para a presença de dois artefactos (desfoque focal e objetos saturados) e associado uma etiqueta “1” se presente e “0” se não.
Instrumento	Microscópio automático - ImageXpress Micro XLS
Escala	$\times 20$ magnificação ($0.656 \mu\text{m}/\text{pixel}$)
Amostra	375 lâminas em 384 formato poço, para um conjunto de dados de 988 994 campos de vista

6.4. Avaliação do modelo de metadados

A avaliação do modelo de metadados foi realizada com o objetivo de recolher a opinião de investigadores relativamente à utilidade e usabilidade do modelo. A informação recolhida é importante para, no futuro, serem realizadas atualizações e modificações que melhorem a sua utilização. Esta informação pode também ser importante para adaptar o modelo de metadados a um contexto específico, onde seja necessário a introdução de descritores específicos.

A avaliação (Anexo IV), devido à pandemia de covid-19, desenvolveu-se na plataforma *Google Forms*. Aos investigadores, foi pedido para descreverem uma imagem usada como dado de investigação, utilizando os descritores do modelo de metadados proposto. Antes de realizarem a tarefa de descrição, foi disponibilizada e aconselhada a leitura da documentação de apoio para utilizarem de forma correta o modelo. Depois de descreverem as imagens, foi pedido para responderem a cinco questões. Estas questões versavam sobre os descritores do modelo que consideravam mais e menos úteis, se acharam difícil a utilização do modelo, se este era útil e ainda se acrescentariam algum descritor.

Os investigadores escolhidos para realizarem esta tarefa, foram os investigadores que participaram nas entrevistas. Escolheram-se estes investigadores devido à sua familiarização com a temática desta dissertação e com o modelo de metadados. Porém, apenas dois investigadores aceitaram participar na tarefa, um das CEE e outro das CVS.

Os investigadores consideraram úteis diferentes descritores. O motivo para a escolha dos descritores úteis prendeu-se com o facto de serem genericamente aplicáveis. Outro motivo apresentado foi a captação de informação de contextualização e de como reproduzir a imagem. Os investigadores acharam alguns descritores menos úteis pois numa primeira análise não contribuíam para a compreensão da imagem e por serem os menos aplicáveis genericamente. Os investigadores ao considerarem úteis diferentes descritores, mostram que é importante haver alguma diversidade de descritores à sua disposição.

Na utilização do modelo foram encontradas algumas dificuldades. Um investigador achou o descritor “fonte” e “recurso relacionado” redundante. O outro investigador teve dificuldades na identificação do tipo de imagem e considerou que alguns descritores não eram aplicáveis às suas imagens.

O modelo foi considerado útil por ambos os investigadores, pois permite ter uma ideia e abrangente do contexto, propósito e significado das imagens. Os descritores do modelo

foram vistos como suficientes, os investigadores não acrescentariam mais nenhum descritor.

A avaliação geral do modelo desenvolvido foi positiva, embora seja necessário a realização de mais sessões de avaliação. Um maior número de avaliações permitirá obter informação mais aprofundada relativamente ao desempenho do modelo de metadados.

7. Proposta de boas práticas para a gestão de imagens em investigação

O resultado de imagens, bem organizadas, documentadas, preservadas, acessíveis, validadas e controladas em relação à sua precisão, são imagens de alta qualidade e valor para a ciência. Para além disso, uma boa gestão dos dados aumenta a eficiência da investigação, contribuindo para resultados baseados em evidência sólida e para a redução de custos assim como de tempo (Eynden et al., 2011).

O presente capítulo apresenta um conjunto de recomendações destinadas a ajudar os investigadores na gestão dos seus dados de investigação em formato de imagem. Estas recomendações visam tornar as imagens facilmente acessíveis e reutilizáveis, através de um conjunto de práticas que promova a boa gestão dos recursos imagéticos. As recomendações foram desenvolvidas tendo em conta algumas práticas recomendadas nas normas de gestão de dados textuais e numéricos. Algumas destas práticas podem ser vistas como de aplicação transversal na gestão de dados de investigação, pois existem aspetos que são comuns à gestão de qualquer tipo de dados. Na gestão de imagens é necessário ter especial atenção às questões relacionadas com o armazenamento e partilha. Os dados recolhidos através do questionário e das entrevistas auxiliaram também na elaboração destas orientações. Estes dados permitiram conhecer e avaliar o grau de familiarização dos investigadores relativamente às práticas de gestão de dados em formato de imagem. Com base nestes dois blocos de informação, procedeu-se à elaboração da proposta de boas práticas para a gestão de imagens na investigação.

Ao longo deste capítulo, as palavras dados e dados de investigação referem-se à utilização de dados imagéticos enquanto registos de apoio à investigação.

7.1. Planeamento

O planeamento da gestão de imagens deve começar antes do começo da recolha de imagens. Planear antecipadamente a organização, documentação e tratamento dos dados de investigação ao longo do projeto de investigação traz benefícios para o investigador. Não planear devidamente a forma como se gerem as imagens, pode levar a que no fim da investigação algumas imagens e anotações desapareçam ou que falem permissões para analisar ou publicar um conjunto de imagens.

O conteúdo do plano de gestão de dados depende dos dados, das necessidades do projeto de investigação e requisitos externos, como por exemplo, requisitos de financiadores da investigação. Geralmente, o plano deve abordar quais os dados a criar e como vai ser feita a organização, documentação, armazenamento, partilha e gestão dos dados depois do

projeto acabar. Ao longo de cada etapa devem ser documentados os recursos necessários, como, calendarização das tarefas e dispositivos necessários (Briney, 2015).

O planeamento não se resume à fase inicial da investigação, este deve consultado ao longo do projeto de investigação para garantir que se está a fazer o que foi delineado. Outro ponto importante a ter em conta é o facto do plano não ser estático, se for necessário o plano pode ser modificado no decorrer da investigação.

Assim, algumas considerações a ter no planeamento da gestão de dados em formato de imagem são:

- Fazer uma descrição geral das imagens a produzir ou recolher;
- Quais as imagens que são necessárias para a investigação. Importante mencionar se as imagens serão produzidas no contexto do projeto ou utilizadas imagens já existentes;
- Como será a metodologia de recolha das imagens;
- Explicar como a consistência e qualidade das imagens será garantida;
- Qual o volume e formato das imagens e as suas implicações no armazenamento, backups e reutilização;
- Considerar como as imagens serão armazenadas assim como procedimentos de backup e segurança;

A criação de um plano de gestão de imagens é útil pois fornece um guia para a gestão dos dados durante e depois da investigação terminar. É também importante para a partilha de dados, uma vez que funciona como um roteiro para preparar as imagens para a partilha. Em projetos de investigação colaborativos, o plano contribui para definir uma estratégia de gestão de imagens igual para todos (Briney, 2015).

Como foi verificado através do questionário e das entrevistas, os investigadores não costumam planear nem usar orientações que orientem a produção e uso das imagens. No entanto, a pré-existência de um guia de boas práticas é considerado útil pelos investigadores. Este ajuda-os a salvaguardarem as características e a preparar as imagens para publicação.

7.2. Documentação e Metadados

Um aspeto essencial para os dados de investigação serem partilháveis e utilizados por outros investigadores, é assegurar que estes podem ser compreendidos e interpretados corretamente. Isto pode ser atingido através de descrição clara e detalhada dos dados, anotações e informação de contextualização. A documentação que acompanha os dados deve ter informação sobre o contexto e método de recolha dos dados, estrutura do

conjunto de dados, procedimentos de validação e verificação da qualidade, alterações na versão do conjunto de dados e informação sobre o acesso e uso dos dados (Eynden et al., 2011).

Uma boa documentação das imagens utilizadas como dados de investigação deve incluir a seguinte informação:

- Contextualização da recolha das imagens: informação sobre o projeto de investigação, âmbito e objetivos;
- Método de produção das imagens: processo de produção, instrumentos usados bem como *hardware* e *software*, amostra;
- Método de análise;
- Processos de validação, tratamento e verificação da qualidade das imagens;
- Informação relativa às condições de acesso e uso.

Quando as imagens são compiladas de outros conjuntos de imagens é preciso ter especial atenção na documentação de aspetos relacionados com a proveniência das imagens. Alguns pontos a ter atenção são:

- Local de onde foi retirada a imagem;
- Autores da imagem;
- Qualidade das imagens;
- Consentimento para utilização.

Esta informação deve estar contida preferencialmente nos metadados relativos às imagens ou num documento que oriente a sua utilização. A informação pode estar também em cadernos de laboratórios usados pelos investigadores.

Na descrição das imagens através de metadados devem ser utilizados normas de metadados. Estas normas de metadados definem um conjunto de elementos descritivos que podem ser usados na descrição dos dados de investigação. Antes de usar um modelo de metadados é necessário pensar qual o modelo a usar. Atualmente existe uma grande variedade de modelos, uns de aplicação genérica e outros específica para determinado domínio científico.

Uma boa ferramenta para ajudar na escolha do modelo de metadados é o diretório de modelos de metadados fornecido pela *Research Data Alliance*⁷ (RDA). O diretório contém modelos genéricos e específicos, ferramentas e casos de uso para ajudar os investigadores a descobrir o modelo mais adequado aos seus dados. Alguns modelos de

⁷ <http://rd-alliance.github.io/metadata-directory/>

metadados mais utilizados em contexto de investigação são: Dublin Core, CERIF, DDI, DIF, EML, entre outros.

Relativamente a modelos de metadados que foquem diretamente a descrição de dados de investigação em formato de imagem, pouco modelos existem. O OME-TIFF e o *Flexible Image Transport System* (FITS) são formatos de ficheiros imagéticos que possuem metadados embebidos para a descrição da imagem. O OME-TIFF é específico para o domínio da biologia, já o FITS é direcionado para dados imagéticos astronómicos. Na descrição das imagens pode ser utilizado o modelo de metadados produzido no decurso da presente investigação, cuja aplicação é transversal a todas as áreas científicas.

A informação recolhida no estudo exploratório permitiu perceber que poucos investigadores estão sensibilizados relativamente à importância da descrição através de metadados. Também se verificou que a grande maioria deles não conhece nem utiliza modelos de metadados.

Na gestão das imagens podem ser utilizados sistemas de informação para facilitar esta tarefa. Tipicamente, estes sistemas acompanham todo o processo de gestão e permitem o trabalho colaborativo. Estas soluções evitam que os dados sejam documentados e fiquem apenas no computador de um investigador, estando em rede e disponível, sob permissões de acesso, a vários investigadores. Promovem o trabalho em conjunto, um acompanhamento mais próximo dos dados e uma metodologia mais concreta para a documentação e metadados. Um exemplo é o Dendro.

7.3. Armazenamento e Backup

O acesso a longo prazo e a prevenção de perdas de informação requiere a existência de uma boa estratégia de armazenamento. Igualmente importante, é garantir a acessibilidade das imagens a longo prazo, para isso é preciso ter atenção aos formatos utilizados. As imagens devem estar também bem organizadas e estruturadas, assim como com nomes e versões claras (Eynden et al., 2011).

O armazenamento é um fator central na gestão dos dados de investigação. Uma boa prática é manter os dados em três localizações distintas, duas no local de trabalho e uma fora. Atualmente existem muitas soluções de armazenamento que permitem atingir as três cópias, as melhores opções são o computador, disco externo, servidor local e armazenamento em fita magnética (Briney, 2015). O computador e o disco externo são os locais de armazenamento mais utilizados pelos investigadores.

A existência de ficheiros e estruturas de pastas bem organizadas tornam mais fácil a tarefa de localizar e aceder aos dados. A organização dos ficheiros deve ser desenvolvida

de forma a que funcione no projeto de investigação e ser utilizada de maneira consistente. Os ficheiros podem estar organizados por tipo de dados, atividade da investigação, materiais da investigação, documentação e publicações. Em projetos colaborativos, a necessidade de uma organização estruturada é ainda maior.

Os nomes atribuídos aos ficheiros devem providenciar indicações sobre o seu conteúdo, identificar claramente e ajudar a classificar os ficheiros. O nome do ficheiro pode conter o acrónimo do projeto, iniciais do investigador, informação sobre o tipo de ficheiro, versão e data. São consideradas boas práticas:

- Nomes breves, mas informativos;
- Usar o nome dos ficheiros para criar uma classificação geral dos ficheiros;
- Evitar usar espaços e caracteres especiais;
- Evitar nomes muito longos.

A versão dos ficheiros pode ser indicada através da introdução da data no nome do ficheiro, numeração da versão (exemplo, v1, v2, v3) e descrição da versão (rascunho ou final). Pode também ser criado um documento que contenha anotações sobre versões, datas, autores e alterações efetuadas (Eynden et al., 2011).

A gestão das imagens revela-se um processo difícil quando existe um grande número de imagens, devido ao espaço limitado para armazenamento. A escolha de formatos de armazenamento com compressão com perda, resulta em perdas na claridade, resolução, camadas e fidelidade da imagem. Por este motivo, preferencialmente a imagem deve ser guardada num formato sem perda (.tiff, .raw) ou balanceada a compressão sem perdas contra possíveis problemas de armazenamento ou de compatibilidade que possam surgir no futuro. É recomendado que os investigadores minimizem o número de compressões realizadas, para não comprometer a qualidade das imagens (Vejvoda et al., 2016). Em relação ao volume ocupado pelas imagens na investigação, a maioria dos investigadores não o consegue quantificar. Isto afeta negativamente a capacidade de planear a estratégia de armazenamento.

A realização regular de *backups* é essencial para salvaguardar a perda de informação contra eventos maliciosos ou acidentais. O procedimento de *backup* depende do contexto do projeto, do valor dos dados e do nível de risco. Algumas boas práticas a considerar são:

- Quais os ficheiros a salvaguardar, ficheiros específicos ou o todo o sistema computacional;

- A frequência de realização dos backups, depois de cada modificação no ficheiro ou em intervalos regulares;
- O modo de realização dos backups, preferencialmente deve ser automático;
- A estratégia para todos os sistemas onde estão armazenados os dados;
- A organização e etiquetagem clara dos ficheiros resultantes do backup.

O *backup* pode ser feito em locais como o disco externo pessoal ou do projeto de investigação, *cloud*, rede interna ou numa plataforma colaborativa. Os dados recolhidos demonstraram que os investigadores não costumam realizar *backups* regularmente, o que constitui um risco para a investigação.

7.4. Partilha

A partilha é o culminar dos processos de descrição, documentação, armazenamento e organização. Estes processos são críticos para a partilha de dados, uma vez que asseguram que estes estão devidamente contextualizados, acessíveis e reutilizáveis a longo prazo.

A partilha pública dos dados traz benefícios para a ciência, mas também para os investigadores. A partilha contribui para reprodutibilidade dos dados, aumento das citações e colaborações e ainda redução de gastos e tempo na produção de dados.

Nesta fase deve ser considerado onde, como, quando e quais as imagens a partilhar. Devem também ser considerados aspetos legais e éticos relacionados com a partilha como, anonimização, consentimento para partilha, restrições no uso das imagens e período de embargo.

As imagens que devem ser partilhadas são aquelas utilizadas para suportar artigos científicos ou aquelas que permitem reproduzir os resultados da investigação. Em relação ao formato em que as imagens devem ser partilhadas, é uma boa prática partilhar no formato que seja mais útil para os outros reutilizarem. Devido à perda de informação resultante de compressões e conversões de formato, o ideal será a partilha das imagens num formato sem perda de informação (por exemplo .tiff). Se não for possível, deve ser minimizado o número de compressões e conversões.

Preferencialmente, as imagens devem ser colocados em repositórios. O primeiro benefício do depósito num repositório é que depois dos dados serem depositados, o investigador não tem de se preocupar com a manutenção do acesso. O repositório será o responsável por assegurar a acessibilidade aos dados. Os dados aos estarem num repositório têm uma maior probabilidade de serem descobertos e citados. Além disso, os repositórios fornecem ferramentas que permitem ligar o conjunto de dados depositado

ao respetivo artigo, recolher métricas e atribuir *Digital Object Identifiers* (DOIs). A atribuição de DOIs é importante na partilha, pois permite identificar de forma exclusiva e permanente um objeto digital (Briney, 2015). Apesar dos benefícios que a partilha de dados em repositórios possui, a maioria dos investigadores não o faz. Isto foi verificado através do questionário e das entrevistas.

A R3Data⁸ constitui um diretório que ajuda os investigadores a identificar possíveis repositórios para efetuar o depósito dos dados. É possível pesquisar por tema, tipo de conteúdo (por exemplo, imagens) e por país. Esta ferramenta identifica mais de 2500 repositórios de dados científicos, sendo por isso importante definir quais os critérios para a escolha do repositório. Alguns critérios para selecionar o repositório podem ser:

- Quem é a instituição responsável pelo repositório;
- Confiabilidade e reputação do repositório;
- O repositório fomenta a longevidade dos dados, por exemplo através de DOIs e tarefas de curadoria;
- O repositório é de fácil utilização e possui ferramentas de citação e métricas (Briney, 2015);
- Verificar se o repositório tem capacidade para acolher o conjunto de imagens.

Alguns dos repositórios genéricos mais conhecidos são o Figshare, Dataverse e Zenodo. Estes repositórios possuem dados de investigação de todas as áreas científicas. Existem também repositórios temáticos dedicados a domínios científicos e temas específicos, tipo de conteúdo assim como repositórios institucionais. Dois exemplos de repositórios institucionais são os repositórios de dados de investigação da UP e do INESC-TEC. Os repositórios, por norma, definem um volume máximo para o depósito.

Outras formas de partilhar os dados é através de pedidos efetuados por colegas e depósito dos dados em páginas web pessoais. Estas opções não são as ideias pois requerem mais trabalho por parte do investigador. Além disso, não oferecem a estabilidade e funcionalidades proporcionadas por um repositório.

Um aspeto importante na partilha é a atribuição de uma licença de utilização. Esta licença dita a forma como os dados são utilizados. A *Creative Commons* e o *Open Data Commons* definem um conjunto de licenças que os investigadores podem usar nos seus dados (Tabela 21).

⁸ <https://doi.org/10.17616/R3D>

Tabela 21 - Tipos de licenças de utilização. Adaptado de Briney (2015).

Tipo de licença	Licença <i>Creative Commons</i>	Licença <i>Open Data Commons</i>	Descrição
Domínio Público	CCo	PDDL	Conteúdo está em domínio público sem restrições de uso
Atribuição	CC BY	ODC-By	Todas as formas de reutilização permitidas desde seja atribuído ao criador
Atribuição, partilha igual	CC BY-SA	ODC-ODbL	Reutilização permitida com atribuição e uso de uma licença equivalente no conteúdo produzido com base no original
Atribuição, não comercial	CC BY-NC	-	Reutilização autorizada para fins não comerciais, com atribuição
Atribuição, sem derivação	CC BY-ND	-	Conteúdo original não pode ser modificado e tem de ser atribuído
Atribuição, não comercial, partilha igual	CC BY-NC-SA	-	Tem de ser feita a atribuição, reutilização apenas para fins não comerciais e licença equivalente nos produtos derivados
Atribuição, não comercial, sem derivação	CC BY-NC-ND	-	Não podem ser feitas modificações, nem ser usado comercialmente e tem de ser atribuído

Em relação à escolha da licença de utilização, as licenças recomendadas são a de domínio público e a de atribuição. Quando os dados estão em domínio público não é obrigatório citar, embora por cortesia deva ser feito. Porém, cabe ao investigador decidir a forma como quer que as suas imagens sejam partilhadas.

As imagens constituem um grande desafio na questão de dados sensíveis, pois sejam elas estáticas ou em movimento podem comprometer a identidade de pessoas. Por isso, no momento de partilha deve-se ter em atenção a proteção da identidade dos participantes. Quando necessário devem-se adotar processos que permitam anonimizar as imagens. A anonimização pode ser feita através do desfoque ou pintura das áreas que permitam a

identificação. Este processo é mais complexo nas imagens do que em dados textuais, no último tipo de dados é comum substituir os nomes por caracteres ou números. Assim é essencial assegurar que este processo não afeta a qualidade da imagem.

8. Conclusões e perspectivas de trabalho futuro

A gestão dos dados produzidos durante a investigação é importante para o desenvolvimento da ciência. Uma boa estratégia de gestão de dados para além de melhorar a eficiência do processo de investigação, aumenta o tempo de vida dos dados de investigação.

O uso da imagem como dado de investigação tornou-se mais recorrente, devido ao desenvolvimento tecnológico, nomeadamente dos instrumentos de captura, e a sua capacidade em registar fielmente o que outros meios não conseguem. Porém, na área da gestão de dados de investigação pouca ou nenhuma atenção tem sido dada à gestão deste tipo de dados.

Com este estudo, verificou-se que os investigadores utilizam as imagens como dado de investigação principalmente no domínio das Ciências da Vida e da Saúde. Porém, não é costume existirem orientações para a sua gestão.

No processo de criação, existe um grande uso de instrumentos digitais. O microscópio sobressai nas Ciências da Vida e da Saúde e a câmara fotográfica tradicional nas Ciências Sociais e Humanidades. Os processos de verificação da qualidade são diversos e nem todos os investigadores o fazem. De igual modo, o processamento e análise é heterogéneo, não se tendo verificado a existência de padrões. Estas foram as fases onde se verificaram mais diferenças. Isto pode ser explicado através da existência de diversos objetos de estudo e metodologias nos diferentes domínios científicos.

Na descrição, os investigadores não utilizam modelos de metadados para os assistir nesta tarefa. Em relação aos descritores foram obtidas respostas variadas sobre os descritores usados e tidos como relevantes, com apenas três descritores a serem consensuais entre os investigadores, título, autor e data.

O armazenamento das imagens é, maioritariamente, realizado no computador. Isto pode ser explicado devido ao fácil acesso, familiaridade e ao seu uso recorrente nos projetos de investigação. De notar que grande parte dos investigadores não está consciente do volume de imagens que produz e utiliza na investigação.

Apesar dos investigadores partilharem as imagens, verificou-se que não o fazem através do depósito em repositórios que asseguram a sua preservação e acessibilidade a longo prazo. Uma possível explicação é o facto dos investigadores considerarem o processo de partilha difícil (Tenopir et al., 2011).

A realização da dissertação na temática da gestão de dados de investigação em formato de imagem é relevante, por existirem poucos trabalhos a abordar este tema. Este trabalho

por um lado, apresenta informação valiosa sobre os processos e hábitos relativos à gestão de dados imagéticos em contexto de investigação. Por outro lado, disponibiliza orientações para os investigadores realizarem essa gestão.

Os objetivos propostos na introdução para esta dissertação foram atingidos. Primeiro, foram sistematizados e analisados os procedimentos relativos à gestão de dados de investigação em formato de imagem. Para isso utilizou-se o questionário e as entrevistas para estudar o uso e produção de imagens em contexto de investigação. Depois da recolha e análise dos dados, procedeu-se ao desenvolvimento do modelo de metadados e das orientações para a gestão de dados imagéticos. Estes dois elementos constituem ferramentas para auxiliar os investigadores a gerirem os seus dados imagéticos de acordo com boas práticas. O trabalho desenvolvido contribui assim para fomentar a utilização de boas práticas na gestão de dados em formato de imagem no contexto de investigação.

Os dados recolhidos através do questionário assim como os dados resultantes da análise encontram-se depositados no repositório de dados de investigação do INESC-TEC⁹.

Este trabalho está associado a algumas limitações. A primeira limitação está relacionada com a divulgação do questionário. No processo de divulgação do questionário existiram dificuldades na obtenção de resposta ao pedido de divulgação do questionário por parte dos centros de investigação e da reitoria da UP. Isto inviabilizou a obtenção de um maior número de respostas. Contudo, as 107 respostas obtidas, constituem um número aceitável para o estudo. A segunda limitação prende-se com a dificuldade em identificar investigadores de diferentes domínios científicos para as entrevistas. Tal como no questionário, esta limitação afetou a abrangência do estudo.

A atual situação de pandemia de covid-19 e as respetivas medidas de segurança dificultaram a realização de algumas atividades. As entrevistas e a avaliação do modelo de metadados tiveram de ser realizadas através de uma plataforma online. Algumas tarefas que envolviam a colaboração de membros do laboratório tornaram-se mais morosas.

No futuro será importante aprofundar o estudo do uso da imagem como dado de investigação, através da recolha de um maior número de respostas. Em paralelo, é também relevante captar a atenção dos investigadores para a temática da gestão de dados de investigação e, em especial, para a inclusão das imagens quando utilizadas como dado.

Será igualmente importante avaliar a aplicabilidade das práticas recomendadas para a gestão de dados em formato de imagem. A avaliação permitirá conhecer a disposição dos investigadores em seguir as práticas recomendadas. Permite também aprimorar o

⁹ <https://doi.org/10.25747/7ma9-9132>

conjunto de práticas recomendadas conforme a opinião dada pelos investigadores. Outro possível caminho a seguir, será aplicar as práticas recomendadas e guiar os investigadores até ao depósito das imagens num repositório.

De igual modo será importante aprofundar a opinião investigadores sobre o modelo de metadados desenvolvido, nomeadamente através de um maior número de avaliações. A opinião dada pelos investigadores constituirá uma oportunidade para refinar e, se existir necessidade, adaptar a ontologia a contextos de investigação específicos.

Referências bibliográficas

- Abell, C. (2010). The epistemic value of photographs. In C. Abell & K. Bantinaki (Eds.), *Philosophical Perspectives on Depiction* (pp. 82–103).
<https://doi.org/10.4324/9780203790762>
- Access to Biological Collection Data task group. (2007). Access to Biological Collection Data (ABCD), Version 2.06. Retrieved December 12, 2019, from
<http://www.tdwg.org/standards/115>
- Ball, A. (2012). *Review of Data Management Lifecycle Models*. Retrieved from University of Bath website:
<https://purehost.bath.ac.uk/ws/files/206543/redm1rep120110ab10.pdf>
- Banks, M. (2007). *Using visual data in qualitative research* (1st ed.; U. Flick, Ed.). London: SAGE Publications.
- Bell, G. (2009). Foreword. In *The Fourth Paradigm: Data-Intensive Scientific Discovery* (pp. xi–xvii). Retrieved from
<https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/M091000H.pdf>
- Besser, H. A. (1990). Visual Access to Visual Images : The UC Berkeley Image Database Project Visual Access to Visual Images. *Library Trends*, 38(4).
- Borgman, C. L. (2012, June). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, Vol. 63, pp. 1059–1078. <https://doi.org/10.1002/asi.22634>
- Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, 7(1–2), 17–30.
<https://doi.org/10.1007/s00799-007-0022-9>
- Bray, M.-A., Gustafsdottir, S. M., Rohban, M. H., Singh, S., Ljosa, V., Sokolnicki, K. L., ... Carpenter, A. E. (2017). A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *GigaScience*, 6(12).
<https://doi.org/10.1093/gigascience/giw014>
- Briney, K. (2015). *Data Management for Researchers: Organize, Maintain and Share your Data for Research Success*. Exeter, UK: Pelagic Publishing.
- CASRAI. (2015). Research data. Retrieved November 11, 2019, from
https://dictionary.casrai.org/Research_data

- Castro, J. A., Silva, J. R. da, & Ribeiro, C. (2014). Creating lightweight ontologies for dataset description practical applications in a cross-domain research data management workflow. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 313–316. <https://doi.org/10.1109/JCDL.2014.6970185>
- Chen, H.-L., & Rasmussen, E. (1999). Intellectual Access to Images. *Library Trends*, 48(2), 291–302.
- Choo, C. W. (2002). *Information management for the intelligent organization : the art of scanning the environment* (3rd ed). Medford, NJ: Information Today.
- Corcho, O. (2006). Ontology Based Document Annotation: Trends and Open Research Problems. *International Journal of Metadata, Semantics and Ontologies*, 1.
- Cox, A. M., & Tam, W. W. T. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, 70(2), 142–157. <https://doi.org/10.1108/AJIM-11-2017-0251>
- Cromey, D. W. (2010). Use and Manipulation of Scientific Digital Images. *Sci Eng Ethics*, 16(4), 639–667. <https://doi.org/10.1007/s11948-010-9201-y>
- Cromey, D. W. (2013). Digital Images Are Data: And Should Be Treated as Such. *Methods in Molecular Biology*, 931(1), 153–167. https://doi.org/10.1007/978-1-62703-056-4_1
- DataONE. (2019). Data Life Cycle. Retrieved December 19, 2019, from <https://www.dataone.org/data-life-cycle>
- Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image Retrieval : Ideas , Influences , and Trends of the New Age. *ACM Computing Surveys*, 40(2), 1–60. <https://doi.org/10.1145/1348246.1348248>
- Datta, R., Li, J., & Wang, J. Z. (2005). Content-Based Image Retrieval - Approaches and Trends of the New Age. In H. Zhang, J. Smith, & Q. Tian (Eds.), *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval* (pp. 253–262). Retrieved from <https://dl.acm.org/doi/10.1145/1101826.1101866>
- Day, M. (2005). Metadata. In S. Ross & M. Day (Eds.), *DCC Digital Curation Manual*.
- DDI Alliance. (2019). XML Schema for DDI-Lifecycle - Version 3.2. Retrieved December 4, 2019, from <https://ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/>

- Digital Curation Centre. (2019a). DCC Curation Lifecycle Model. Retrieved December 22, 2019, from <http://www.dcc.ac.uk/resources/curation-lifecycle-model>
- Digital Curation Centre. (2019b). Glossary. Retrieved November 11, 2019, from <http://www.dcc.ac.uk/digital-curation/glossary#D>
- Directory Interchange Format. (2018). Writer's Guide. Retrieved from Global Change Master Directory - National Aeronautics and Space Administration website: <https://gcmd.nasa.gov/DocumentBuilder/defaultDif10/guide/whatisadif.html>
- Dondero, M. G. (2009). The Semiotics of Scientific Image: from Production to Manipulation. *The American Journal of Semiotics*, 4, 1–19.
- Dublin Core Metadata Initiative. (2019). Metadata Basics. Retrieved December 3, 2019, from <https://www.dublincore.org/resources/metadata-basics/>
- Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata Principles and Practicalities. *D-Lib Magazine*, 8(4). Retrieved from <http://www.dlib.org/dlib/aprilo2/weibel/o4weibel.html>
- Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5), 667–690. <https://doi.org/10.1177/0306312711413314>
- Enser, P. (2000). Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. *Journal of Information Science*, 26(4), 199–210.
- Enser, P., Sandom, C. J., Hare, J. S., & Lewis, P. H. (2007). Facing the reality of semantic image retrieval. *Journal of Documentation*, 63(4), 465–481. <https://doi.org/10.1108/00220410710758977>
- euroCRIS. (2019). Main features of CERIF. Retrieved December 4, 2019, from <https://www.eurocris.org/cerif/main-features-cerif>
- Eynden, V. Van den, Corti, L., Woollard, M., Bishop, L., & Horton, L. (2011). *Managing and sharing data: best practices for researchers*. Essex.
- Fegraus, H., Andelman, S., Jones, M. B., & Schildhauer, M. (2005). Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *The Bulletin of the Ecological Society of America*, 86(3), 158–168. [https://doi.org/doi:10.1890/0012-9623\(2005\)86\[158:MTVOED\]2.0.CO;2](https://doi.org/doi:10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2)
- FORCE11. (2019). The FAIR Data Principles . Retrieved November 16, 2019, from

- <https://www.force11.org/group/fairgroup/fairprinciples>
- Furtado, F., Príncipe, P., & Carvalho, J. (2017). *Kit sobre dados de investigação*. Retrieved from <http://repositorium.sdum.uminho.pt/handle/1822/46351>
- Genova, F., Hanahoe, H., Laaksonen, L., Morais-Pires, C., Wittenburg, P., & Wood, J. (2014). *The Data Harvest : How sharing research data can yield knowledge, jobs and growth*. (December), 40. Retrieved from <https://europe.rd-alliance.org/documents/publications-reports/data-harvest-how-sharing-research-data-can-yield-knowledge-jobs-and>
- Goodrum, A. A. (2000). Image Information Retrieval: An Overview of Current Research. *Informing Science*, 3(2), 5.
- Gordon, A. S. (2001). Browsing image collections with representations of common-sense activities. *Journal of the American Society for Information Science and Technology*, 52(11), 925–929. <https://doi.org/10.1002/asi.1143>
- Heery, R., & Patel, M. (2000). Application Profiles: Mixing and Matching Metadata Schemas. *Ariadne*, (25). Retrieved from <http://www.ariadne.ac.uk/issue/25/app-profiles/>
- Heidorn, P. B. (1999). Image retrieval as linguistic and nonlinguistic visual model matching. *Library Trends*, 48(2), 303–325.
- Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation*, 3(1), 134–140.
- Higman, R., Bangert, D., & Jones, S. (2019). Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights the UKSG Journal*, 32. <https://doi.org/10.1629/uksg.468>
- Japan Electronics and Information Technology Industries Association. (2002). *Exchangeable image file format for digital still cameras: Exif Version 2.2* (p. 154). p. 154. Retrieved from <https://www.exif.org/Exif2-2.PDF>
- Jones, M. B., O'Brian, M., Mecum, B., Boettiger, C., Schildhauer, M., Maier, M., ... Chong, S. (2019). *Ecological Metadata Language version 2.2.0*. <https://doi.org/10.5063/f11834t2>
- Jörg, B. (2010). CERIF: The Common European Research Information Format Model. *Data Science Journal*, 9(July), 24–31. <https://doi.org/10.2481/dsj.cris4>
- Layne, S. S. (1994). Some Issues in the Indexing of Images. *Journal of the American*

- Society for Information Science*, 45(8), 583–588.
- Lyon, L. (2016). Transparency: The emerging third dimension of open science and open data. *LIBER Quarterly*, 25(4), 153–171. <https://doi.org/10.18352/lq.10113>
- Matusiak, K. K. (2006). Towards user-centered indexing in digital image collections. *OCLC Systems & Services: International Digital Library Perspectives*, 22(4), 283–298. <https://doi.org/10.1108/10650750610706998>
- Ménard, E. (2010). Image indexing and retrieval: Challenges and new perspectives. *Canadian Journal of Information and Library Science*, 34(3), 245–248. <https://doi.org/10.1353/ils.2010.0002>
- Meskin, A., & Cohen, J. (2008). Photographs as Evidence. In S. Walden (Ed.), *Photography and Philosophy, Essays on the Pencil of Nature* (p. 338). Blackwell Publishing.
- Michener, W. K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, 1(1), 3–7. <https://doi.org/10.1016/j.ecoinf.2005.08.004>
- Möbner, N. (2013). Photographic Evidence and the Problem of Theory-Ladenness. *Journal for General Philosophy of Science*, 44(1), 111–125. <https://doi.org/10.1007/s10838-013-9219-3>
- NASA. (2019). Directory Interchange Format (DIF) Standard. Retrieved December 3, 2019, from <https://earthdata.nasa.gov/esdis/eso/standards-and-references/directory-interchange-format-dif-standard>
- National Science Board. (2005). *Long-Lived Digital Data Collections : Enabling Research and Education in the 21 st Century*. Retrieved from <https://www.nsf.gov/geo/geo-data-policies/nsb-0540-1.pdf>
- OECD. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding* (p. 24). p. 24. OECD Publishing.
- OECD. (2015). *Making Open Science a Reality*. (pp. 1–108). pp. 1–108. <https://doi.org/10.1787/5jrs2f963zs1-en>
- Pennock, M. (2007). Digital Curation : A Life-Cycle Approach to Managing and Preserving Usable Digital Information. *Library & Archives*, (1), 1–3.
- Pinto, M. A. (2016). *Gestão de Serviços de Informação (PowerPoint de apoio à disciplina de Gestão de Serviços de Informação, lecionada na FLUP, LCI)* (p. 128). p. 128.

- Pinto, M. A., & Silva, A. M. da. (2005). Um Modelo Sistêmico E Integral De Gestão Da Informação Nas Organizações. *2º Congresso Internacional de Gestão Da Tecnologia e Sistemas de Informação*, 24. São Paulo.
- Pryor, G. (2018). Why manage research data? *Managing Research Data*, 1–16.
<https://doi.org/10.29085/9781856048910.002>
- Qin, J., & Li, K. (2013). 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*, 25–34.
- Quivy, R., & Campenhoudt, L. Van. (1995). *Manual de investigação em Ciências Sociais* (4ª). Lisboa: Gradiva.
- Rafferty, P. (2019). Disrupting the Metanarrative : A Little History of Image Indexing and Retrieval. *Knowledge Organization*, 46(1), 4–15.
<https://doi.org/10.5771/0943-7444-2019-1-4.Abstract>
- Ribeiro, C., Rodrigues, E., Fernandes, E. M., & Saraiva, R. (2010). *Os repositórios de dados científicos: estado da arte*. Retrieved from <https://repositorio-aberto.up.pt/handle/10216/23806>
- Roger, K. (2017). The Fringe Value of Visual Data in Research: How Behind Is Academia? *International Journal of Qualitative Methods*, 16(1).
<https://doi.org/10.1177/1609406917736668>
- Royal Society. (2012). *Science as an open enterprise*. Retrieved from <https://royalsociety.org/topics-policy/projects/science-public-enterprise/Report/>
- Sandweiss, M. A. (2007). Image and Artifact: The Photograph as Evidence in the Digital Age. *Journal of American History*, 94(1), 193–202.
<https://doi.org/10.2307/25094789>
- Silva, J. R. (2014). Fotografia e ciência: A utopia da imagem objetiva e seus usos nas ciências e na medicina. *Boletim Do Museu Paraense Emilio Goeldi: Ciências Humanas*, 9(2), 343–360. <https://doi.org/10.1590/1981-81222014000200006>
- Structural Reform Group. (2004). *DDI VERSION 3.0 CONCEPTUAL MODEL*. Retrieved from <https://ddialliance.org/sites/default/files/Concept-Model-WD.pdf>
- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6), 1–21. <https://doi.org/10.1371/journal.pone.0021101>

- Tešić, J. (2005). Metadata Practices for Consumer Photos. *IEEE Multimedia*, 12(3), 86–93. <https://doi.org/10.1109/MMUL.2005.50>
- Treloar, A., & Wilkinson, R. (2008). Rethinking metadata creation and management in a data-driven research world. *Proceedings - 4th IEEE International Conference on EScience, EScience 2008*, 782–789. <https://doi.org/10.1109/eScience.2008.41>
- UK Data Archive. (2019). Research data lifecycle. Retrieved December 19, 2019, from <https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx>
- Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a Standard for the Social Sciences. *International Journal of Digital Curation*, 3(1), 107–113. <https://doi.org/10.2218/ijdc.v3i1.45>
- Vejvoda, B., Burpee, K. J., & Lackie, P. (2016). Image Management as a Data Service. *IASSIST Quarterly*, 40(27–34).
- Whyte, A., & Tedds, J. (2011). *Making the Case for Research Data Management*. DCC Briefing Papers. Retrieved from <http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm>
- Wilder, K. E. (2009). Photography and the art of science. *Visual Studies*, 24(2), 163–168. <https://doi.org/10.1080/14725860903106161>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 1–9. <https://doi.org/10.1038/sdata.2016.18>
- Willis, C., Greenberg, J., & White, H. (2012). Analysis and Synthesis of Metadata Goals for Scientific Data. *Journal of the American Society for Information Science and Technology*, 63(8), 1505–1520. <https://doi.org/10.1002/asi.22683>
- Zhou, W., Li, H., & Tian, Q. (2017). *Recent Advance in Content-based Image Retrieval: A Literature Survey*. 1–22. Retrieved from <http://arxiv.org/abs/1706.06064>

ANEXOS

ANEXO I – Questionário

Planeamento

Como são obtidas as imagens?

- Produzidas no contexto da investigação
- Adquiridas em vendedores
- Fornecidas por terceiros

Costuma realizar um documento que oriente a utilização das imagens, como por exemplo, um plano de gestão de dados?

- Nunca
- Raramente
- Ocasionalmente
- Frequentemente
- Sempre

Criação, recolha e compilação

Que instrumentos são utilizados para produzir as imagens?

- Máquina fotográfica
- Microscópio
- Computador
- Aparelhos de imagiologia médica

Realiza algum tipo de manipulação/edição na imagem?

- Não realizo manipulação/edição da imagem
- Recorto a imagem
- Faço ajustes simples (contraste, brilho, saturação, etc) em toda a imagem
- Uso filtros para melhorar a qualidade de imagem
- Clono ou copio objetos de outras imagens e componho uma imagem nova

Se manipula/edita, onde a realiza?

- Imagem original
- Cópia da imagem original

Assegurar a qualidade

Costuma existir a validação? por quem? como é o processo?

Quais os processos existentes para assegurar a qualidade da imagem?

Que fatores influenciam a qualidade da imagem?

Processamento/Análise

Que programas utiliza para analisar as imagens?

Que tipo de análise faz à imagem?

Descrever

As imagens são acompanhadas por anotações?

Nunca

Raramente

Ocasionalmente

Frequentemente

Sempre

Onde são realizadas as anotações?

Na própria imagem

Caderno

Ficheiro digital

Que normas de metadados utiliza para descrever as imagens?

Não utilizo metadados

Dublin Core

CERIF

DDI

DIF

ABCD

EML

EXIF

Que informação considera ser importante para descrever a imagem?

Título

Autor

Data

Descrição

Instrumento de captura

Fonte

Direitos

Método

Amostra

Utiliza algum software que ajude a descrever as imagens?

Nunca

Raramente
Ocasionalmente
Frequentemente
Sempre

Documenta as manipulações/edições a que a imagem é submetida?

Nunca
Raramente
Ocasionalmente
Frequentemente
Sempre

Armazenar

Como são armazenadas as imagens?

Computador
Disco externo
Pen drive
Cloud

Realiza backups das imagens?

Diariamente
Semanalmente
Mensalmente
Não realizo backups

Qual o formato em que são armazenadas as imagens?

TIFF
JPEG
PNG
BMP
RAW

Qual o volume estimado de imagens armazenadas?

Descobrir/Partilhar

Quando são as imagens partilhadas?

Decorrer da investigação
Final da investigação
Só algum tempo depois da investigação ter terminado

Quem define e autoriza a partilha?

Responsável pela investigação
Investigador

Instituição

Entidade financiadora

A que restrições de acesso se encontram sujeitas as imagens?

Livre acesso e utilização

Livre acesso mas restrições à utilização

Restrições ao acesso e utilização

Onde são partilhadas?

Repositórios institucionais

Repositórios temáticos

Arquivos de imagens

Nos artigos científicos existe alguma ligação ao local onde os dados estão armazenados?

Com quem são partilhadas?

Entre colegas do grupo de investigação

Comunidade restrita de investigadores

Instituição

Entidade financiadora

Público em geral

ANEXO II - Guião da entrevista

Investigador -

Área científica a que pertence -

Grupo de investigação -

Contacto -

Data -

Para responder às seguintes questões baseie-se nos últimos 3 projetos de investigação em que trabalhou.

Planeamento

1. De forma simples quais são as fases pela qual a imagem passa durante a investigação?
2. Costuma elaborar um documento que oriente a produção e utilização de imagens na investigação?
 - a. Se sim, quais os benefícios que traz?
 - b. Se não, qual o motivo para não elaborar?
3. Gostaria de ter recomendações para a gestão de imagens em investigação? Em que fase/fases estas recomendações teriam mais valor?

Criação, recolha e compilação

1. Como se desenvolve o processo de produção de imagens durante a investigação?

Assegurar a qualidade

1. Se existirem processos para assegurar a qualidade, que atividades costumam ser realizadas?

Processamento/Análise

1. Como é o processo de análise (extração de informação) das imagens?

Descrever

1. Em relação à descrição das imagens, como descreve as imagens utilizadas na investigação?
2. Quando utiliza uma imagem não produzida por si, que informação necessita para a compreender e utilizar?
3. Na sua opinião, que benefícios traria um modelo de metadados para a descrição de dados de investigação em formato de imagem? Teria interesse em utilizar?

Armazenar

1. Costumam definir orientações sobre como devem estar organizadas as imagens?
2. Como costumam estar organizadas as imagens no local de armazenamento? (pasta do projeto, pasta com dados relacionados, etc..)

3. Nos seus projetos de investigação existe algum local central onde as imagens se encontrem guardadas?

Descobrir/Partilhar

1. Se partilha as imagens, com quem costuma partilhar e qual a motivação para a partilha?
2. Alguma vez depositou as imagens utilizadas na investigação em algum repositório ou arquivo de imagens?
 - a. Se sim, pode explicar como foi o processo?
 - b. Se não, qual o motivo para não o fazer?

ANEXO III – Documento para importação no Dendro.

```
/**
 * Elements of the dadosimageticos
 */
Elements.dadosimageticos =
{
  amostra :
  {
    type : DbConnection.string,
    control : Config.controls.markdown_box
  },
  autor :
  {
    type : DbConnection.string,
    control : Config.controls.input_box
  },
  coberturaEspacial :
  {
    type : DbConnection.string,
    control : Config.controls.input_box
  },
  data :
  {
    type : DbConnection.string,
    control : Config.controls.input_box
  },
  descricao :
  {
    type : DbConnection.string,
    control : Config.controls.markdown_box
  },
  direitosUtilizacao :
  {
    type : DbConnection.string,
```

```

        control : Config.controls.input_box
    },
    escala :
    {
        type : DbConnection.string,
        control : Config.controls.input_box
    },
    fonte :
    {
        type : DbConnection.string,
        control : Config.controls.input_box
    },
    formato :
    {
        type : DbConnection.string,
        control : Config.controls.input_box
    },
    instrumento :
    {
        type : DbConnection.string,
        control : Config.controls.input_box
    },
    metodologia :
    {
        type : DbConnection.string,
        control : Config.controls.markdown_box
    },
    periodoTemporal :
    {
        type : DbConnection.string,
        control : Config.controls.input_box
    },
    qualidade :
    {

```

```

    type : DbConnection.string,
    control : Config.controls.markdown_box
  },
  tipo
  {
    type : DbConnection.string,
    control : Config.controls.input_box
  },
  titulo:
  {
    type : DbConnection.string,
    control : Config.controls.input_box
  },
}

```

ANEXO IV – Avaliação do modelo de descrição de dados em formato de imagem

No seguimento da entrevista realizada e como foi mencionado na sessão, uma das finalidades da entrevista foi contribuir para o desenvolvimento de um modelo de metadados para a descrição de dados em formato de imagem.

O objetivo a atingir com esta atividade é recolher feedback relativo à utilização do modelo de metadados, através da descrição de uma imagem exemplo utilizada na investigação.

Os passos a realizar são os seguintes (no máximo devem demorar 10 minutos a realizar):

- 1 - Escolher uma imagem utilizada como dado na sua investigação. A imagem escolhida não tem que revelar detalhes sobre a investigação.
- 2 - Realizar a descrição de acordo com os descritores fornecidos.
- 3 - Responder a 5 breves perguntas no final relativas ao modelo de metadados.

No link seguinte encontra-se um exemplo prático e documentação de apoio à descrição que deve ser lida antes de a imagem ser descrita:
https://drive.google.com/file/d/1r2nModo4z1f_zxtpJT3iu3KeoilXXTua/view?usp=sharing

Nome do investigador:

Para contextualizar a descrição, e se consentir, insira a sua imagem aqui (Apenas será visualizada por mim:

DESCRITORES

Descritor	Descrição	Informação sobre a imagem
Título	Nome dado ao conjunto de imagens	
Autor	Entidade responsável pela autoria do recurso	
Cobertura espacial	Características espaciais relevantes para o conjunto de imagens. Pode ser um local ou as coordenadas geográficas	
Período temporal	Um período temporal relevante para o conjunto de imagens. Pode ser uma data ou período de tempo	
Data	Um ponto ou período de tempo associado a um evento no ciclo de vida da imagem	

Descrição	Uma descrição sobre a imagem	
Metodologia	Metodologia e o processamento envolvidos na produção da imagem	
Direitos de utilização	Uma declaração sobre os direitos de propriedade associados à imagem	
Fonte	Um recurso de onde a imagem deriva	
Formato	Formato do ficheiro ou meio físico em que se encontra a imagem	
Tipo	Natureza ou género da imagem	
Instrumento	O nome do instrumento usado para produzir a imagem	
Escala	Especifica os valores numéricos ou textuais definidos na escala de medida	
Qualidade	Informação sobre a qualidade dos dados ou quaisquer procedimentos de garantia de qualidade seguidos na produção da imagem	
Recursos relacionados	Link ou apontamento para um recurso estritamente relacionado com imagem	
Amostra	Descreve a amostra utilizada na experiência	
Edição	Uma declaração sobre as edições/manipulações realizadas na imagem	

AVALIAÇÃO DO MODELO DE METADADOS

1. Quais os metadados mais úteis?

1.1. Qual o motivo pelo qual escolheu estes descritores?

2. Quais os metadados menos úteis?

2.1. Qual o motivo pelo qual escolheu estes descritores?

3. Durante a descrição encontrou alguma dificuldade na utilização do modelo de metadados?

4. Na sua opinião, o modelo apresentado é útil e adequado para a descrição das imagens?

5. Acrescentaria algum descritor ao modelo? Se sim, quais?