UNIVERSIDADE DE BRASÍLIA
FACULDADE DE CIÊNCIA DA INFORMAÇÃO
CURSO DE GRADUAÇÃO EM BIBLIOTECONOMIA

LUCAS HENRIQUE ALVES DA SILVA

GESTÃO DE DADOS CIENTÍFICOS SOB A ÓTICA DA CIÊNCIA ABERTA:
UMA ANÁLISE DE PLANOS DE GESTÃO DE DADOS EUROPEUS

BRASÍLIA
2019

Lucas Henrique Alves da Silva

GESTÃO DE DADOS CIENTÍFICOS SOB A ÓTICA DA CIÊNCIA ABERTA: UMA ANÁLISE DE PLANOS DE GESTÃO DE DADOS EUROPEUS

Monografia apresentada à Faculdade de Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de Bacharel em Biblioteconomia.

Orientadora: Prof.ª Dr.ª Fernanda Passini Moreno

Brasília
2019

S586g Silva, Lucas Henrique Alves da, 1997-

Gestão de dados científicos sob a ótica da Ciência Aberta: uma análise de planos de gestão de dados europeus / Lucas Henrique Alves da Silva. - Brasília, 2019.

115 f. : il.

Orientadora: Prof.ª Dr.ª Fernanda Passini Moreno.

Monografia (Bacharelado em Biblioteconomia) - Universidade de Brasília, Faculdade de Ciência da Informação, 2019.

Bibliografia: p. 83-89.

1. Comunicação científica. 2. Ciência aberta. 3. Dados científicos. 4. Gestão de dados científicos. I. Moreno, Fernanda Passini, orient. II. Título.

CDU: 002:004

Título: Gestão de dados científicos sob a ótica da ciência aberta: uma análise de planos de gestão de dados europeus.

Aluno: Lucas Henrique Alves da Silva.

Monografia apresentada à Faculdade de Ciência da Informação da Universidade de Brasília, como parte dos requisitos para obtenção do grau de Bacharel em Biblioteconomia.

Brasília, 20 de setembro de 2019.

Fernanda Passini Moreno - Orientadora
Professora da Faculdade de Ciência da Informação (FCI/UnB)
Doutora em Ciência da Informação

João de Melo Maricato - Membro
Professor da Faculdade de Ciência da Informação (FCI/UnB)
Doutor em Ciência da Informação

Michelli Pereira da Costa - Membro
Professora da Faculdade de Ciência da Informação (FCI/UnB)
Doutora em Ciência da Informação

Ao meu pai, Lindomar Bento,
minha inspiração de resiliência e determinação,
que fez todos os esforços para que eu tivesse uma boa educação
e que sempre me apoiou em todas as minhas decisões.

AGRADECIMENTOS

Agradeço aos meus pais, **Vânia Leite** e **Lindomar Bento**, cujo amor, dedicação e apoio incondicional foram primordiais para que eu conseguisse percorrer o trajeto que me proporcionou experiências e conquistas.

À minha irmã, **Letícia Hellen**, por ser sempre uma pessoa doce e companheira, alguém que eu sinto que inspiro e sou inspirado.

À minha família, por sempre ter acreditado em mim e me fazer sentir amado e querido durante toda a minha vida, apesar do meu jeito esquisito de ser e da distância geográfica.

À minha amiga **Lívia Santos**, que esteve comigo durante toda a tempestade de 2012, prova viva de que o tempo e a distância não desfazem amizades verdadeiras.

À minha amiga **Joyce Wneuryann**, que também surgiu na minha vida em 2012, mostrando que eu não estava sozinho.

À minha orientadora, **Fernanda Moreno**, que com sua competência profissional e imensa paciência me orientou tanto na iniciação científica quanto neste trabalho de conclusão de curso, compartilhando comigo ideias e conhecimentos novos que aumentaram ainda mais meu interesse pela Biblioteconomia e pela Ciência da Informação.

A **Augusto Pippi**, com quem compartilhei uma filosofia de vida extremamente semelhante, e que me ajudou a impulsionar minha produção acadêmica na reta final deste trabalho, no momento em que eu mais precisava de concentração e foco. Obrigado por tudo.

Aos amigos que fiz durante a graduação:

Yonara Karine, com quem pude compartilhar alegrias, angústias, anseios e sentimentos; que tornou minha vida mais feliz e leve com sua personalidade única e espontânea; que me inspirou em vários aspectos; e que me ouviu, aconselhou e estendeu a mão sempre que necessitei.

Cristiane Mendes, que esteve do meu lado em meu momento mais turbulento em Brasília, que me acolheu como uma irmã, e com quem estabeleci uma conexão inexplicável.

Denise Oliveira, baixinha e *baby* como eu, que sempre demonstrou disponibilidade para me ouvir, me compreender e me aconselhar, capaz de perceber o que eu estava sentindo nos meus gestos mais sutis. Amiga nos quatro anos de graduação, dois anos de estágio e para a vida toda.

Larissa de Araújo, pela amizade durante esses anos, e por ser um exemplo de dedicação, meticulosidade, cautela e sensatez para mim.

Beatriz Santos, que entrou na minha vida já nos primeiros dias de aula no primeiro semestre, e cuja serenidade tornou meus dias mais suaves e tranquilos.

Alcemir Gomes, pela mais filosófica e turbulenta amizade entre Câncer e Escorpião e INFJ-ENFJ que o mundo já concebeu. Obrigado por ter me ouvido e me compreendido tão acuradamente, por me decifrar sem que eu precisasse pronunciar uma palavra, por levantar meu astral mesmo eu tendo sempre sete pedras na mão. Mesmo que a gente discuta e passe meses sem se falar um com o outro, ainda assim vamos saber que a amizade continua firme e forte.

Aos meus colegas de trabalho da Biblioteca do Superior Tribunal de Justiça, onde tive o prazer de estagiar durante dois anos:

Allan Rafael, **Roberta Marins** e **Betânia Lima**, bibliotecários incríveis que muito me ensinaram tanto sobre a profissão quanto sobre os mais diversos aspectos da vida; **Patrícia Rabello**, por inspirar sempre um comportamento pautado na boa educação, no decoro e no respeito; **Taty Bu** e **Jeolane Marinho**, pelas companhias agradáveis nas minhas tardes; **Tauane Fonseca**, pelas inúmeras conversas a respeito de vários assuntos, e pela amizade que criamos dentro de pouco tempo de convivência; **Letícia Cintra**, com quem pude trabalhar junto durante um ano, pela empatia e gentileza demonstrada para comigo durante todo o tempo em que trabalhamos juntos; e **Lara Cristina**, prova viva de que uma amizade maravilhosa pode surgir mesmo sem quase nada em comum.

Obrigado a todos vocês por fazerem parte da minha vida e por terem contribuído à sua maneira na minha trajetória acadêmica, profissional e pessoal.

"Estamos vivendo em uma era técnica. Muitas pessoas estão convencidas de que a ciência e a tecnologia encerram as respostas para todas as nossas perguntas. Nós apenas deveríamos deixar os cientistas e técnicos prosseguirem com seu trabalho, e eles criarão o céu aqui na terra."

Yuval Noah Harari - Sapiens

Meu Deus, olhe para o que somos agora,
Sem arrependimento por todas as coisas que fizemos.
Obrigado por todas as dúvidas, e por todos os questionamentos,
Por toda a solidão e por todo o sofrimento,
Por todo o vazio e pelas cicatrizes deixadas por dentro.
Isso me inspirou, um ímpeto para lutar.
Obrigado pela convicção, pelo propósito encontrado junto,
Pela força e coragem, isso em mim eu nunca tinha descoberto.
Às vezes eu desejo que você pudesse me ver agora,
No lugar certo, onde eu sabia que pertencia.
Às vezes eu desejo que você possa algum dia entender,
Fechar o capítulo e esquecer o passado.
Mas nada mudaria, pois fazemos o melhor que podemos,
E somos medidos pelas nossas ações.
Vamos esperar nossa vez e deixar o futuro nos mostrar como imortais,
Em grandes lendas a serem contadas.

VNV Nation – Gratitude (2011), tradução livre

RESUMO

A gestão de dados de pesquisa é uma temática que tem ganhado cada vez mais projeção na era do quarto paradigma da Ciência, caracterizado pelo grande uso de tecnologia da informação para o armazenamento, preservação e compartilhamento de dados científicos. Com base nisso, esta pesquisa teve como objetivo geral analisar planos de gestão de dados elaborados por pesquisadores europeus, buscando compreender como esses documentos são estruturados. Para isso, realizou-se uma revisão da literatura nacional e estrangeira referente à Comunicação Científica, Ciência Aberta, *e-Science* e dados de pesquisa nos seus mais diversos aspectos. Selecionaram-se cinco planos de gestão de dados de diferentes entidades financiadoras de pesquisas para análise e comparação, a partir da plataforma do DCC. A metodologia empregada foi qualitativa, descritiva e documental. Constatou-se que os planos de gestão de dados abordam aspectos relativos principalmente aos seguintes tópicos: coleta de dados, metadados, armazenamento, preservação, compartilhamento, ética e direitos de propriedade intelectual. Concluiu-se que a devida gestão de dados científicos proporciona benefícios como economia de tempo e recursos, transparência científica e reutilização de dados, cabendo aos profissionais da informação oferecer suporte aos pesquisadores nas práticas de gestão de dados, em âmbito institucional.

Palavras-chave: Comunicação científica. Ciência aberta. Gestão de dados científicos. Plano de gestão de dados. Compartilhamento de dados.

ABSTRACT

Research data management is a theme that has increasingly gained prominence in the era of the fourth science paradigm, characterized by the great use of information technology for the storage, preservation and sharing of scientific data. Based on this, this research aimed to analyze data management plans prepared by European researchers, seeking to understand how these documents are structured. To this end, a review of the national and foreign literature on Scientific Communication, Open Science, e-Science and research data in its various aspects was performed. Five data management plans from different research funders were selected for analysis and comparison from the DCC platform. The methodology employed was qualitative, descriptive and documentary. It was found that data management plans address aspects relating mainly to the following topics: data collection, metadata, storage, preservation, sharing, ethics and intellectual property rights. Thus, it can be concluded that proper management of scientific data provides benefits such as time and resource saving, scientific transparency and data reuse.

Keywords: Scientific communication. Open science. Research data management. Data management plan. Data sharing.

LISTA DE FIGURAS

FIGURA 1 – ESCOPO TEMÁTICO DO ESTUDO 20
FIGURA 2 – RAZÕES PARA O COMPARTILHAMENTO DE DADOS DE PESQUISA 31
FIGURA 3 – PRINCIPAIS BARREIRAS AO COMPARTILHAMENTO DE DADOS DE PESQUISA 32
FIGURA 4 – MODELO DE CITAÇÃO DE CONJUNTOS DE DADOS EM TRABALHOS ACADÊMICOS 37
FIGURA 5 – FATORES QUE EXERCEM INFLUÊNCIA SOBRE A PRÁTICA DE REUSO DE DADOS 38
FIGURA 6 – MODELO DE CICLO DE VIDA DOS DADOS DA UNIVERSITY OF VIRGINIA 39
FIGURA 7 – MODELO DE CICLO DE VIDA DOS DADOS DO DATAONE 40
FIGURA 8 – FLUXO DA GESTÃO DE DADOS DE PESQUISA EM INSTITUIÇÕES ACADÊMICAS 42
FIGURA 9 – EXEMPLO DE INFORMAÇÕES RELEVANTES SOBRE OS DADOS DE UMA PESQUISA 75
FIGURA 10 – RESPONSÁVEIS PELO ARMAZENAMENTO DOS DADOS DO PROJETO AUTOPOST 77

LISTA DE QUADROS

QUADRO 1 – ETAPAS DO CICLO DE VIDA DOS DADOS 40
QUADRO 2 – BENEFÍCIOS DOS REPOSITÓRIOS DE DADOS DE PESQUISA 47
QUADRO 3 – FUNÇÕES DO BIBLIOTECÁRIO NO AUXÍLIO À COMUNIDADE CIENTÍFICA 48
QUADRO 4 – CARACTERIZAÇÃO DA PESQUISA 51
QUADRO 5 – PROCEDIMENTOS METODOLÓGICOS 53
QUADRO 6 – *CHECKLIST* DE ELABORAÇÃO DE UM PGD 54
QUADRO 7 – *CHECKLIST* DO PGD DO AHRC 59
QUADRO 8 – *CHECKLIST* DO PGD DO BBSRC 62
QUADRO 9 – *CHECKLIST* DO PGD DO EPSRC 65
QUADRO 10 – *CHECKLIST* DO PGD DO ESRC 68
QUADRO 11 – *CHECKLIST* DO PGD DO H2020 73
QUADRO 12 – QUADRO COMPARATIVO DOS PGDS 79

LISTA DE ABREVIATURAS E SIGLAS

AHRC	*Arts and Humanities Research Council*
APA	*American Psychological Association*
BBSRC	*Biotechnology and Biological Sciences Research Council*
BRAPCI	Base de Dados em Ciência da Informação
DataONE	*Data Observation Network for Earth*
DCC	*Digital Curation Centre*
DOI	*Digital Object Identifier*
ENANCIB	Encontro Nacional de Pesquisa em Ciência da Informação
EPSRC	*Engineering and Physical Sciences Research Council*
ESRC	*Economic and Social Research Council*
FAIR	*Findable, Accessible, Interoperable and Reusable*
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FIPS	*Federal Information Processing Standard*
FOSTER	*Facilitate Open Science Training for European Research*
FTP	*File Transfer Protocol*
H2020	*Horizon 2020*
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
ISO	International Organization for Standardization
OAI	*Open Archives Initiative*
OAI-PMH	*Open Archives Initiative Protocol for Metadata Harvesting*
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
ORCID	*Open Researcher and Contributor Identification*
PGD	Plano de Gestão de Dados
ProIC	Programa de Iniciação Científica
RDA	*Research Data Alliance*
Re3data	*Registry of Research Data Repositories*
UE	União Europeia
UnB	Universidade de Brasília
UFPE	Universidade Federal de Pernambuco
UFRGS	Universidade Federal do Rio Grande do Sul
UKDA	*UK Data Archive*
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO 16

1.1 OBJETIVOS 17

1.1.1 OBJETIVO GERAL 17

1.1.2 OBJETIVOS ESPECÍFICOS 17

1.2 JUSTIFICATIVA 17

2 REVISÃO DE LITERATURA 20

2.1 O SISTEMA DA COMUNICAÇÃO CIENTÍFICA 21

2.2 OS PRINCÍPIOS DA CIÊNCIA ABERTA 23

2.3 O PARADIGMA DA *E-SCIENCE* 25

2.4 A PROBLEMÁTICA DOS DADOS DE PESQUISA 27

2.4.1 COMPARTILHAMENTO DE DADOS CIENTÍFICOS 29

2.4.2 A DISCUSSÃO EM TORNO DOS DADOS ABERTOS 33

2.4.3 TIPOLOGIA DOS DADOS DE PESQUISA 34

2.4.4 REUSO DE DADOS, METADADOS E CITAÇÃO 35

2.5 A GESTÃO DE DADOS DE PESQUISA E O CICLO DE VIDA DOS DADOS 39

2.6 PLANO DE GESTÃO DE DADOS: CONCEITO E FUNÇÕES 43

2.7 BIBLIOTECAS UNIVERSITÁRIAS E SERVIÇOS RELACIONADOS À GESTÃO DE DADOS 45

2.8 REPOSITÓRIOS DE DADOS CIENTÍFICOS 46

2.9 A ATUAÇÃO DO BIBLIOTECÁRIO NA GESTÃO DE DADOS DE PESQUISA 48

3 PROCEDIMENTOS METODOLÓGICOS 50

3.1 CARACTERIZAÇÃO DA PESQUISA 50

3.2 FONTES DE INFORMAÇÃO 51

3.3 AMBIENTE DE PESQUISA 52

3.4 *CHECKLIST* PARA ANÁLISE DOS PLANOS DE GESTÃO DE DADOS 53

4 ANÁLISE DOS PLANOS DE GESTÃO DE DADOS 58

4.1 ANÁLISE DOS DOCUMENTOS 58

4.1.1 PGD DO ARTS AND HUMANITIES RESEARCH COUNCIL 58

4.1.2 PGD DO BIOTECHNOLOGY AND BIOLOGICAL SCIENCES RESEARCH COUNCIL 61

4.1.3 PGD DO ENGINEERING AND PHYSICAL SCIENCES RESEARCH COUNCIL 64

4.1.4 PGD DO ECONOMIC AND SOCIAL RESEARCH COUNCIL 68

4.1.5 PGD DO HORIZON 2020 72

4.2 ANÁLISE COMPARATIVA DOS DOCUMENTOS 77
5 DISCUSSÃO DOS RESULTADOS 80
6 CONSIDERAÇÕES FINAIS 82
REFERÊNCIAS 83
ANEXO A – PGD DO AHRC 90
ANEXO B – PGD DO BBSRC 94
ANEXO C – PGD DO EPSRC 97
ANEXO D – PGD DO ESRC 100
ANEXO E – PGD DO H2020 103

1 INTRODUÇÃO

Com o crescimento da produção científica e a conscientização de seu impacto no avanço do conhecimento e da própria sociedade, especialmente a partir da segunda metade do século XX, emergiu a necessidade de se promover uma ampliação do acesso à Ciência e ao conhecimento científico, além de uma preocupação crescente com a transparência, o armazenamento e a preservação dos dados coletados pelas pesquisas científicas, muitas vezes financiadas com recursos públicos. Com o objetivo de democratizar o acesso ao conhecimento e incentivar a prática de compartilhamento de dados científicos, surgiu o movimento da Ciência Aberta (*Open Science*).

Como produto primário da Ciência, os dados de pesquisa ganham notoriedade especial num contexto de constante evolução tecnológica, otimização de recursos e esforços, e cooperação científica, o que acarreta a necessidade de tomar medidas de gestão desses conjuntos de dados coletados nas investigações científicas, sejam elas realizadas no âmbito das Ciências Humanas e Sociais, Exatas ou Biológicas.

Em vista disso, a gestão de dados de pesquisa desponta como um procedimento fundamental para coleta, armazenamento, preservação e disseminação dos dados oriundos da pesquisa, o que implica a necessidade dos pesquisadores de elaborar um plano de gestão de dados, já exigido por muitas entidades financiadoras europeias, cuja prática tende a se expandir pelos outros continentes, o que constitui uma modernização da prática científica, alinhada aos propósitos do quarto paradigma da Ciência.

Este estudo busca realizar uma análise dos planos de gestão de dados elaborados por pesquisadores europeus, especificamente britânicos, em virtude das exigências das agências de fomento, tendo como finalidade compreender como esses planos são estruturados, quais tipos de informação são descritos e identificar categorias que permitam delimitar um padrão de plano de gestão de dados eficaz.

Na primeira etapa deste trabalho, realizou-se uma pesquisa bibliográfica no intuito de situar conceitualmente e compreender o fenômeno da gestão de dados de pesquisa. O *corpus* bibliográfico foi composto por artigos de periódicos nacionais e estrangeiros, anais de congressos brasileiros, dissertações, teses, livros, *websites* e documentos de universidades e entidades de apoio à Ciência Aberta. Tais documentos foram recuperados principalmente a partir de pesquisas na base de dados *Library Literature & Information Science Full Text*, nos portais de eventos do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB), além das referências pertinentes indicadas nos textos lidos. Foram realizadas também pesquisas no

Google, focando-se nas informações fornecidas por plataformas especializadas nas temáticas da Ciência Aberta e da gestão de dados científicos, tais como o *Digital Curation Centre* (DCC), *Data Observation Network for Earth* (DataONE), FOSTER, entre outros.

O presente trabalho estrutura-se da seguinte maneira: 1) no capítulo de Introdução, contextualiza-se o tema do estudo, define-se o problema a ser estudado, elencam-se os objetivos geral e específicos a serem cumpridos na consecução deste trabalho, e apresenta-se uma justificativa sobre a relevância do estudo a ser realizado; 2) na seção de Revisão de Literatura, discute-se acerca dos conceitos gerais e específicos associados à temática do trabalho, como Comunicação Científica, Ciência Aberta, *e-Science*, dados de pesquisa, gestão de dados de pesquisa, plano de gestão de dados, ciclo de vida dos dados e repositórios de dados científicos; 3) na seção de Procedimentos Metodológicos, caracteriza-se a pesquisa realizada; 4) no capítulo de Análise e Discussão dos Resultados, apresentam-se e examinam-se os planos de gestão de dados selecionados para o estudo; 5) por fim, na seção de Considerações Finais, levantam-se os principais aspectos identificados e as conclusões formadas após a realização do estudo.

1.1 Objetivos

1.1.1 Objetivo geral

- Analisar planos de gestão de dados elaborados por pesquisadores europeus, buscando compreender como esses documentos são estruturados.

1.1.2 Objetivos específicos

- Analisar planos de gestão de dados indicados pelo DCC, identificando os tipos de informação considerados relevantes;
- Comparar os planos de gestão de dados analisados, observando a consonância entre os documentos examinados e os quesitos elencados pelo DCC.

1.2 Justificativa

Obedecendo ao curso evolutivo dos fatos, dos objetos, das sociedades e do conhecimento, a Ciência tem sofrido profundas transformações em seu sistema, especialmente em

virtude tanto das oportunidades oferecidas pelas novas tecnologias quanto das demandas impostas por fatores internos e externos à comunidade científica.

Atualmente marcada por um grande intercâmbio e compartilhamento de dados entre pesquisadores, pela investigação colaborativa e pelo acesso cada vez menos restritivo à informação – práticas que prometem ser ainda mais estimuladas nas décadas de 2020 e posteriores –, a Ciência tem apresentado novas facetas que merecem ser analisadas, como a gestão de dados científicos.

Durante a graduação, o autor deste trabalho teve contato com a temática da gestão de dados de pesquisa a partir da participação no Programa de Iniciação Científica da universidade (ProIC/UnB), no qual realizou-se uma pesquisa que buscou investigar os serviços de apoio à gestão de dados científicos oferecidos por bibliotecas universitárias espanholas. Ao longo da pesquisa realizada, despontou-se o interesse em estender o estudo a respeito do assunto, haja vista a recência do tema, a tendência do florescimento de sua discussão no meio acadêmico e os incentivos que vêm sendo tomados internacionalmente – e nacionalmente, ainda que em fase incipiente – acerca da preservação e do compartilhamento de dados brutos de pesquisa.

Como se evidencia na literatura, "dados de pesquisa historicamente não têm recebido o cuidado, em termos de indexação, arquivamento e disseminação" (SALES; SAYÃO, 2018, p. 4186). Em decorrência disso, várias iniciativas têm sido tomadas nas últimas duas décadas, especialmente nos contextos europeu e norte-americano, em torno da gestão de dados de pesquisa, a qual desenha um ciclo de etapas mais ou menos bem definidas que devem ser cumpridas pelos pesquisadores, conjuntamente às instituições acadêmicas, a fim de garantir boas práticas de coleta, arquivamento, preservação e compartilhamento de dados.

Na nova era da Ciência, também denominada de quarto paradigma – que será discutido ao longo deste trabalho –, os dados de pesquisa ganharam um protagonismo muito maior em comparação com outrora, quando eram mal arquivados, muitas vezes perdidos ou destruídos com o passar do tempo, e vistos como atores secundários no produto final das investigações científicas.

Na atualidade, a Ciência clama mais do que nunca por transparência, integridade, otimização de tempo e recursos, cooperação, entre outros requisitos. Nesse contexto, os dados primários de pesquisa têm recebido maior destaque, não só por permitir o reuso dos dados por outros pesquisadores, mas também por garantir maior economia, transparência e validação à Ciência. Por essa razão, a efetiva e responsável gestão desses dados apresenta-se como fundamental tanto à modernização do modo de fazer Ciência quanto para a memória acadêmica, tendo em vista as importantes funções exercidas pelos repositórios de dados.

Apesar de no Brasil a prática de gestão de dados científicos ainda ser incipiente comparado ao cenário estrangeiro, na Europa muitas iniciativas têm sido tomadas a respeito, com muitas entidades financiadoras – a exemplo do programa *Horizon 2020* – exigindo previamente dos pesquisadores a elaboração de um plano de gestão de dados detalhando questões como tipos e formatos de dados a serem coletados, propostas de arquivamento, preservação e compartilhamento, além de especificações referentes às licenças de uso e direitos autorais. No caso brasileiro, podem-se citar iniciativas recentes como as do IBICT, que publicou o "Plano de Dados Abertos 2017-2018", visando contribuir para a abertura de dados na instituição e para a transparência na administração pública, e a da FAPESP, que determinou a obrigatoriedade de se anexar um plano de gestão de dados a certas modalidades de projeto e chamadas de propostas (HENNING *et al.*, 2019).

À vista disso, este trabalho visa contribuir para o crescimento da discussão a respeito da gestão de dados científicos no cenário brasileiro, estreitando um pouco o foco na direção dos planos de gestão de dados, que se interpõem como um instrumento primordial na consecução das etapas de gestão de dados.

2 REVISÃO DE LITERATURA

Para a contextualização disciplinar e compreensão conceitual da temática escolhida para se abordar neste estudo, buscou-se realizar uma revisão de literatura que compreendesse tanto os conceitos mais genéricos quanto aqueles mais específicos, relacionados ao tema em questão. De acordo com Creswell (2010, p. 51), a revisão de literatura é uma etapa da pesquisa que cumpre propósitos importantes como os de compartilhar "com o leitor os resultados de outros estudos que estão intimamente relacionados àquele que está sendo realizado", além de proporcionar uma estrutura teórica capaz de evidenciar a relevância da pesquisa. A fundamentação teórica em um estudo é essencial para explicar ou compreender os fenômenos ou processos que se pretendem investigar (MINAYO, 2009).

A Figura 1 sintetiza a esquematização teórica deste trabalho, embora caiba ressaltar que um conceito pode não se encontrar totalmente inserido em outro, mas sim trata-se de uma espécie de funil no qual vários elementos se misturam, fazendo com que falar de um implica discutir também sobre os outros.

FIGURA 1 – Escopo temático do estudo

Fonte: Elaboração própria.

Nas seções a seguir, discute-se acerca da teoria encontrada a partir da pesquisa bibliográfica a respeito dos seguintes tópicos: Comunicação Científica, Ciência Aberta, *e-Science*,

Dados de Pesquisa, Gestão de Dados de Pesquisa, Plano de Gestão de Dados, Ciclo de Vida dos Dados, Repositórios de Dados, e Bibliotecário na Gestão de Dados.

2.1 O sistema da Comunicação Científica

A Ciência não funciona como um sistema hermético que possui um fim em si mesmo. De acordo com Meadows (1999, p. 161), a comunicação dos resultados de uma pesquisa é uma atividade indissociável de seu processo de realização, implicando uma sequência cíclica na qual uma etapa se segue à outra, o que significa dizer que realizar uma pesquisa e não comunicar seus resultados constitui-se uma atividade incompleta e desprovida de sentido. Conforme define Leite (2011, p. 30),

> A comunicação científica está, portanto, inexoravelmente ligada às atividades de produção do conhecimento científico. Tais atividades somente são viabilizadas porque, subjacente à pesquisa propriamente dita, são promovidos fluxos de informação de modo que, mediante processos e estruturas de comunicação científica, é possível que pesquisadores acessem, usem, gerem e disseminem informação continuamente e em uma dinâmica cíclica.

Tendo um crescimento mais evidente a partir do século XVII, quando começaram a surgir os primeiros periódicos científicos, a comunicação científica tornou-se um campo mais complexo e sofreu maiores avanços na segunda metade do século XX, após a Segunda Guerra Mundial, com o aprimoramento das tecnologias de informação e comunicação (SILVA *et al.*, 2017).

Valois *et al.* (1989, p. 28) apontam que emergiram "interesses, necessidades, meios, conteúdos, usuários e contextos que ampliam as possibilidades de penetração e expansão por segmentos e grupos sociais além do *locus* da produção científica". Para os autores, "a comunicação científica, em princípio, não difere em sua estrutura interna das demais trocas de informação entre os agentes de um grupo social". Na mesma linha de pensamento, Katz e Martin (1997), em uma breve análise sobre a sociologia da ciência, afirmam que a ciência é uma instituição social cujo avanço está crucialmente ancorado nas interações entre os cientistas, que passaram a formar cada vez mais redes de colaboração.

Entretanto, a produção científica não pode constituir-se em uma atividade desvinculada do contexto social no qual está inserida. Como postula Targino (2000), o conhecimento científico não deve limitar-se aos círculos restritos de pesquisadores, mas sim alcançar a sociedade como um todo, sendo agente de transformações e intervenções sociais, evitando, dessa forma, tornar-se estéril e inútil. A autora assevera, ainda, que os pesquisadores, como força-

motriz da Ciência, devem trabalhar, acima de tudo, por curiosidade intelectual e, principalmente, em prol do avanço do conhecimento e da humanidade, internalizando a consciência de que "é aético executar investigações científicas exclusivamente por dinheiro ou para garantir posição social, tal como é ilícito deixar que interesses subjetivos interfiram na aceitação ou rejeição de uma ideia científica" (TARGINO, 2000, p. 16).

Meadows (1999) reforça esta noção de que a pesquisa científica - e suas atividades de coletar dados, desenvolver teorias e realizar experiências - está intimamente ligada à interação social. O autor reconhece a importância do fator psicológico no âmbito individual, mas ressalva que a comunicação científica é uma atividade de grupo, destacando o fator sociológico e a natureza cooperativa envolvidos no processo.

Ainda a respeito da sociologia da ciência, Bourdieu (1983, p. 122) afirma que "o universo 'puro' da mais 'pura' ciência é um campo social como outro qualquer, com suas relações de força e monopólios, suas lutas e estratégias, seus interesses e lucros". Em conformidade com tal afirmação, Mueller (2006) constata que o sistema de comunicação científica opera como um jogo de forças no qual se contrapõem interesses financeiros, institucionais, nacionais, políticos, econômicos e pessoais, envolvendo principalmente editoras, institutos de pesquisa, universidades e pesquisadores renomados e iniciantes.

Já introduzindo neste estudo preceitos atinentes à Ciência Aberta, conceito que será explanado nas seções subsequentes, cabe destacar desde cedo a participação do Estado como peça-chave no sistema de produção e comunicação científica, especialmente no caso do Brasil, onde o Estado financia, na maior parte das vezes, a realização de pesquisas, congressos e publicações com recursos públicos (MUELLER, 2006), fato que torna ainda mais patente a necessidade de tornar amplamente acessível não apenas todo esse conhecimento produzido, mas também os dados primários coletados no processo, como será justificado mais adiante.

Convém estabelecer, nesse contexto, uma distinção conceitual e ao mesmo tempo uma relação intrínseca entre a *comunicação científica* - realizada, basicamente, a partir do compartilhamento e avaliação de informações especializadas entre os pares - e a *divulgação científica*, a qual cumpre o papel de democratizar o acesso ao conhecimento científico (BUENO, 2010).

A transmissão da informação científica, conforme aponta Meadows (1999), pode ser comparada à evolução de uma epidemia em determinada população, considerando três atores principais: os propagadores, os curados e os vulneráveis. Contextualizando-se o processo no plano disciplinar em questão neste estudo, pode-se equiparar os propagadores com aqueles que possuem a informação e podem compartilhá-la; os curados seriam aqueles que possuem a

informação, mas não podem disseminá-la por alguma razão (tecnológica, institucional, financeira, ética ou pessoal); e os vulneráveis corresponderiam à parcela de indivíduos que ainda não tiveram acesso à informação, mas podem vir a adquiri-la.

Trazendo uma abordagem mais alinhada à realidade dos tempos atuais, Hurd (2000) faz uma projeção, duas décadas antes, de um modelo de comunicação científica para 2020, caracterizado pela forte interdisciplinaridade da pesquisa científica, pelo uso de grandes conjuntos de dados compartilhados e pela existência de repositórios de dados abertos, conectados em rede, capazes de contribuir para o desenvolvimento da ciência em locais anteriormente marginalizados.

2.2 Os princípios da Ciência Aberta

A Ciência Aberta, ou originalmente *Open Science*, compreende em seu escopo diversos conceitos, práticas e propostas que visam, primordialmente, à ampliação do acesso ao conhecimento científico. Para Delfanti e Pitrelli (2015, p. 59), a Ciência Aberta "é um conceito muito amplo, que engloba diversas práticas e ferramentas ligadas à utilização das tecnologias digitais colaborativas e ferramentas de propriedade intelectual alternativas".

Tendo sua gênese em iniciativas como a *Open Archives Initiative* (OAI) e o *Open Access* - na literatura brasileira também referido como Acesso Aberto -, a Ciência Aberta surgiu como uma providência contra o elevado custo de acesso às publicações veiculadas em periódicos, além da conscientização da primordialidade de tornar públicos os dados e resultados das pesquisas científicas custeadas com recursos do Estado (BAPTISTA; COSTA; KURAMOTO; RODRIGUES, 2007; ROCHA; SALES; SAYÃO, 2017). O movimento do Acesso Aberto preconiza a eliminação das restrições de acesso à informação, contrariamente ao modelo tradicional de comunicação científica, caracterizado pelo alto custo de acesso cobrado pelas grandes bases dados (ANDRADE; MURIEL-TORRADO, 2017).

O conceito de Ciência Aberta, de acordo com Sales e Sayão (2018, p. 4183), sustenta que "todos os processos de investigação científica, fluxos, instrumentos, códigos, resultados e metodologias devem ser os mais abertos possíveis" e, além disso, "pressupõe o uso intensivo de redes de computadores e seus meios de comunicação para o compartilhamento e colaboração". Em consonância com os autores, Hourcade (2015, p. 27) enfatiza que

> A ideia de Ciência Aberta remete a práticas de compartilhamento de dados científicos, informações, resultados de pesquisa, procedimentos, e pode ser compreendida como um processo de construção coletiva do conhecimento, sem as barreiras que limitem seu acesso e reutilização. Um movimento, por-

> tanto, que abarca a noção de abertura da ciência e compartilhamento de conhecimento em vários níveis ou etapas da produção científica, indo do laboratório e do projeto de pesquisa até as publicações com resultados finais ou parciais. [...] o *Open Science* se insere em uma proposta que visa o livre acesso de uma forma geral, o da cultura *Open*, que engloba diversos movimentos, como por exemplo o livre acesso a periódicos científicos, o *Software* Livre, o livre acesso à produção cultural.

O termo *open* tem sido empregado para designar iniciativas que se apresentam como uma alternativa ao enclausuramento do conhecimento científico provocado pelas políticas proprietárias que restringem a cópia, a distribuição e o reuso da informação. Do ponto de vista formal, é fundamental mencionar três documentos basilares para a consolidação do movimento do Acesso Aberto: a *Declaração de Budapeste* (2002), o qual trata do *copyright* como um fator fortemente restritivo ao uso da informação; a *Declaração de Berlim* (2003), que trata principalmente de licenças para cópia, uso e distribuição de obras e uso responsável do conhecimento, com foco nas Ciências e Humanidades; e a *Declaração de Bethesda* (2003), outra assertiva importante nesse contexto, a qual buscou garantir o livre acesso à literatura científica, com ênfase na área biomédica, além de destacar o papel dos repositórios abertos na consecução desse objetivo (MACHADO, 2015; ANDRADE; MURIEL-TORRADO, 2017).

Nesse mesmo patamar, Moreno (2018, p. 53) destaca o "*OECD Principles and Guidelines for Access to Research Data from Public Funding*" como outro documento de referência dentro do movimento do Acesso Aberto, publicado em 2007 pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE), como um reforçamento da ideia de "maximizar o valor do investimento público na ciência e que restrições a esse acesso poderiam diminuir a qualidade e a eficiência da investigação e da inovação científica". Também como um exemplo de promoção da Ciência Aberta, a autora menciona o projeto europeu FOSTER, que busca auxiliar a comunidade acadêmica no que tange aos dados abertos e assuntos correlatos.

A discussão a respeito dos dados abertos ganhou maior projeção com o *workshop* de Sebastopol (Califórnia) em 2007, que elencou oito princípios fundamentais para dados abertos governamentais, os quais postularam que "qualquer dado, para ser aberto, deve ser utilizado por qualquer um para qualquer propósito" (MACHADO, 2015, p. 212). Em contrapartida, é oportuno lembrar que existem os casos de exceção, ou seja, os dados que não podem ser divulgados em acesso aberto em virtude de patenteamento ou outro motivo que justifique sua proteção (COSTA; LEITE, 2017).

Constata-se que a Ciência Aberta, ao mesmo tempo em que contesta o modelo preponderante de produção e divulgação científica – muitas vezes atrelado à privatização do conhecimento –, também busca uma conciliação entre a propriedade privada e o livre acesso à

produção intelectual (HOURCADE, 2015; OLIVEIRA, 2016). O potencial choque entre Acesso Aberto e Direitos Autorais é evidenciado no fato de que, no contexto do Acesso Aberto, atividades como acesso, reprodução, tradução e compartilhamento de dados e informações não deveriam ser ações que tivessem que se deparar com barreiras legais em seu caminho. No entanto, os direitos de propriedade intelectual continuam constituindo um dos principais entraves à democratização do acesso e do uso da informação.

A despeito disso, Oliveira (2016, p. 36) ressalta que a Ciência Aberta, apesar das barreiras que encontra, "conduz a novas pesquisas, investigações e hipóteses a partir do uso, reuso e reprodutibilidade dos dados científicos, sob a égide do quarto paradigma, o qual fomenta o surgimento de um fazer científico baseado no compartilhamento e na colaboração".

Além das implicações discutidas acima, é imprescindível também destacar a contribuição fundamental que as agências de fomento desempenham no estímulo à pesquisa por meio dos financiamentos. Somado a isso, há também o caráter social por trás da atividade científica, posto que a sociedade constitui, incontestavelmente, uma das principais beneficiárias por direito dos frutos gerados pela Ciência, qualquer que seja a área de conhecimento na qual se originem descobertas e inovações, afinal, a mesma investe vultosos recursos que são aplicados nas pesquisas e demanda um retorno por isso (SILVA *et al.*, 2017).

Antes de partir para a discussão dos tópicos mais específicos deste estudo, é pertinente traçar uma relação direta entre acesso aberto e dados abertos, os quais podem ser armazenados, preservados e disseminados por meio de repositórios de dados, que são capazes de proporcionar interoperabilidade e uma maior segurança aos dados, garantindo sua preservação e seu compartilhamento a longo prazo, ao contrário de tempos passados, em que predominavam práticas inadequadas de arquivamento de dados e enorme dispêndio de esforços, recursos e conhecimentos (LAFUENTE, 2006).

2.3 O paradigma da *e-Science*

A transição entre as décadas de 1990 e de 2000 foi marcada por profundas transformações na esfera tecnológica, o que evidentemente provocou impactos diretos no modo de fazer Ciência. Essas mudanças ensejaram a eclosão de um novo paradigma para a Ciência, o da *exploração de dados*, também denominado *e-Science* ou quarto paradigma (precedido pelos paradigmas experimental, teórico e computacional, cronologicamente).

De acordo com Gray (2007, p. 18), o quarto paradigma da Ciência "prevê a captura de dados por instrumentos ou sua geração através de simulações, seguida por um processamento

por *software* e armazenamento da informação ou conhecimento resultantes em computadores".

Em outras palavras, a *e-Science* refere-se a uma infraestrutura computacional intensiva, implantada em rede, que utiliza volumosos conjuntos de dados, assegurando a coleta, o processamento, a preservação, a análise e o armazenamento de grande quantidade de dados, além do compartilhamento dos dados entre equipes de pesquisa, sem barreiras de natureza geográfica (ALBAGLI; APPEL; MACIEL, 2013; OLIVEIRA, 2016).

Albagli, Appel e Maciel (2013) apontam alguns fatores que exercem influência sobre a aderência da *e-Science* às práticas da Ciência Aberta, variáveis sejam elas institucionais, normativas, culturais ou técnicas, destacando-se também os obstáculos operacionais relativos à interoperabilidade entre sistemas e aos padrões de metadados. Os autores salientam que, a despeito da estreita relação entre o aparato estrutural da *e-Science* e as práticas de colaboração científica, nem sempre as pesquisas científicas realizadas sob os desígnios da *e-Science* podem ser caracterizadas como sendo de acesso livre e, por conseguinte, serem produto da Ciência Aberta.

No que se diz respeito aos elementos que constituem esse fenômeno, Oliveira (2016) identifica três pilares sobre os quais se apoiam a denominada *e-Science*, são eles:

I. *Ciclo de vida dos dados*: modelo que garante um efetivo gerenciamento dos dados primários, considerando as três fases da pesquisa: antes, durante e após a coleta dos dados; o ciclo de vida dos dados será discutido com mais detalhes nas seções seguintes;

II. *Ciberinfraestrutura*: base tecnológica sustentável para viabilizar o uso, reuso e reprodutibilidade dos dados de pesquisa;

III. *Compartilhamento*: colaboração entre pesquisadores, profissionais, instituições e países, no que concerne ao compartilhamento dos dados primários.

Ferreira, Villalobos e Moura (2016) ressaltam as novas ferramentas tecnológicas surgidas com o advento da *e-Science* como fortes catalisadoras na transição entre o fazer científico tradicional, alicerçado na investigação isolada (caracterizada por cada pesquisador em sua bolha e pela perda de grande volume de dados primários em seus arquivos pessoais), e o fazer científico contemporâneo, fundamentado na investigação colaborativa (caracterizada pela intensa comunicação entre sujeitos e pela consolidação de melhores práticas de preservação e compartilhamento de dados brutos, principalmente por meio de repositórios).

Na seção subsequente, pretende-se realizar uma discussão mais exaustiva a respeito do assunto mais nuclear deste estudo, que são os dados de pesquisa e suas diversas facetas, implicações, processos, instrumentos e fatores (positivos e negativos) envolvidos.

2.4 A problemática dos dados de pesquisa

A comunidade científica sempre conferiu maior atenção e valor aos resultados e às conclusões das pesquisas do que aos dados primários subjacentes às publicações científicas. Conforme reporta Bell (2011), até o século XX os dados coletados e produzidos nas pesquisas geralmente subsistiam por tempo indeterminado – podendo ser uma curta ou longa duração – ocultados e inacessíveis nos cadernos de laboratório pessoais dos cientistas, sujeitos ao desaparecimento e à destruição, sendo que, na melhor das hipóteses, eram armazenados em mídias magnéticas também suscetíveis à obliteração ou depositados na biblioteca da universidade até que se decidisse pelo seu descarte.

Corroborando este fato, Rocha, Sales e Sayão (2017) apontam que esses registros de dados ainda hoje permanecem escondidos em computadores e mídias pessoais, perenizando uma "cultura de segredo" que, conforme defendem, necessita ser progressivamente superada nos dias atuais.

A discussão referente aos dados de pesquisa no âmbito da Ciência da Informação remonta à década de 1970, porém até o presente momento não se tem observado uma produção científica sistemática a respeito do tema, sendo ainda poucos os estudos que discorrem sobre o assunto no escopo da Comunicação Científica (COSTA; LEITE, 2017).

Dados de pesquisa podem ser definidos como qualquer material resultante da coleta ou geração de dados primários, sejam eles de natureza quantitativa ou qualitativa, ou também dados derivados de fontes pré-existentes, que são analisados no decurso de um projeto de pesquisa (GRANT, 2017). Como exemplos mais frequentes de dados de pesquisa, pode-se citar documentos de texto, cadernos de laboratório, questionários, fotografias, dados estatísticos, dados experimentais, medições, observações anotadas em trabalhos de campo, dados de levantamento, gravações de entrevistas, slides, modelos, metodologias, fluxos de trabalho, procedimentos e protocolos (WILEY, 2014; KRUSE, THESTRUP, 2018).

Como se pode constatar, esses dados podem ser gerados por meio de procedimentos das mais variadas naturezas, podendo assumir diversos formatos diferentes. No contexto da Ciência contemporânea, alguns instrumentos de alta tecnologia têm se destacado na geração de uma massiva quantidade de dados, como telescópios, satélites, aceleradores, supercompu-

tadores, redes de sensores e simuladores (CASTELLI; MANGHI; THANOS, 2013). Todavia, diariamente dados são coletados ou produzidos no seio da pesquisa de todas as áreas do conhecimento, seja as Ciências Exatas, Biológicas ou Humanas e Sociais.

Ao contrário dos séculos passados, no século XXI o novo paradigma da Ciência exige que o vasto volume de dados científicos capturados pelos mais diversos instrumentos seja submetido a uma gestão adequada que garanta sua preservação a longo prazo e a acessibilidade pelo público (BELL, 2011). E, tendo em vista que boa parte dos recursos investidos em pesquisa é de origem pública, esse obscurecimento dos dados científicos precisa ceder lugar às práticas de reutilização de dados e colaboração mútua entre pesquisadores e instituições (ROCHA; SALES; SAYÃO, 2017).

O protagonismo assumido pelos dados primários no meio científico desde a virada do terceiro milênio é um fenômeno que tem caracterizado de modo marcante a pesquisa da era da *e-Science*, chegando a assumir os moldes de uma revolução que "é catalisada por novos métodos, instrumentos e ferramentas e está apoiada por progressos alcançados na capacidade computacional, no armazenamento digital e na simulação por programas de computador" (SALES; SAYÃO, 2018, p. 4182).

Após séculos de Ciência, esses dados primários começaram a ser vistos não apenas como meros subprodutos das publicações acadêmicas, mas sim como elementos de elevado valor dentro da comunicação científica, possuindo até mesmo metadados que lhe conferem a possibilidade de serem compartilhados, descobertos, acessados, validados, reutilizados e preservados a longo prazo (CASTELLI; MANGHI; THANOS, 2013). No contexto do quarto paradigma, não apenas os resultados e as conclusões da pesquisa devem ser divulgados, mas também seus dados brutos, anotações e toda a parafernália envolvida no decurso da pesquisa, posto que a prática científica passou a ser mais colaborativa, transparente e célere (SILVA *et al.*, 2017).

A informação veiculada nos artigos de periódicos não são os dados de pesquisa propriamente ditos, mas interpretações destes. De acordo com Lafuente (2006), aproximadamente 80% dos dados gerados em laboratório não são divulgados em nenhum meio acessível, o que segundo o autor beira o "escandaloso", tendo em vista o desperdício de recursos que esse subaproveitamento de dados sugere. Gráficos, tabelas e diagramas são as fisionomias mais comuns que os dados primários assumem nas publicações científicas; no entanto, tais representações são insuficientes quando entram em debate questões relativas à validação e credibilidade dos dados, uma vez que a comunidade científica pode ter o interesse de verificá-los ou contrastá-los com outros dados (LAFUENTE, 2006; CURTY; AVENTURIER, 2017). Con-

forme assinala Meadows (1999, p. 196), “o autor pode omitir pontos de dados que se desviam grandemente da média. Essas alterações, se forem exageradas, podem começar a tocar as raias da fraude”, o que chama a atenção para a polêmica da falsificação de dados.

A respeito desse quesito, Sales e Sayão (2018, p. 4193) frisam que “os dados publicados em versões de alta densidade, como gráficos e tabelas, são produtos finais de pesquisa valiosos, porém ocultam a história completa da pesquisa”.

A necessidade de se publicar os dados originais na íntegra de modo avulso à publicação do artigo em periódicos e eventos também encontra respaldo em Meadows (1999, p. 34), quando este afirma que “um antigo provérbio diz que uma figura vale mais que mil palavras”. Na mesma linha de raciocínio, não seria um equívoco traçar um paralelo e afirmar que um conjunto de dados primários de uma única pesquisa pode comunicar mais do que mil artigos científicos publicados.

2.4.1 Compartilhamento de dados científicos

O marco inicial do compartilhamento de dados de pesquisa por via eletrônica data de aproximadamente 50 anos, quando cientistas da computação começaram a compartilhar arquivos anonimamente por meio do protocolo FTP, utilizado para a transferência de arquivos entre computadores em rede (MACHADO, 2015).

Conforme salientam Costa e Leite (2018, p. 4457), “a comunicação dos dados de pesquisa é apontada como uma das condições para avaliação dos resultados da pesquisa e para o desenvolvimento acelerado, colaborativo e eficaz da ciência como um todo”. Logo, o compartilhamento de dados afigura-se como uma das etapas primordiais da gestão de dados alinhada à filosofia da Ciência Aberta. Um dos principais argumentos em defesa da visibilização dos dados de pesquisa é o de que estes

> [...] constituem matérias-primas importantes para a ecologia da ciência e são essenciais para novos ciclos de criação de conhecimento científico, pois fornecem insumos para um processo iterativo no ciclo de vida da investigação, permitindo a continuidade da descoberta científica e da inovação tecnológica (CURTY; AVENTURIER, 2017, p. 4).

A pesquisa científica assume, historicamente, uma conduta muito mais orientada ao resultado do que ao processo, o que explica sua predileção pelos produtos finais (artigos publicados em periódicos, patentes e protótipos, por exemplo) em detrimento do cuidado com os dados coletados ou produzidos no decorrer da investigação, os quais constituem as testemu-

nhas mais fidedignas e inequívocas da trajetória de erros e acertos trilhada pela Ciência em seu progresso (SAYÃO; SALES, 2016a).

Os artigos publicados em periódicos, os trabalhos apresentados em eventos científicos – mais tarde organizados em anais – e os livros técnico-científicos, apesar de tradicionalmente disseminarem os avanços do conhecimento, fazem-no de modo condensado, muitas vezes privando os dados originais de serem analisados sob outros pontos de vista, em outros contextos, e de dar origem a novas ideias que *a priori* não se cogitavam (CURTY; AVENTURIER, 2017). Não obstante, ainda prevalece o enquadramento dos dados de pesquisa na categoria de material suplementar, que ao contrário das publicações em periódicos, não são passíveis de serem divulgados, descobertos, citados, preservados nem incluídos nos esquemas de recompensa (SALES; SAYÃO, 2018). Contudo, esse cenário tem sofrido mudanças, especialmente com as iniciativas que vêm sendo tomadas para promover a prática de compartilhamento de dados e sua devida citação.

Algumas áreas do conhecimento são mais receptivas do que outras quanto ao compartilhamento de dados, como é o caso da Astronomia, cuja divulgação de dados e metodologias contribui com o próprio desenvolvimento da área, e as Ciências da Saúde, cujas descobertas são de incontestável interesse de toda a população, ao contrário de outros campos, que são mais fechados à ideia de compartilhamento, principalmente por razões culturais e financeiras, como a Engenharia, para citar um exemplo (SAYÃO; SALES, 2016b; COSTA; LEITE, 2017).

Analisando o contexto global, é notável o pioneirismo da Europa nas iniciativas voltadas à comunicação dos dados de pesquisa, como se pode verificar na formulação de princípios e no surgimento de grupos de apoio à Ciência Aberta, enquanto nos Estados Unidos e na América Latina ainda é tímido o engajamento nessa questão, num ponto de vista comparativo (COSTA; LEITE, 2017; MORENO, 2018). Conforme frisa Machado (2015, p. 223),

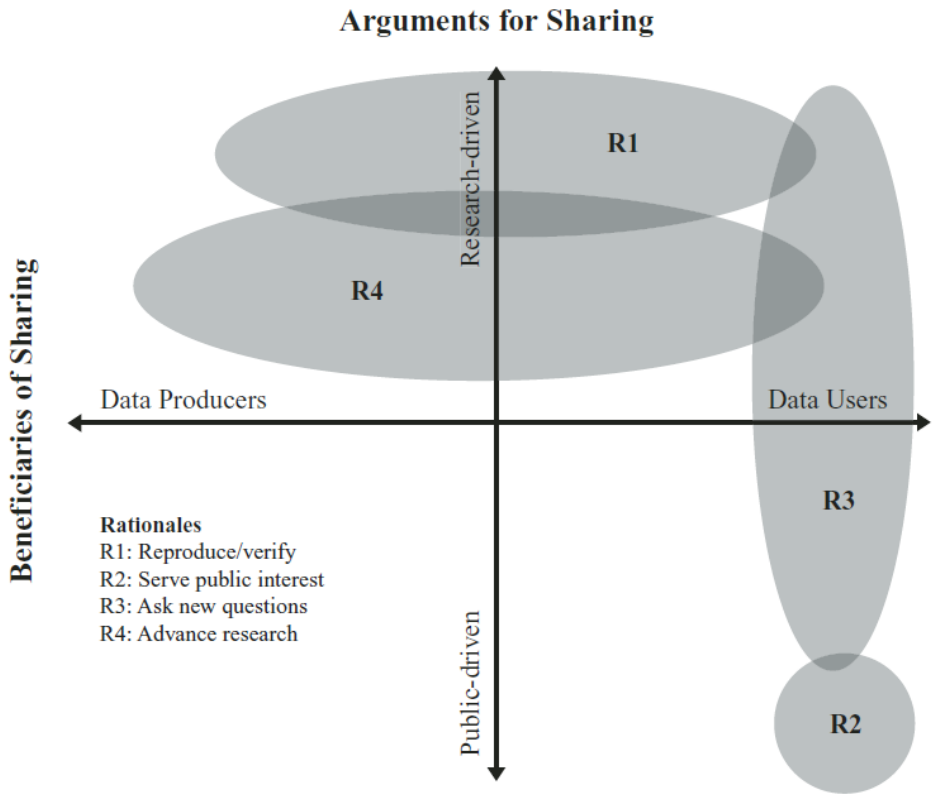
> É necessário um arcabouço legal que sustente e incentive sua disponibilização [dos dados] – como uma lei que garanta o acesso à informação pública e o acesso a dados obtidos com financiamento público. Também são necessárias políticas científicas que apoiem sua disponibilização de forma ativa e sob licenças livres. E há que vencer as resistências culturais, pois abertura de dados e informação tende a reduzir as assimetrias entre os usuários de tais dados, reduzir privilégios entre os que têm acesso e concentram mais informação e conhecimento.

A transição de um paradigma para outro, seja em qual matéria for, normalmente depende de uma conjuntura favorável à renovação. No caso do compartilhamento de dados, a proeminência do assunto ganha força principalmente com incentivos coletivos, legais, institu-

cionais e governamentais que encorajem as atividades nesse sentido, além de eventualidades que surgem para desenterrar as motivações faltantes para conferir mais visibilidade à questão (MACHADO, 2015; OLIVEIRA, 2016). A título de exemplo, podem-se mencionar os surtos dos vírus *zika* e do ebola em 2014, que desencadearam discussões nos anos subsequentes a respeito do compartilhamento de dados para a superação dessas epidemias, culminando no acordo internacional denominado *Statement on Data Sharing in Public Health Emergencies* em 2016, firmado por 33 instituições, incluindo ONGs, institutos de pesquisa e grandes editores científicos (COSTA; LEITE, 2017).

De acordo com Borgman (2012, p. 1066, tradução nossa), "as pressões exercidas para encorajar o compartilhamento de dados surgem de várias origens: agências de financiamento, conselhos de pesquisa, editores de periódicos, professores, o público em geral e os próprios pesquisadores". A autora lista pelo menos quatro razões relatadas com frequência para justificar o compartilhamento de dados, são elas: R1) reprodutibilidade e verificação; R2) interesse público; R3) engendramento de novas questões de pesquisa; e R4) avanço da pesquisa (Figura 2).

FIGURA 2 – Razões para o compartilhamento de dados de pesquisa



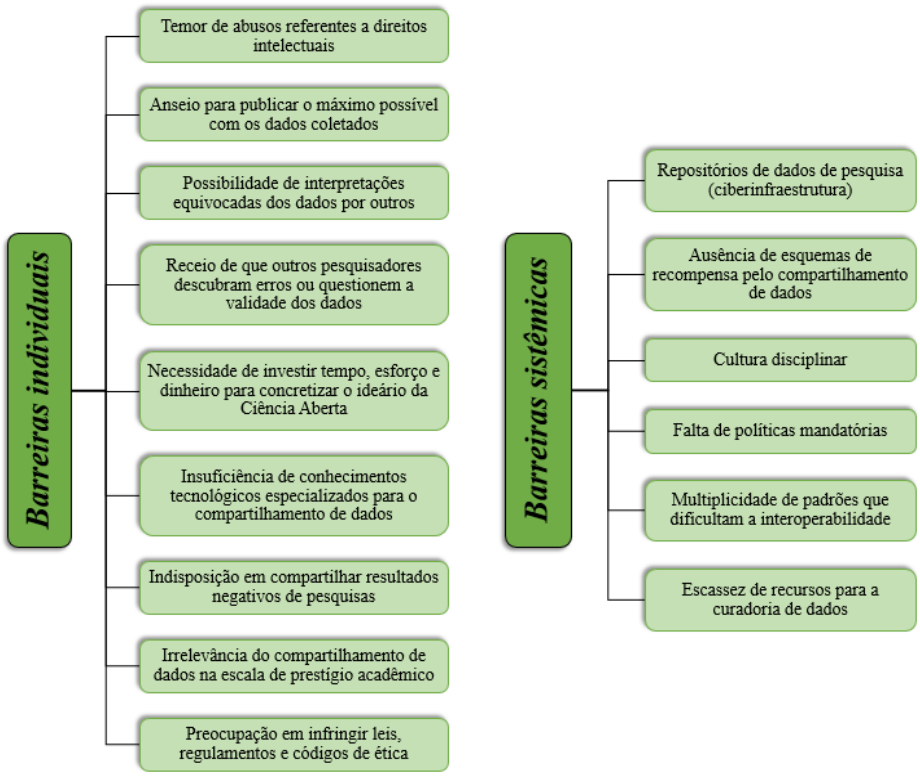
Fonte: Borgman (2012).

Sob essa perspectiva, como se observa no eixo vertical, os argumentos em defesa do compartilhamento podem ser de dois interesses principais: do público ou da pesquisa. No que tange aos beneficiários do compartilhamento, a autora considera dois grupos: os produtores de dados e os usuários de dados.

Nos quadrantes relativos aos produtores e usuários de dados em interseção com o interesse direcionado à pesquisa, inserem-se com grande peso tanto os argumentos em prol da reprodutibilidade e verificabilidade dos dados (R1) quanto o que defende o avanço da pesquisa (R4). Já no quadrante referente ao interesse do público e dos usuários de dados, inserem-se os argumentos defensores do interesse coletivo (R2) e da possibilidade de se lançarem novas perguntas sobre os dados compartilhados (R3). A autora ressalta que a R1 é a mais delicada das quatro razões, devido ao fato de que a criatividade na Ciência muitas vezes consiste na identificação de um novo método para abordar um problema antigo, algo que a simples reprodução de dados pode embaraçar.

Em contrapartida, se por um lado o compartilhamento de dados científicos possui seus argumentos em favor, existem também argumentos em contrário. A Figura 3 sumariza alguns dos principais motivos apontados na literatura que atravancam a prática de compartilhamento.

FIGURA 3 – Principais barreiras ao compartilhamento de dados de pesquisa



Fonte: Elaboração própria com base em Sales e Sayão (2018).

2.4.2 A discussão em torno dos Dados Abertos

Conforme comentado na seção anterior, a abertura dos dados de pesquisa encontra obstáculos que podem ser maiores ou menores de acordo com a área do conhecimento. Em contrapartida, observa-se um esforço global que visa difundir a filosofia dos dados abertos no âmbito da ciência e da tecnologia, arguindo em defesa da visibilidade da pesquisa científica, da otimização de recursos (financeiros, humanos e de tempo), da transparência dos dados e da verificabilidade do conhecimento produzido, buscando salvaguardar os dados de pesquisa como um bem público global, principalmente aqueles gerados a partir de investimentos estatais (BERTIN *et al.*, 2017).

Convém esclarecer, antes de tudo, que a iniciativa dos dados abertos contempla um conjunto de pré-requisitos a serem cumpridos, os quais incluem: disponibilidade, acessibilidade, reuso, redistribuição, participação universal, completeza, primariedade, atualização, legibilidade por máquina, indiscriminação e liberdade de licenças (OPEN KNOWLEDGE FOUNDATION, 2010). Tais pré-requisitos coadunam com os princípios FAIR, acrônimo que, em português, significa dados *localizáveis*, *acessíveis*, *interoperáveis* e *reutilizáveis*. Idealizados em uma conferência internacional em 2014, os princípios FAIR consistem em um conjunto de 15 diretrizes que visam nortear os processos de descoberta, acesso, interoperabilidade e compartilhamento de dados de pesquisa (HENNING *et al.*, 2018).

É válido ressaltar que, nacionalmente, o engajamento na questão dos dados abertos tem muito mais força em matéria de dados governamentais do que de dados científicos. Em termos de legislação, o Brasil promulgou em 2016 o Decreto nº 8.777, que instituiu a Política de Dados Abertos do Poder Executivo Federal, traçando os seguintes objetivos:

> I - promover a publicação de bases de dados de órgãos e entidades da administração pública federal direta, autárquica e fundacional sob a forma de dados abertos;
> II - aprimorar a cultura de transparência pública;
> III - franquear aos cidadãos o acesso, de forma aberta, aos dados produzidos ou acumulados pelo Poder Executivo federal, sobre os quais não recaia vedação expressa de acesso;
> IV - facilitar o intercâmbio de dados entre órgãos e entidades da administração pública federal e as diferentes esferas da federação;
> V - fomentar o controle social e o desenvolvimento de novas tecnologias destinadas à construção de ambiente de gestão pública participativa e democrática e à melhor oferta de serviços públicos para o cidadão;
> VI - fomentar a pesquisa científica de base empírica sobre a gestão pública;
> VII - promover o desenvolvimento tecnológico e a inovação nos setores público e privado e fomentar novos negócios;

> VIII - promover o compartilhamento de recursos de tecnologia da informação, de maneira a evitar a duplicidade de ações e o desperdício de recursos na disseminação de dados e informações; e
> IX - promover a oferta de serviços públicos digitais de forma integrada (BRASIL, 2016).

O decreto determina, ainda, que a implementação desta Política será executada no âmbito de cada órgão ou entidade da Administração Pública Federal, a partir da elaboração de "planos de dados abertos" por essas unidades. A lei ainda conceitua dados abertos como "acessíveis ao público, representados em meio digital, estruturados em formato aberto, processáveis por máquina, referenciados na internet e disponibilizados sob licença aberta que permita sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte" (BRASIL, 2016).

Como exemplo de cumprimento do decreto, pode-se citar a criação do Plano de Dados Abertos da UFPE, fundamentado sobre o ciclo de vida dos dados abertos, nas boas práticas para publicação de dados na *web* e numa sólida estrutura de governança (LÓSCIO; VILA NOVA, 2017).

Todavia, é importante deixar claro que a abertura de dados também encerra alguns desafios, tais como o devido cuidado com aqueles considerados sigilosos por razões de competitividade, como também a garantia da qualidade dos dados publicados, visto que algumas instituições possuem problemas na atualização e na apresentação dos dados, principalmente quando não realizam um planejamento prévio que considere a sustentabilidade da iniciativa de abertura de dados (BERTIN *et al.*, 2017; LÓSCIO; VILA NOVA, 2017). Tais ponderações coincidem com as considerações levantadas por Sayão e Sales (2016a), que chegaram à conclusão de que a disponibilização de dados na *web* sem uma devida contextualização estrutural e semântica podem trazer consequências indesejáveis, como impossibilitar a interpretação, o reuso, a transmissão do conhecimento, a reinterpretação dos dados em outros contextos, além de ocasionar a redução de seu valor na pesquisa interdisciplinar.

2.4.3 Tipologia dos dados de pesquisa

Os dados de pesquisa são classificados na literatura em três categorias, de acordo com sua natureza e forma de coleta, as quais são: dados experimentais, dados computacionais e dados observacionais.

Os *dados experimentais* são produzidos a partir de estudos realizados em laboratório, um ambiente controlado, e incluem desde medições de reações química a estudos comporta-

mentais controlados (BORGMAN, 2012; SAYÃO; SALES, 2016a). A pesquisa experimental é um tipo de trabalho muitas vezes passível de ser reproduzido por outros pesquisadores em ocasiões posteriores à original, desde que sejam conhecidos os instrumentos e as condições necessárias. Em virtude disso, a preservação de dados experimentais a longo prazo precisa contrabalançar os custos do armazenamento e da preservação com os custos da reprodução do experimento (SAYÃO; SALES, 2013).

Os *dados computacionais* são gerados a partir da execução de um modelo ou de uma simulação computacional, seja uma realidade física ou virtual, o que faz com que sua replicação futura demande um rico conjunto de metadados relativos ao *hardware*, ao *software* e aos dados de entrada, o que muitas vezes pode dispensar, portanto, a preservação a longo prazo dos dados da simulação propriamente ditos (BORGMAN, 2012; SAYÃO; SALES, 2013).

Por fim, os *dados observacionais* são aqueles oriundos da observação de fenômenos estudados em lugares e períodos específicos, podendo incluir medições climáticas, pesquisas de comportamento, entrevistas, etnografias, entre outros (BORGMAN, 2012; SAYÃO; SALES, 2016a). É relevante destacar que os dados observacionais, por serem revestidos desse caráter instantâneo e praticamente irrepetível, "guardam uma importância crítica que os qualificam como registros históricos que não podem ser coletados uma segunda vez e, portanto, devem ser arquivados para sempre" (SAYÃO; SALES, 2013, p. 6).

2.4.4 Reuso de dados, metadados e citação

Dados científicos podem ter, em muitas das vezes, uma vida útil muito mais longa do que o projeto de pesquisa que lhes deu origem. Em virtude disso, esses dados podem ser reaproveitados pelo pesquisador mesmo após o término do financiamento, ao mesmo tempo em que podem também ser reutilizados por outros investigadores, desde que sejam devidamente organizados, documentados, preservados e compartilhados, de um modo que permita sua utilização póstuma (UK DATA SERVICE, 2015?).

De acordo com Sales e Sayão (2018, p. 4187), a descontinuidade no registro e na comunicação dos dados de pesquisa "fragiliza a ciência enquanto empreendimento social e humanístico", além de levar automaticamente à "duplicação de esforços que consome recursos e alonga desnecessariamente o ciclo de comunicação científica". Segundo os autores, isso ocorre principalmente porque muitas comunidades ainda não possuem expertise nem estrutura necessárias para um gerenciamento de dados efetivo, e o problema também abrange a falta de mecanismos de recompensa aos pesquisadores que se predispõem a divulgar seus dados.

Quanto a isso, Borgman (2012) ressalta que essas comunidades podem combinar diferentes técnicas para identificar, capturar, descrever, analisar, derivar e interpretar os dados, o que faz com que a coleta e o gerenciamento possam ser abordados de diferentes formas.

A respeito dos obstáculos que se interpõem ao reuso, à documentação e à citação de dados científicos, nota-se que

> O problema é principalmente cultural, já que a mudança de normas comportamentais é um processo lento que exige que todos os interessados - sejam eles bibliotecários, gerentes de repositórios ou gestores de dados - compreendam e disseminem os benefícios da citação de dados para os pesquisadores, especialmente no que tange à descoberta e reutilização de dados, bem como aos devidos créditos para aqueles autores que publicam dados de qualidade (CASTELLI; MANGHI; THANOS, 2013, p. 162, tradução nossa).

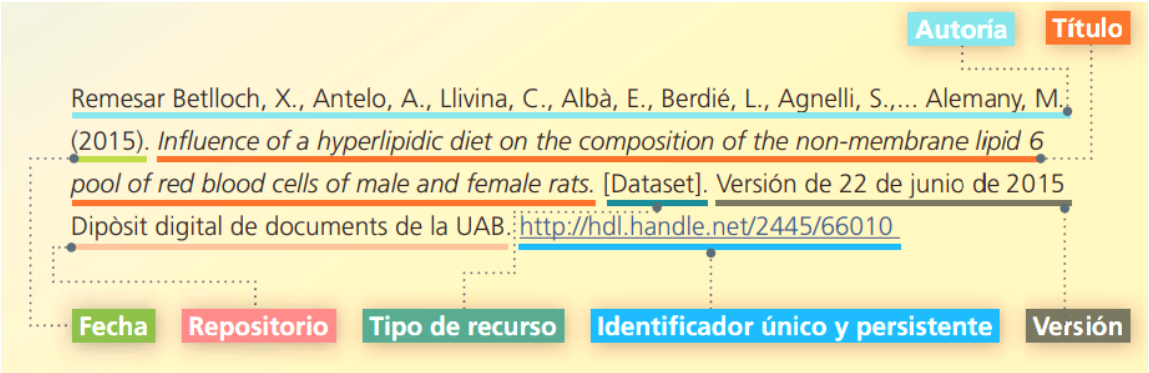
A documentação dos dados de pesquisa corresponde aos metadados necessários para que esses dados sejam descritos apropriadamente, de um modo que possibilite seu gerenciamento ao longo do tempo e sua reutilização futura sem prejuízos informacionais. O padrão de metadados a ser adotado irá depender da área disciplinar na qual a pesquisa é realizada, visto que existem vários esquemas de metadados nas mais diversas disciplinas, sendo que a escolha por um deles dependerá da adequação de seus elementos ao tipo de dado que foi coletado ou gerado.

Os metadados constituem uma parte crucial do tratamento de dados de pesquisa, porquanto a descrição meticulosa desses dados encerra em si uma etapa imprescindível à gestão de dados científicos. As funções e os benefícios dos metadados são múltiplos, podendo-se citar os seguintes: 1) evitar a duplicação de dados; 2) compartilhar as informações de modo estruturado; 3) tornar os dados pesquisáveis, recuperáveis e avaliáveis; 4) ajudar o usuário dos dados a discernir se estes atendem às suas necessidades específicas de informação; 5) permitir a descoberta e o uso de conjuntos de dados de pesquisa (WILEY, 2014). Para dar suporte às exportações desses registros de metadados que se armazenam nos repositórios em formato XML, utiliza-se o protocolo OAI-PMH, o qual começou sendo aplicado em bibliotecas digitais, mas que também se demonstrou de equivalente utilidade nos repositórios de dados de pesquisa (CASTELLI; MANGHI; THANOS, 2013).

Além da documentação dos dados, outro problema a ser discutido é o da citação dos dados em trabalhos científicos. A citação de dados só pode ser viabilizada mediante sua devida descrição por meio de metadados - de modo a permitir que seja elaborada uma referência que possa ser citada, assim como as publicações ordinárias - e sua disponibilização em alguma fonte acessível, como os repositórios de dados.

Entretanto, algumas ações precisam ser tomadas nesse sentido, como associar digitalmente os dados com suas publicações, facilitar o acesso aos dados, aumentar sua aceitação como contribuições tão legítimas e citáveis como os artigos de periódicos e estimular seu arquivamento (de preferência em repositórios acessíveis), facilitando sua descoberta e reutilização (CASTELLI; MANGHI; THANOS, 2013). Promover iniciativas que viabilizem a citação de dados permite que os autores recebam o devido crédito pelo seu trabalho e representa outra etapa essencial na gestão de dados de pesquisa, algo que encontra ratificação na realidade prática, haja vista que muitas universidades que prestam apoio à gestão de dados têm oferecido informações aos pesquisadores acerca da citação de dados. A Figura 4 ilustra um exemplo de como os dados de pesquisa podem ser citados em um trabalho acadêmico, de acordo com as normas da APA.

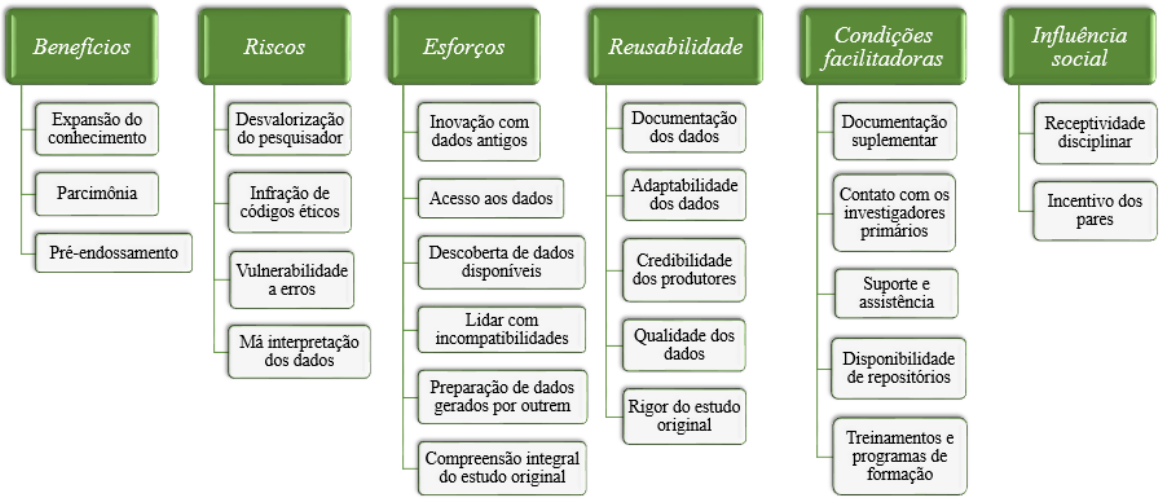
FIGURA 4 – Modelo de citação de conjuntos de dados em trabalhos acadêmicos



Fonte: REBIUN (2016).

Adotar um parâmetro para a citação de conjuntos de dados mostra-se uma estratégia de grande relevância, na medida em que proporciona um meio de atribuir os devidos créditos aos pesquisadores que originalmente os coletou ou produziu, incentivando-se, desse modo, o reconhecimento dos autores. Entretanto, o reuso de dados pode ser uma questão delicada, tendo em vista que pode exigir que os "cientistas sejam capazes de traduzir e recontextualizar os dados primários obtidos por outros pesquisadores, a fim de aplicar para os seus próprios fins, sem que haja má-interpretação ou má-utilização destes" (CURTY, 2016, p. 3300). Muitos são os fatores que influenciam a prática de reutilização dos dados, como se pode constatar nos itens levantados na síntese da Figura 5.

FIGURA 5 – Fatores que exercem influência sobre a prática de reuso de dados



Fonte: Elaboração própria com base em Curty (2016).

Como se pode notar, muitos são os fatores que precisam ser levados em consideração quando se trata de compartilhamento e reutilização de dados de pesquisa. Ainda sobre a questão da citação, é incontestável o fato de que a citação de dados é mais complexa do que a citação de publicações científicas, visto que os conjuntos de dados dificilmente são localizáveis ou revisados pelos pares, estando frequentemente escondidos nos arquivos pessoais dos cientistas (CASTELLI; MANGHI; THANOS, 2013).

Pode-se delimitar pelo menos três macroprocessos que ocorrem na comunicação dos dados de pesquisa, os quais são: produção, compartilhamento e uso (COSTA; LEITE, 2017). Estes dois últimos são objeto de discussões mais longas no âmbito da comunicação científica, algo evidenciado no fato de que "há uma parcela considerável do trabalho científico que não está visível nem para a sociedade em termos de benefícios e qualidade de vida, nem para os pares no contexto da dinâmica de uma comunidade científica" (SALES; SAYÃO, 2018, p. 4183).

Em meio a um volume tão vasto de dados produzidos nas pesquisas científicas, há um segmento da Ciência denominado por Sales e Sayão (2018, p. 4185) como *cauda longa da Ciência*, representada pelos dados científicos produzidos em larga escala e que se encontra dispersa "entre instituições, projetos, laboratórios, pequenos grupos de pesquisa e pesquisadores individuais". Os autores apontam vários motivos que justificam a opacidade dos dados desse segmento da Ciência, dentre os quais podem ser citados: 1) falta de infraestrutura tecnológica e gerencial e de políticas institucionais que assegurem a estabilidade, a persistência e a interoperabilidade dos dados; 2) necessidade de controle de qualidade e a padronização; 3)

falta de políticas voltadas para a publicação de dados; 4) ausência de esquemas de reconhecimento da autoria e de políticas de recompensa pela organização e disseminação dos dados; 5) fatores individuais e comportamentais, como como a falta de interesse dos pesquisadores em divulgar seus dados (SALES; SAYÃO, 2018).

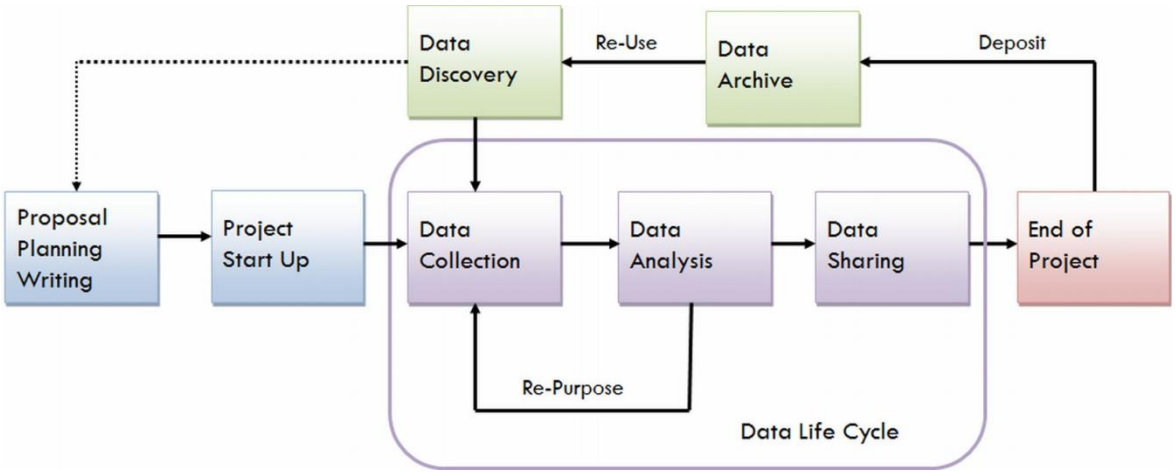
2.5 A gestão de dados de pesquisa e o ciclo de vida dos dados

Gestão de dados de pesquisa (em inglês, *research data management*) pode ser definida como um conjunto de processos que exigem ferramentas e infraestruturas e envolvem serviços que se enquadram nas etapas do ciclo de vida dos dados (DUDZIAK, 2016).

Esses serviços de gestão de dados, geralmente oferecidos por universidades ou entidades associadas aos projetos de pesquisa, começaram a ser oferecidos após as mudanças despertadas pelo quarto paradigma da Ciência, o qual engendrou não apenas uma ciberinfraestrutura capaz de atender às exigências do ciclo de vida dos dados, mas também uma modificação nos costumes das agências de financiamento, que passaram a instituir políticas mandatórias de gestão de dados científicos (CHIWARE; MATHE, 2015; ANJOS; DIAS; RODRIGUES, 2017).

O ciclo de vida dos dados, circuito dentro do qual ocorrem todos os procedimentos relativos à gestão de dados, é um modelo desenvolvido para conferir um caráter sistemático à prática de gerenciamento de dados e a curadoria digital. Nas Figuras 6 e 7 apresentam-se dois modelos de ciclo de vida de dados, ligeiramente diferentes entre si, mas que em essência representam o mesmo conjunto de processos.

FIGURA 6 – Modelo de ciclo de vida dos dados da University of Virginia



Fonte: University of Virginia Library.

FIGURA 7 – Modelo de ciclo de vida dos dados do DataONE

Fonte: Oliveira (2016), adaptado do DataONE (2015).

As oito etapas integrantes do ciclo de vida dos dados formulado pelo DataONE constituem o próprio processo de gerenciamento de dados de pesquisa e apresentam uma série de especificidades que são descritas no Quadro 1.

QUADRO 1 – Etapas do ciclo de vida dos dados

ETAPA	DESCRIÇÃO
Planejar	Relacionado com o plano de pesquisa que será desenvolvido pelo pesquisador. Descreve o mapeamento dos processos metodológicos e técnicos e a ciberinfraestrutura tecnológica necessária para a execução dos estágios subsequentes.
Coletar	Direcionado ao processo de coleta dos dados primários, os procedimentos e processamentos em bancos de dados, *software*, laboratórios ou repositórios. Recomendável que o pesquisador tenha uma infraestrutura preliminar de gestão, com ferramentas de fácil acesso e manejo, a exemplo do *Dropbox* e *Google Drive*. Observar os aspectos relacionados ao volume, extensão e mídia dos dados.
Assegurar	Relacionado à migração dos dados científicos para uma ferramenta compatível que favoreça o compartilhamento, curadoria e preservação a longo prazo em uma única interface. Adotar identificadores para os dados, a exemplo do DOI. Observar procedimentos legais e licenças públicas.
Descrever	Contempla a adoção de um padrão de metadados para viabilizar a descrição minuciosa dos dados de pesquisa.
Preservar	Analisar que perdas de danos são possíveis por diferentes motivos. Inclui *backups* e arquivamento, assim como conversão, reformatação e salvamento de

	dados. Utiliza ferramentas de armazenagem em nuvens, *data centers*, repositórios, entre outros.
Descobrir	Associado à identificação de conjuntos de dados e repositórios que podem agregar valor ao projeto, desde que os dados e metadados sejam detectáveis, reutilizáveis e citáveis.
Integrar	Relacionado a protocolos de interoperabilidade responsáveis pela integração dos metadados com a finalidade de prover visibilidade, novas análises e investigações para o reuso e reprodutibilidade futuras. Adoção do OAI-PMH para possibilitar a interoperabilidade.
Analisar	Relacionado com a análise do projeto de dados por pesquisadores, comunidade científica, parceiros e profissionais, por meio de ferramentas especializadas.

Fonte: Elaboração própria com base em Oliveira (2016).

A gestão de dados científicos apresenta-se como uma realidade da qual não se pode esquivar na era da *e-Science*, já havendo muitas iniciativas no âmbito da União Europeia e dos Estados Unidos no sentido de armazenamento, preservação, compartilhamento e reuso de dados científicos. Contudo, no Brasil os empreendimentos em torno do assunto ainda são incipientes, tendo em vista a inocência de grande parte da comunidade acadêmica a respeito de sua importância, das ferramentas necessárias e da falta de visão de futuro (SAYÃO; SALES, 2013).

Anjos, Dias e Rodrigues (2017), em pesquisa realizada em 2017 com pesquisadores vinculados a programas de pós-graduação em Ciência da Informação de universidades brasileiras, certificaram-se de que cerca de 60% dos pesquisadores realizavam a gestão dos dados de suas pesquisas. O estudo publicou resultados parciais de uma pesquisa que trabalha com um recorte da realidade da gestão de dados científicos na esfera acadêmica brasileira, e procurou investigar também as razões pelas quais esses pesquisadores não realizam a gestão de seus dados. Em linhas gerais, os motivos apontados foram os seguintes:

- Falta de conhecimento acerca da gestão de dados de pesquisa, tendo em vista o caráter recente desta prática no meio científico;
- Desconhecimento de métodos adequados de tratamento e armazenamento;
- Baixa familiaridade com os procedimentos de gestão, faltando conhecer melhor as estratégias necessárias voltadas ao preparo e à disponibilização dos dados de pesquisa em repositórios;
- Ausência de uma política mandatória que torne obrigatória a gestão de dados, fazendo com que se proceda da forma tradicional: guardar todos os dados brutos (incluindo

questionários e transcrições de entrevistas) em pastas físicas e digitais, seja em computadores pessoais ou armazenamento em nuvem, como no *Dropbox*;

- Hábito de apenas concluir a pesquisa, publicar seus resultados e descartar os demais dados;
- Falta de recursos - principalmente tempo - para manter em funcionamento os sistemas que permitem acesso aos dados;
- Crença de que a gestão de dados de pesquisa ainda não se consolidou como uma atividade relacionada ao fazer científico;
- Falta de tradição da prática de gestão de dados na área da Ciência da Informação, inexistindo solicitações para compartilhar os dados das pesquisas;
- E, por último, o desinteresse pessoal, que pode ter causas de naturezas diversas.

Como se pode observar, várias podem ser as razões para não se realizar a gestão apropriada dos dados de pesquisa no contexto acadêmico brasileiro. Gradativamente, as instituições de ensino superior e de pesquisa têm reconhecido a importância de formular estratégias que visem à maximização do valor dos dados gerados nas atividades de pesquisa, inclusive sobre a disponibilização e preservação desses dados, além da inevitabilidade de se adequar aos requisitos que têm sido estabelecidos pelas entidades financiadoras (PRÍNCIPE; SARAIVA, 2015). A Figura 8 ilustra como se dá a dinâmica da gestão de dados num âmbito institucional, especificamente aplicável às universidades, nas quais a prática de gerenciamento de dados é amparada pelas bibliotecas.

FIGURA 8 – Fluxo da gestão de dados de pesquisa em instituições acadêmicas

Fonte: Sayão e Sales (2016a).

Os cientistas sempre diferiram quanto ao modo de organização e recuperação de seus dados de pesquisa, podendo seus sistemas de arquivamento variar do caprichado ao caótico, a

depender do pesquisador (MEADOWS, 1999). O individualismo tornou-se um comportamento ultrapassado na era do quarto paradigma da Ciência, o qual propicia toda uma estrutura tecnológica para o armazenamento seguro e para o compartilhamento de dados científicos. Conforme assinala Corrêa (2016, p. 394), "se os pesquisadores armazenam os dados digitais em servidores ou discos rígidos sem executar regularmente as ações de preservação necessárias, com o tempo seus dados se tornarão inutilizáveis" (CORRÊA, 2016, p. 394).

Muito mais do que apenas uma questão de compartilhamento, é preciso adotar procedimentos, padrões e técnicas internacionais que viabilizem e facilitem o processo de gestão de dados, o que implica uma atenção especial com os metadados, os quais possuem a capacidade de garantir a confiabilidade de dados sensíveis, a citação e os créditos dos autores (OLIVEIRA, 2016). No plano econômico, o reuso de dados de difícil reprodução pode diminuir o custo financeiro das pesquisas, ao passo que no plano acadêmico, a divulgação dos dados primários juntamente com as publicações permite a verificação pelos pares, o que possibilita que seja averiguada a consistência dos dados, assegurando a qualidade da produção científica e encurtando o ciclo clássico de comunicação científica (SAYÃO; SALES, 2013; ROCHA; SALES; SAYÃO, 2017). Reforçando as ideias expostas, convém salientar que

> [...] a crescente importância dos dados de pesquisa, somado às políticas mandatórias e de incentivos ao livre acesso e ao compartilhamento colocadas pelas agências de fomento e pelas instituições de pesquisa, somado às exigências de publicações dos dados por uma parcela considerável de editores de periódicos científicos, impulsiona o surgimento de serviços e sistemas em escala mundial voltados para hospedagem de dados científicos. Diante desse quadro se torna difícil para pesquisadores, agências financiadoras, editores e instituições acadêmicas selecionar repositórios apropriados para armazenar e descobrir dados de pesquisa (SAYÃO; SALES, 2016a, p. 109).

2.6 Plano de gestão de dados: conceito e funções

Um plano de gestão de dados (de modo abreviado, PGD; em inglês, *data management plan*) é um documento formal elaborado pelo pesquisador no início de seu projeto de pesquisa no qual são descritas as diretrizes para o ciclo de vida dos dados, descrevendo-se como os dados serão tratados durante e após a pesquisa (SAYÃO; SALES, 2015; MONTEIRO; SANT'ANA, 2018; BIBLIOTECA CEPAL, 2019).

> O PGD auxilia tanto os pesquisadores que coletam e manipulam conjuntos de dados quanto àqueles profissionais que atuam nos repositórios de dados científicos e fornece diretrizes para todo o ciclo de vida dos dados, com destaque para a indicação dos tipos e formatos dos dados, os métodos de compartilhamento de dados e as políticas para reutilização e redistribuição de dados (MONTEIRO; SANT'ANA, 2018, p. 163-164).

Frequentemente exigidos por agências de fomento como instrumentos mandatórios, a exemplo do europeu H2020 e da norte-americana NSF, os planos de gestão de dados exigem esforços tanto dos pesquisadores quanto das instituições às quais estes se encontram associados, no que tange ao gerenciamento, compartilhamento e preservação dos dados de pesquisa, incluindo o depósito desses dados em repositórios confiáveis (COX; PINFIELD; SMITH, 2016; OLIVEIRA, 2016; MORENO, 2018).

Em pesquisa realizada em 2016, Monteiro e Sant'Ana (2018) verificaram que, entre as 100 melhores universidades do mundo – ranqueadas pelo *webometrics.info*[1] –, 55 dispunham de repositórios de dados científicos e apenas 36 desses possuíam políticas de gestão de dados próprias, o que evidencia que mesmo as instituições mais prestigiadas do mundo acadêmico ainda estão iniciando em processo de implantação de repositórios e, portanto, ainda iniciando suas ações em prol do gerenciamento de dados.

Um dos principais aspectos abordados nos planos de gestão de dados é a documentação e a adoção de algum esquema de metadados. A eminência dos metadados justifica-se pelo fato de que estes permitem buscas estruturadas e consistentes em repositórios, propiciando a descoberta, uma das etapas do ciclo de vida dos dados (SAYÃO; SALES, 2015). Portanto, é dedutível que

> Uma documentação exaustiva dos dados é a chave para a compreensão do significado deles agora e no futuro. Sem uma descrição minuciosa do contexto tecnológico dos arquivos de dados, do contexto no qual os dados foram criados ou coletados, das medidas que foram feitas, dos detalhes espaciais e temporais, dos instrumentos usados, dos parâmetros e unidades e da qualidade dos dados e da sua proveniência, é improvável que os dados possam ser descobertos, interpretados, gerenciados e efetivamente usados e reusados. Os metadados cumprem essa tarefa, porque eles são a documentação dos dados (SAYÃO; SALES, 2015, p. 20).

Em suma, um PGD encarrega-se de descrever os métodos de coleta de dados e formatos de arquivo a serem adotados, além de detalhar informações sobre o armazenamento, compartilhamento e descarte, direitos de propriedade intelectual, implicações éticas e ferramentas a serem utilizadas (MONTEIRO; SANT'ANA, 2018).

---

[1] Ranking acadêmico das instituições de ensino superior do mundo. Realizado pelo Cybermetrics Lab (Conselho Nacional de Pesquisa da Espanha, CSIC), fornece informações confiáveis, multidimensionais, atualizadas e úteis sobre o desempenho das universidades de todo o mundo com base em sua presença e impacto na *web*.
Disponível em: http://webometrics.info/en/Methodology. Acesso em: 19 jul. 2019.

2.7 Bibliotecas universitárias e serviços relacionados à gestão de dados

Em ação conjunta com os pesquisadores e as agências de fomento, muitas universidades têm prestado serviços de apoio à gestão de dados científicos. Como exemplos de serviços, Tenopir *et al.* (2017, p. 27) listam os seguintes:

1) Criação e gerenciamento de repositórios de dados institucionais;
2) Fornecimento de ferramentas para mineração e visualização de dados;
3) Treinamento para pesquisadores sobre atividades de gerenciamento de dados;
4) Orientação sobre políticas institucionais;
5) Auxílio na criação de PGDs e metadados;
6) Assistência em questões de propriedade intelectual e privacidade.

Como se pode notar, vários são os tipos de auxílio que as bibliotecas universitárias podem oferecer aos pesquisadores para prestar suporte às fases do ciclo de vida dos dados. Idealmente, esses serviços "devem ser prestados em estreita colaboração com pesquisadores e podem incluir o desenvolvimento de planos de gestão para documentar e organizar os dados através do desenvolvimento de ferramentas ou recursos para armazenar dados de forma segura" (CORRÊA, 2016, p. 388).

Um exemplo brasileiro de serviço de apoio à gestão de dados de pesquisa é o Centro de Documentação e Acervo Digital da Pesquisa, órgão auxiliar da Faculdade de Biblioteconomia e Comunicação da UFRGS, cujo objetivo é "dar suporte à pesquisa científica e tecnológica através da digitalização, gestão e curadoria de ativos digitais de pesquisa, assim como avançar o estado da arte em digitalização e curadoria de ativos digitais" (ROCHA; CAREGNATO; GABRIEL JUNIOR, 2018, p. 3).

A respeito do programa *Horizon 2020*, os pesquisadores participantes são solicitados a depositar seus dados juntamente com seus metadados associados e, caso aplicável, indicar também as ferramentas de software necessárias para replicar e validar os dados, além de ser recomendado o uso de repositórios de dados confiáveis para armazenar e divulgar esses dados (VERHAAR *et al.*, 2017). Nota-se, portanto, um enaltecimento dos repositórios de dados como uma das ferramentas primordiais no sucesso da gestão de dados científicos.

Outro aspecto importante é o gerenciamento de dados de acordo com os princípios FAIR, cujas recomendações enfatizam questões relativas a metadados e protocolos, promovendo, por meio da disseminação desses princípios, a filosofia e as práticas da Ciência Aberta (HENNING *et al.*, 2019).

2.8 Repositórios de dados científicos

Repositórios de dados podem ser definidos como serviços digitais que podem ser classificados como institucionais, temáticos, disciplinares ou científicos, sendo responsáveis por proporcionar o agrupamento, a descrição, o acesso contínuo e a preservação a longo prazo de conjuntos de dados ou publicações (SAYÃO; SALES, 2016a; CURTY; AVENTURIER, 2017).

Os repositórios de dados científicos tiveram sua gênese com a emergência da necessidade de gestão de dados, estando desde então vinculados a universidades e instituições de pesquisa, viabilizando a disponibilização dos dados para a comunidade científica ou para o público em geral com o mínimo possível de restrições (MONTEIRO; SANT'ANA, 2018).

> [Os repositórios de dados] tornam uma parte importante da atividade científica – antes oculta e sem lugar apropriado – visível e aberta para toda a sociedade; apoiam a validação e a revisão de publicações científicas; se tornam também parte da memória digital mais fidedigna da ciência, posto que podem registrar os percursos de erros e acertos que fazem parte do ciclo de geração de conhecimento científico; contribuem ainda para que os pesquisadores que coletam, geram e organizam dados possam ser identificados, reconhecidos e citados pelo trabalho que fica oculto no contexto de uma ciência voltada para o resultado final - a descoberta e a publicação (SAYÃO; SALES, 2016a, p. 111).

Como administradoras dos repositórios de dados, as instituições mantenedoras dessas coleções assumem a função de aumentar a visibilidade de seus pesquisadores e da própria instituição, na medida em que promovem o acesso a esses dados e possibilita que estes sejam descobertos, avaliados e citados, contribuindo, dessa forma, para o aumento do reconhecimento internacional da produção científica dos países nos quais estão localizadas (PAVÃO *et al.*, 2012).

Como exemplo da importância da visibilização dos repositórios de dados científicos na atual era da Ciência, criou-se o *re3data.org*, um diretório que reúne mais de 2.300 repositórios[2] de dados das mais variadas áreas do conhecimento e dos mais diversos países (CURTY; AVENTURIER, 2017; MORENO, 2018). Dentre suas curiosidades, destaca-se o esquema de ícones que identificam cada um dos repositórios em termos de acesso, licenças, políticas e padrões, informações que saltam à vista de maneira simples e rápida (SAYÃO; SALES, 2016a). Os benefícios dos repositórios de dados são numerosos, como se pode contemplar no Quadro 2.

[2] Último levantamento realizado em setembro de 2019.

QUADRO 2 – Benefícios dos repositórios de dados de pesquisa

ASPECTO	BENEFÍCIOS
Visibilidade dos dados	Permite que os dados sejam consultados e citados com mais frequência.
Compartilhamento de dados	Capacidade de agregação e organização de recursos informacionais dispersos no tempo e no espaço, facilitando sua divulgação.
Crédito ao autor dos dados	Permite que os autores dos dados sejam reconhecidos, citados, avaliados e recompensados pelo trabalho de coleta, geração e organização dos dados.
Preservação digital	Oferece um ambiente tecnológico, gerencial e de padronização propício para a preservação de longo prazo dos dados de pesquisa de valor contínuo.
Memória científica e transparência	Contribui para a formação da memória científica das instituições, complementando os repositórios institucionais focados em publicações.
Segurança dos dados	Oferece sistema de armazenamento seguro, esquemas de *backup* e segurança física que se contrapõem ao armazenamento informal em mídias portáteis e computadores pessoais.
Disponibilidade	Permite que os dados estejam disponíveis *on-line* para serem acessados, baixados, visualizados e processados.
Curadoria digital	Proporciona um ambiente apropriado para os processos de avaliação, de adição de valor, reformatação, agregação e recriação de dados.
Serviços inovadores	Abre possibilidades de criação de novos serviços de informação para pesquisadores, gestores e financiadores de pesquisa a partir da análise e integração dos dados arquivados com fontes internas e externas à instituição.
Reuso dos dados	Possibilita a realização de novas pesquisas de caráter interdisciplinar, minimiza a duplicação de esforços e otimiza os investimentos.
Redes de repositórios	Permite, por meio de protocolos de interoperabilidade, como o OAI-PMH, a formação de redes de repositórios de dados.
Indicador de qualidade e produtividade da instituição	Evidencia a qualidade e relevância das atividades de pesquisa por meio da organização e do arquivamento das coleções de dados.

Fonte: Elaboração própria com base em Sayão e Sales (2016).

Tendo em vista essas vantagens proporcionadas pelos repositórios, torna-se evidente que, além de oferecer uma infraestrutura tecnológica para a consecução de etapas vitais do ciclo de vida dos dados, os repositórios de dados científicos providenciam o acesso, mineração, exploração, reprodução, compartilhamento e citação dos dados, além da própria valida-

ção dos resultados das pesquisas e verificação de sua consistência, promovendo a transparência científica (SAYÃO; SALES, 2016a; COSTA; LEITE, 2017).

2.9 A atuação do bibliotecário na gestão de dados de pesquisa

Considerando a discussão exposta nas seções anteriores, não é difícil dar-se conta da imensa contribuição que pode ser prestada pelos profissionais da informação no processo de gestão de dados científicos e no ato de torná-los abertos para recuperação e reuso, seja por meio da catalogação apropriada dos dados, seja por meio de ações de preservação digital (BERTIN *et al.*, 2017; ROCHA; SALES; SAYÃO, 2017).

O Quadro 3 apresenta brevemente apenas algumas das formas pelas quais os bibliotecários podem auxiliar os pesquisadores no que concerne à gestão de dados de pesquisa.

QUADRO 3 – Funções do bibliotecário no auxílio à comunidade científica

ETAPA	FUNÇÃO DO BIBLIOTECÁRIO
Obtenção de dados	● Auxiliar o desenvolvimento de um plano de gestão de dados e estratégia de metadados; ● Identificar o repositório adequado para as necessidades dos investigadores.
Preservação de dados	● Auxiliar a migração para formatos adequados de preservação, incluindo a criação de *backups* e metadados adicionais; ● Assegurar que o dados permaneçam disponíveis pelo tempo que for necessário.

Fonte: Elaboração própria com base em Corrêa (2016).

De acordo com Bertin *et al.* (2017, p. 5), no que tange à abertura dos dados científicos, "o desenvolvimento de competências ou a contratação de perfis profissionais específicos como os do 'bibliotecário de dados' e 'engenheiro de dados' mostra-se essencial". Na mesma linha de pensamento, Sayão e Sales (2015) ressaltam que os bibliotecários são capacitados para trabalhar na gestão de dados de pesquisa em virtude de seus conhecimentos em gestão de informação, metadados, descoberta de recursos e preservação digital.

Entretanto, tal pressuposto choca com o ponto de vista de Chiware e Mathe (2015), os quais acreditam que a falta de compreensão da diversidade dos dados e a insuficiência das habilidades dos bibliotecários constitui um dos maiores desafios das bibliotecas universitárias no oferecimento de serviços de apoio à gestão de dados.

Para Dudziak (2016), representam as principais competências do bibliotecário gestor de dados de pesquisa o fornecimento do acesso a dados, a defesa e o apoio à gestão de dados, e a gestão de coleções de dados. A autora assinala que “o papel das bibliotecas e dos bibliotecários envolve a identificação e localização de dados, apoio às condições de acesso e reutilização de dados, suporte à citação e referenciação, até a correta organização e preservação” (DUDZIAK, 2016).

Nesse sentido, Sales e Sayão (2018, p. 4195) defendem a necessidade de profissionalização da gestão de dados de pesquisa, a qual abrangeria ações rigorosas para assegurar o acesso a longo prazo aos dados de pesquisa “de valor constante, com graus aceitáveis de autenticidade e integridade e que privilegiem a proveniência - valor informacional que possibilita a reconstrução da geração dos dados”.

A disposição para disponibilizar os dados de pesquisa pode variar conforme o campo do conhecimento. Constatou-se, em pesquisa realizada na University of Colorado em 2011, que pesquisadores da área de Ciências da Terra tendem a ser mais abertos à disponibilização, ao passo que aqueles da área de Ciências Exatas, a qual se caracteriza pelo alto nível de competitividade, normalmente são mais fechados em relação à gestão de dados, especialmente no tocante ao compartilhamento (CORRÊA, 2016).

A despeito das barreiras encontradas, Corrêa (2016, p. 388) frisa que “a assistência do bibliotecário aos pesquisadores ajuda a avaliar o que realmente necessitam entender, a compreenderem como organizar uma variedade de tipos de dados e tomar as decisões corretas sobre o acesso e preservação de dados para os seus projetos”.

3 PROCEDIMENTOS METODOLÓGICOS

Esta seção descreve a metodologia empregada na consecução do presente estudo, desde suas características mais básicas até os procedimentos realizados. Entende-se por metodologia o passo a passo da execução da pesquisa, o que diz respeito aos métodos, técnicas e criatividade de quem está pesquisando, o que significa dizer que a metodologia abarca as concepções teóricas, a realidade empírica e a interpretação do investigador (MINAYO, 2009).

3.1 Caracterização da pesquisa

Para uma compreensão mais lúcida do enquadramento metodológico deste estudo, pode-se classificá-lo quanto à sua natureza, seu objetivo e sua fonte de coleta de dados.

Quanto à sua *natureza*, esta pesquisa pode ser tida como *qualitativa*, caracterizada pela observação e análise de um dado fenômeno sob a ótica subjetiva do pesquisador (GIBBS, 2009; APPOLINÁRIO, 2012). No caso em questão, o fenômeno estudado refere-se à estruturação de planos de gestão de dados de pesquisa.

A complexidade da pesquisa qualitativa reside na "proximidade entre observador e objeto observado, [...] normalmente guiada seguindo dois caminhos: o da *simplificação do objeto* e o da *redução da extensão do domínio observado*" (CARDANO, 2017, p. 24-25, grifo do autor). Neste trabalho, o objeto de estudo são os dados de pesquisa, e sua simplificação consiste na escolha de um de seus aspectos para ser alvo de escrutínio, que no caso são os planos de gestão de dados. A redução da extensão do domínio corresponde à escolha de documentos disponibilizados apenas no âmbito do *Digital Curation Centre*, excluindo-se itens pertencentes a outras fontes.

Segundo Cardano (2017, p. 25), a pesquisa qualitativa normalmente ocorre a partir da "focalização sobre poucos casos, dos quais se propõe a individuar e representar os mínimos detalhes". De acordo com o autor, as técnicas de pesquisa qualitativa priorizam o aprofundamento dos detalhes do objeto estudado e a constituição de amostras pequenas. No contexto das Ciências Sociais Aplicadas, a estreita proximidade entre observador e objeto permite ao primeiro "evidenciar os traços, as peculiaridades, as diferenças que separam os diversos objetos em estudo muito mais facilmente" (CARDANO, 2017, p. 24).

Quanto ao *objetivo* a que se propõe, a presente pesquisa pode ser designada como *descritiva*, visto que nesse tipo de estudo o autor escolhe observar, registrar, narrar e descrever um fato ou fenômeno sem interferir na realidade estudada, mantendo uma postura de especta-

dor analítico (CERVO; BERVIAN; SILVA, 2007; APPOLINÁRIO, 2012). Neste trabalho, busca-se descrever a forma como os planos de gestão de dados têm sido elaborados e estruturados, atentando-se aos tipos de informações levantadas e ao nível de exaustividade com que estas são detalhadas.

Por fim, em relação à *fonte* de coleta de dados, esta pesquisa pode ser classificada como *documental*, na qual "a fonte de informação primordial é constituída por documentos, sejam eles em fontes impressas, sejam eletrônicas" (APPOLINÁRIO, 2012, p. 66). Aqui foram investigados documentos (planos de gestão de dados) em formato de texto, havendo o propósito comum às pesquisas documentais, que é descrever e comparar costumes, tendências, diferenças e outras características (CERVO; BERVIAN; SILVA, 2007).

Convém salientar que o tipo de dados estudados neste trabalho pode ser classificado, de acordo com Cardano (2017), como *dados naturais*, cuja disponibilidade é pré-existente, não dependendo da intervenção do pesquisador para serem gerados. Tal fato condiz com a premissa do pesquisador como um instrumento fundamental na pesquisa qualitativa e documental, na qual pode coletar dados individualmente por meio do exame de documentos (CRESWELL, 2010).

No Quadro 4, sintetiza-se em duas colunas a classificação deste trabalho quanto aos três aspectos discutidos nesta seção.

QUADRO 4 – Caracterização da pesquisa

ASPECTO	TIPO DE PESQUISA
Quanto à *natureza*	Pesquisa qualitativa
Quanto ao *objetivo*	Pesquisa descritiva
Quanto à *fonte de coleta de dados*	Pesquisa documental

Fonte: Elaboração própria.

3.2 Fontes de informação

A priori, realizou-se uma pesquisa bibliográfica exaustiva nas principais bases de dados nacionais e internacionais especializadas em Ciência da Informação, como a BRAPCI e a *Library Literature & Information Science Full Text*.

Como expressões de busca, empregaram-se as seguintes no campo de pesquisa geral: '*research data*', '*research data management*', '*data management plan*', '*scientific communi-*

cation', '*open data*', '*data lifecycle*', 'dados de pesquisa', 'gestão de dados de pesquisa', 'plano de gestão de dados', 'comunicação científica', 'dados abertos' e 'ciclo de vida dos dados'. Em virtude do caráter relativamente recente da eclosão da temática no âmbito da pesquisa acadêmica, não se mostrou necessário aplicar um filtro de data para a seleção dos artigos, visto que grande parte da literatura começou a ser publicada a partir de 2010.

De acordo com Cervo, Bervian e Silva (2007, p. 60), "a pesquisa bibliográfica procura explicar um problema a partir de referências teóricas publicadas em artigos, livros, dissertações e teses". Portanto, para a consecução deste estudo foi a realizada a leitura de artigos de periódicos (em língua portuguesa, inglesa e espanhola), livros indicados nas referências bibliográficas dos artigos lidos, teses e dissertações sobre Ciência Aberta defendidas na década de 2010, além de trabalhos publicados em eventos.

Além das bases de dados supracitadas, realizou-se também uma minuciosa análise dos trabalhos publicados nos últimos dez anos nos eventos do ENANCIB, especificamente aqueles enquadrados no escopo dos seguintes grupos de trabalho: Organização e Representação do Conhecimento (GT-2) e Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação (GT-7), nos quais foi possível recuperar contribuições relevantes a respeito da temática de gestão de dados de pesquisa no âmbito da literatura científica brasileira.

Complementarmente, obtiveram-se informações pertinentes também nos *websites* de bibliotecas de universidades espanholas que prestam apoio significativo à Ciência Aberta e à gestão de dados, como a Universidad de Alcalá e a Universidad de Cantabria.

3.3 Ambiente de pesquisa

O universo da presente pesquisa é composto pelos planos de gestão de dados elaborados por pesquisadores no geral. Em vista desta enorme abrangência, selecionou-se como amostra um conjunto de cinco planos de gestão de dados, cada um pertencente a uma entidade financiadora diferente listada no *website* do *Digital Curation Centre*, escolhido por ser considerado um centro de referência mundial em preservação digital e curadoria de dados de pesquisa. As entidades escolhidas foram as seguintes:

- *Arts and Humanities Research Council* (AHRC);
- *Biotechnology and Biological Sciences Research Council* (BBSRC);
- *Engineering and Physical Sciences Research Council* (EPSRC);
- *Economic and Social Research Council* (ESRC);

- *Horizon 2020* (H2020);

Todas as entidades mencionadas acima são europeias, sendo que os quatro primeiros conselhos mencionados são excepcionalmente britânicos, ao passo que o H2020 abrange toda a União Europeia. As entidades financiadoras foram escolhidas em virtude tanto de seu *status* científico quanto da iniciativa comum de disponibilizar em acesso público PGDs de seus pesquisadores. Optou-se por trabalhar com uma amostra estrangeira devido à falta de sistematização de planos de gestão de dados no contexto brasileiro em plataformas acessíveis.

No Quadro 5, apresentam-se resumidamente os procedimentos metodológicos realizados neste estudo, estabelecendo-se uma correlação com os objetivos específicos delimitados no início deste trabalho.

QUADRO 5 – Procedimentos metodológicos

OBJETIVOS ESPECÍFICOS	AMOSTRA	FONTE DE INFORMAÇÃO	TÉCNICA DE COLETA
Analisar planos de gestão de dados indicados pelo DCC, identificando os tipos de informação considerados relevantes.	PGDs disponibilizados por agências de financiamento europeias.	*Digital Curation Centre*.	Pesquisa documental (análise descritiva)
Comparar os planos de gestão de dados analisados, observando a consonância entre os documentos examinados e os quesitos elencados pelo DCC.	PGDs disponibilizados por agências de financiamento europeias.	*Digital Curation Centre*.	Pesquisa documental (análise comparativa)

Fonte: Elaboração própria.

3.4 *Checklist* para análise dos planos de gestão de dados

Como ponto de partida para a compreensão da forma como um PGD deve ser estruturado e elaborado, além de parâmetro de análise da amostra de PGDs selecionada nesta pesquisa, adotou-se o *checklist* criado e atualizado regularmente pelo DCC (2013) para auxiliar pesquisadores, entidades de financiamento à pesquisa e instituições na gestão de dados. O documento destaca oito categorias de informações que devem constar em um PGD que vise ser "completo". Essas categorias são referentes aos seguintes aspectos:

1. Dados Administrativos *(Administrative Data)*;
2. Coleta de Dados *(Data Collection)*;

3. Documentação e Metadados *(Documentation and Metadata)*;
4. Ética e Direitos Autorais *(Ethics and Legal Compliance)*;
5. Armazenamento e *Backup (Storage and Backup)*;
6. Seleção e Preservação *(Selection and Preservation)*;
7. Compartilhamento de Dados *(Data Sharing)*;
8. Responsabilidades e Recursos *(Responsibilities and Resources)*.

No Quadro 6, descrevem-se as questões que idealmente devem ser respondidas na elaboração de um PGD, estratificando-se cada quesito de acordo com as categorias acima listadas, no intuito de destacar os grandes aspectos envolvidos na gestão de dados de pesquisa.

QUADRO 6 – *Checklist* de elaboração de um PGD

ITEM	ASPECTOS E QUESTÕES A SEREM CONSIDERADOS
Dados Administrativos	
ID	*ResearcherID*[3] do pesquisador atribuído pela entidade ou instituição.
Entidade Financiadora	Nome da agência que financia a pesquisa.
Número Concedido	Identificação atribuída pela entidade financiadora, quando aplicável.
Nome do Projeto	Título atribuído pelo(s) pesquisador(es) ao projeto de pesquisa.
Descrição do Projeto	- Qual a natureza do projeto de pesquisa? - Quais problemas de pesquisa estão sendo abordados? - Para quais propósitos os dados serão coletados ou criados?
Pesquisador Principal	Nome do(s) principal(is) pesquisador(es) do projeto.
ORCID[4]	Identificador digital do(s) pesquisador(es) principal(is) do projeto.
Contatos do Projeto	Telefone e e-mail dos pesquisadores.

[3] *ResearcherID* é um identificador único que permite criar um perfil *online* para mostrar o histórico de publicações de um pesquisador, contribuindo para solucionar o problema da ambiguidade do nome do autor no meio acadêmico, evitando sua identificação errônea, e permitindo que o pesquisador gerencie sua lista de publicações, rastreie suas contagens de citações e identifique outros pesquisadores que possuem linhas de pesquisa similares. Disponível em: http://www.sibi.usp.br/apoio-pesquisador/identificacao-pesquisadores/researcher-id/. Acesso em: 2 ago. 2019.
https://www.pucpr.br/wp-content/uploads/2017/01/tutorial-researcherid-orcid.pdf. Acesso em: 2 ago. 2019.

[4] ORCID "é um identificador digital único, gratuito e persistente, que distingue um acadêmico/pesquisador de outro e resolve o problema da ambiguidade e semelhança de nomes de autores e indivíduos, substituindo as variações de nome por um único código numérico. [...] Dessa forma, facilita o registro de informações e automatiza a atualização das publicações e produções". Disponível em: http://www.sibi.usp.br/apoio-pesquisador/identificacao-pesquisadores/orcid-2/orcid-caracteristicas/. Acesso em: 2 ago. 2019.

Data da Prim. Versão	Data da primeira versão do PGD.
Data da Última Versão	Data da última vez em que o PGD foi modificado.
Políticas Relacionadas	- Existem procedimentos existentes para se basear? - A instituição possui diretrizes de gestão de dados? - A instituição tem uma política de proteção de dados a ser seguida? - A instituição ou entidade financiadora possui uma política de gestão de dados de pesquisa? - Existem padrões formais que serão adotados?
Coleta de Dados	
Tipo de dados	- Qual o tipo, formato e volume dos dados? - Os formatos e *softwares* escolhidos permitem o compartilhamento e acesso dos dados a longo prazo? - Há dados pré-existentes que podem ser reutilizados?
Forma de coleta	- Quais padrões ou metodologias serão usados? - Como serão estruturadas e denominadas as pastas e os arquivos? - Como se lidará com a atualização dos *softwares* escolhidos? - Quais processos de garantia de qualidade serão adotados?
Documentação e Metadados	
Tipo de documentação	- Quais informações são necessárias para que os dados sejam inteligíveis e interpretados no futuro? - Como será a captura ou criação da documentação e dos metadados? - Qual padrão de metadados será utilizado e por quê?
Ética e Direitos Autorais	
Questões Éticas	- Obteve-se permissão para preservar e compartilhar dos dados? - Como será protegida a identidade dos participantes da pesquisa (quando aplicável)? - Como os dados confidenciais serão manipulados para garantir que sejam armazenados e transferidos com segurança?
Direitos Intelectuais	- Quem é o proprietário dos dados? - Como os dados serão licenciados para permitir sua reutilização? - Existem restrições quanto à reutilização de dados por terceiros? - O compartilhamento de dados será postergado em virtude de possível patenteamento?
***Armazenamento e* Backup**	
Modo de Guarda	- Há armazenamento suficiente ou será necessário incluir taxas por serviços adicionais? - Como será feito o *backup* dos dados? - Quem será responsável pelo *backup* e recuperação? - Como os dados serão recuperados em caso de um incidente?
Acesso e Segurança	- Quais são os riscos à segurança de dados e como eles serão contornados? - Como será controlado o acesso para manter os dados seguros? - Como se garantirá o acesso dos colaboradores com segurança?

	- Como se garantirá a transferência de dados de campo (se for o caso) para um sistema seguro?
Seleção e Preservação	
Retenção de Dados	- Quais dados devem ser retidos ou destruídos para fins contratuais, legais ou regulatórios? - Como será a decisão sobre quais dados serão mantidos? - Quais são os possíveis usos dos dados para fins de pesquisa? - Por quanto tempo os dados serão retidos e preservados?
Plano de Preservação	- Em que repositório ou arquivo os dados serão guardados? - Haverá custos para arquivamento no repositório? - Haverá custos de tempo e esforço na preparação dos dados para compartilhamento?
Compartilhamento de Dados	
Modo de Compartilhamento	- Como os potenciais usuários poderão descobrir os dados? - Com quem os dados serão compartilhados e em que condições? - Os dados serão compartilhados por meio de repositórios ou outros mecanismos? - Será atribuído um identificador persistente[5] para o conjunto de dados?
Restrições	- Qual medida será tomada para superar ou minimizar as restrições? - Por quanto tempo será preciso o uso exclusivo dos dados e por quê? - Será necessário algum contrato para o compartilhamento dos dados?
Responsabilidades e Recursos	
Responsáveis	- Quem será responsável pela implementação do PGD e por assegurar que este seja revisado e corrigido? - Quem será responsável por cada atividade da gestão dos dados? - Como as responsabilidades serão divididas entre projetos de pesquisa colaboradores? - A posse dos dados e a responsabilidade pelo seu gerenciamento constarão em algum contrato ou outro acordo?
Recursos Adicionais	- Será necessário algum treinamento de pessoal? - Será necessário um *hardware* ou *software* excepcional a ser utilizado pelo repositório? - Serão cobradas taxas por parte do repositório?

Fonte: DCC (2013), tradução livre.

O *checklist* acima, traduzido e adaptado do DCC, pode ser tomado como referência para a elaboração de PGDs, por parte de agências financiadoras, instituições acadêmicas e

[5] *Identificador persistente* pode ser definido como um código alfanumérico que proporciona uma identificação unívoca a um recurso digital, que permanece o mesmo para sempre, independentemente de sua localização. "O uso de um identificador persistente assegura que, mesmo quando um documento é movido, ou sua propriedade é transferida, os *links* para ele permaneçam efetivamente acionáveis" (SAYÃO, 2007, p. 68).

repositórios de dados, no que tange à formulação de políticas[6] de gestão de dados ou *templates* para construção de PGDs.

No capítulo seguinte serão analisados os PGDs selecionados como amostra a partir dos documentos disponibilizados pelo DCC, utilizando-se como critérios de julgamento os requisitos elencados no Quadro 6.

[6] Cabe estabelecer, neste estudo, a diferença entre plano de gestão de dados e políticas de gestão de dados. O primeiro refere-se ao documento elaborado pelo(s) pesquisador(es) no âmbito de seu próprio projeto de pesquisa, ao passo que o segundo diz respeito às diretrizes estabelecidas por entidades ou repositórios para orientar passo a passo a gestão dos dados.

4 ANÁLISE DOS PLANOS DE GESTÃO DE DADOS

Conforme assinala Gibbs (2009, p. 9), “uma parte importante da pesquisa qualitativa está baseada em texto e na escrita, desde notas de campo e transcrições até descrições e interpretações, e, finalmente, à interpretação dos resultados e da pesquisa como um todo”. Sendo assim, este capítulo dedica-se à apresentação dos documentos selecionados para estudo, acompanhada de sua análise descritiva e interpretativa.

4.1 Análise dos documentos

Os documentos escolhidos para serem examinados e discutidos neste estudo foram selecionados por intencionalidade, visto que em uma pesquisa de natureza qualitativa a seleção da amostra ocorre de acordo com o que o investigador julga que irá contribuir mais satisfatoriamente para a compreensão do problema de pesquisa (CRESWELL, 2010).

A análise dos PGDs foi estruturada utilizando-se como critério de divisão a entidade financiadora, no intuito de realizar em um primeiro momento uma análise focalizada para posteriormente proceder com uma análise mais holística, considerando todos os PGDs abordados.

4.1.1 PGD do Arts and Humanities Research Council

O *Arts and Humanities Research Council* é uma entidade que promove pesquisas nos campos das Ciências Humanas e das Artes, priorizando pesquisas relacionadas à humanidade, cidadania, filosofia, cultura, religião, direitos humanos e criatividade artística, aplicados ao desenvolvimento social, humanitário e econômico (ARTS AND HUMANITIES RESEARCH COUNCIL, 2013).

O PGD do AHRC escolhido para análise pertence ao projeto intitulado “*Virtual Holocaust Memory: from Testimony to Holography*” e sua versão completa encontra-se documentada no Anexo A. O plano está subdividido em 8 itens, contando seções e subseções, sendo estruturado da seguinte forma:

- *Section 1. Summary of Digital Outputs and Digital Technologies*
- *Section 2.a. Technical Methodology: Standards and Formats*
- *Section 2.b. Technical Methodology: Hardware and Software*

- *Section 2.c. Technical Methodology: Data Acquisition, Processing, Analysis and Use*
- *Section 3. Technical Support and Relevant Experience*
- *Section 4: Preservation, Sustainability and Use*
- *Section 4.a. Preserving Your Data*
- *Section 4.b. Ensuring Continued Accessibility and Use of Your Digital Outputs*

A fim de detalhar as informações contidas neste PGD, será utilizado o *checklist* já apresentado no Quadro 6, no intuito de verificar a presença ou a ausência dos pré-requisitos apontados pelo DCC. No Quadro 7, mostra-se de forma sistematizada a correspondência entre os itens levantados pelo DCC e as informações constantes no PGD em questão.

QUADRO 7 – *Checklist* do PGD do AHRC

ITEM	P/A*	ITEM	P/A*
ID	✗	Forma de coleta	✓
Entidade Financiadora	✓	Tipo de documentação	✓
Número Concedido	✗	Questões Éticas	✓
Nome do Projeto	✓	Direitos Intelectuais	✗
Descrição do Projeto	✗	Modo de Guarda	✓
Pesquisador Principal	✗	Acesso e Segurança	✓
ORCID	✗	Retenção de Dados	✗
Contatos do Projeto	✗	Plano de Preservação	✓
Data da Primeira Versão	✗	Modo de Compartilhamento	✗
Data da Última Versão	✗	Restrições	✗
Políticas Relacionadas	✗	Responsáveis	✓
Tipo de dados	✓	Recursos Adicionais	✗

Legendas: **P** - Informação presente ✓ | **A** - Informação ausente ✗.
Fonte: Elaboração própria.

- *Dados Administrativos:* a primeira constatação a ser feita a partir da análise deste PGD é que não constam no documento as informações referentes à categoria de *Dados Administrativos*, o que pode ser justificado pela convenção da anonimização dos autores do projeto, visto que o

PGD foi divulgado pela entidade financiadora apenas como um exemplo, o que pode dispensar a necessidade de veicular o documento com os dados pessoais do(s) pesquisador(es) envolvido(s).

- *Coleta de Dados:* a respeito do *tipo, formato e volume dos dados,* faz-se menção no PGD a cada um desses quesitos, visto que estes constituem informações essenciais a serem abordadas. A *forma de coleta* de dados mencionada é por meio de gravação de videoconferências de seminários nas universidades de Leeds, Sheffield e York em 2015 e 2016, além de eventos no *Johannesburg Holocaust and Genocide Centre* entre abril e maio de 2016. As informações sobre os dados a serem coletados são as seguintes:

- *Tipo:* vídeo;
- *Formato:* MP4
- *Volume:* 15 vídeos, cada um tendo entre 5 e 30 minutos;
- *Tamanho:* cada vídeo terá entre 50 e 500 MB, sendo estimado em aproximadamente 3.75 GB o tamanho da soma de todos os vídeos.

- *Documentação e Metadados:* as informações a respeito dos metadados a serem adotados para a descrição dos vídeos é um tanto vaga e imprecisa, sendo mencionado apenas que os arquivos de vídeo serão indexados com "metadados relevantes" tanto no *Youtube* quanto no *WordPress*.

- *Ética e Direitos Autorais:* as questões éticas são tratadas neste PGD quando se relata que os entrevistados serão consultados antes mesmo das gravações, a fim de se obterem as permissões necessárias para um futuro compartilhamento dos vídeos. Já sobre direitos intelectuais, não se relata de forma direta nada a respeito.

- *Armazenamento e* Backup: quanto a esses dois aspectos, o PGD pronuncia-se de forma clara e precisa, afirmando que os arquivos em MP4 serão armazenados gratuitamente em um novo servidor de mídia da University of Leeds denominado *MediaStore*, ao passo que as cópias de *backup* serão armazenadas na University of Leeds SAN *(Storage Area Network)*. Acrescenta-se também que esses dispositivos são configurados para uma resiliência e proteção de alto padrão, incluindo *backup* diário.

- *Seleção e Preservação:* afirma-se que a vida útil do conteúdo do site do *WordPress* que abrigará os vídeos da pesquisa é de no mínimo 5 anos. Não se faz menção a repositórios de dados, subentendendo-se que o próprio site criado para a os fins do projeto é que será responsável pelo armazenamento e divulgação dos dados. Também é dito que esse *website* será desenvolvido de modo a permitir manutenções futuras por outros sujeitos que não aqueles envolvidos na criação inicial.

- *Compartilhamento de Dados:* por se tratar de um conteúdo veiculado por intermédio do *Youtube* e *Wordpress* (hospedado em um servidor da universidade), entende-se que os vídeos poderão ser acessados por todos os públicos gratuitamente e sem restrições.

- *Responsabilidades e Recursos:* outro aspecto, sobre o qual são tecidos vários detalhes, diz respeito aos responsáveis por cada processo previsto no projeto. Sobre a filmagem e edição dos vídeos, a responsabilidade é atribuída à *Leeds Media Services*, uma empresa corporativa local conhecida e que possui os *hardwares* e *softwares* necessários. Sobre o desenvolvimento do *website*, a responsabilidade é designada a um membro da *University's Blended Learning Educational Support Team*, pormenorizando-se, inclusive, informações referentes ao currículo do desenvolvedor. Quanto ao abastecimento de conteúdo do *website* e salvamento dos arquivos, a responsabilidade fica a cargo do pesquisador principal do projeto e de seus auxiliares. Além disso, relata-se que a equipe *Research Data Leeds* será responsável por oferecer apoio à implementação do plano técnico.

4.1.2 PGD do Biotechnology and Biological Sciences Research Council

O *Biotechnology and Biological Sciences Research Council* é um órgão britânico de apoio à pesquisa na área de Ciências Biológicas, que visa avançar tecnologicamente nos campos da agricultura, bioprocessos, produtos químicos, alimentos, saúde, farmácia e outras áreas afins à biotecnologia, colaborando com a competitividade econômica e a melhoria da qualidade de vida no Reino Unido (BIOTECHNOLOGY AND BIOLOGICAL SCIENCES RESEARCH COUNCIL, 2019).

O PGD selecionado para análise denomina-se "*Drosophila Genetics*", projeto que trata do estudo genético de moscas. O texto integral deste PGD pode ser consultado no Anexo B deste trabalho. O texto encontra-se subdividido nas seguintes seções, além da parte inicial na qual constam as informações mais básicas:

- *Data areas and data types.*
- *Standards and metadata.*
- *Relationship to other data.*
- *Secondary use.*
- *Methods for data sharing.*
- *Proprietary data.*
- *Timeframes.*
- *Formats.*

O Quadro 8 sintetiza a verificação dos requisitos realizada durante a análise do PGD. Não obstante o pequeno número de páginas, observou-se a riqueza de detalhes acerca dos aspectos mais importantes de um plano técnico, ainda que de forma abreviada. Após o quadro, apresentam-se as considerações feitas a partir do exame do plano em discussão.

QUADRO 8 – *Checklist* do PGD do BBSRC

ITEM	P/A*	ITEM	P/A*
ID	✗	Forma de coleta	✓
Entidade Financiadora	✓	Tipo de documentação	✓
Número Concedido	✗	Questões Éticas	✗
Nome do Projeto	✓	Direitos Intelectuais	✓
Descrição do Projeto	✓	Modo de Guarda	✓
Pesquisador Principal	✗	Acesso e Segurança	✓
ORCID	✗	Retenção de Dados	✓
Contatos do Projeto	✗	Plano de Preservação	✓
Data da Primeira Versão	✗	Modo de Compartilhamento	✓
Data da Última Versão	✗	Restrições	✓
Políticas Relacionadas	✗	Responsáveis	✓
Tipo de dados	✓	Recursos Adicionais	✗

Legendas: **P** - Informação presente ✓ | **A** - Informação ausente ✗.
Fonte: Elaboração própria.

- *Dados Administrativos:* observa-se um acréscimo em relação ao PGD analisado anteriormente, referente à descrição do projeto, que apesar de breve, demonstra a importância de uma descrição para situar o leitor em termos disciplinares. Também não constam dados relativos à equipe participante do projeto.

- *Coleta de Dados:* o texto é claro e direto a respeito das características dos dados e do modo como estes serão coletados. Mais uma vez, a título de exemplo, apresentam-se abaixo tais informações:

- *Tipo dos dados:* medições experimentais, modelos, registros e imagens (de microscopia e de *western blot*[7]).
- *Formato:*
    - Imagens - .tif
    - Dados de planilhas - .csv
    - Dados textuais - .txt
- *Volume:* não previsto.
- *Tamanho:*
    - Dados microscópicos - Entre 100 GB e 1 TB;
    - Dados do *western blot* - Cerca de 1 GB;
    - Outros dados - Até 10 MB.

- *Documentação e Metadados:* o texto mostra-se altamente meticuloso a respeito da questão da documentação dos dados, fazendo menção a protocolos, controle e manutenção de ferramentas e instrumentos, e padronização da nomeação dos arquivos. Quanto aos metadados, o PGD não aponta nenhum padrão específico a ser adotado, afirmando apenas que as imagens microscópicas já carregam consigo os dados técnicos descritivos. Acrescenta-se, ao longo do documento, que os metadados serão registrados tanto no *DataRegistry* da universidade quanto no DataCite e que poderão ser publicamente pesquisáveis e descobertos (em conformidade com os princípios FAIR discutidos na Revisão de Literatura).

[7] *Western blotting* é uma técnica analítica empregada pela Biologia Molecular para detectar proteínas específicas em uma amostra de tecido biológico (ESLAMI; LUJAN, 2010). Disponível em: https://www.jove.com/video/2359/western-blotting-sample-preparation-to-detection. Acesso em: 30 jun. 2019.

- *Ética e Direitos Autorais:* por não se tratar de uma pesquisa que envolve pessoas como fonte de coleta de dados, o PGD não trata de aspectos éticos. Porém, no que concerne aos direitos intelectuais, relata-se que não é esperado que o estudo gere dados passíveis de patenteamento ou que precisem ser protegidos por direitos de propriedade.

- *Armazenamento e* Backup: o PGD declara que os dados oriundos da pesquisa serão depositados no *Enlighten: Research Data*, repositório de dados institucional da University of Glasgow, de acordo com as políticas da entidade financiadora e da universidade. Quanto ao *backup*, afirma-se brevemente que este será realizado por meio de armazenamento comercial digital e que os dados serão auditados duas vezes por ano, em conformidade com a ISO 27001.

- *Seleção e Preservação:* a medida tomada no sentido de preservação dos dados é a atribuição de um DOI para o conjunto de dados, no intuito de propiciar a identificação, o acesso e a citação dos dados. No que tange à retenção dos dados, o PGD prevê um prazo de 10 anos a partir da data de depósito, sendo que os dados devem ser disponibilizados apenas após a publicação do artigo para o qual foram coletados.

- *Compartilhamento de Dados:* a única referência ao compartilhamento dos dados ao longo do PGD é o depósito no repositório institucional *Enlighten: Research Data*, que pode prover o acesso aos dados a outros pesquisadores e interessados. Além disso, o autor do PGD manifesta uma consciência de que os dados microscópicos podem vir a ser reutilizados futuramente em uma abordagem distinta.

- *Responsabilidades e Recursos:* a única responsabilidade abordada no plano remete ao armazenamento dos dados, já abordado acima.

4.1.3 PGD do Engineering and Physical Sciences Research Council

Constituindo o principal órgão de financiamento à pesquisa na área de Engenharia e Ciências Físicas no Reino Unido, o *Engineering and Physical Sciences Research Council* destaca-se na produção científica situada no escopo da engenharia estrutural, fabricação, matemática, materiais avançados, química, tecnologias da saúde, elétrica, eletrônica, energia,

tecnologia da informação, entre outros campos afins (ENGINEERING AND PHYSICAL SCIENCES RESEARCH COUNCIL, 2019).

O PGD representante deste órgão na amostra deste estudo é intitulado de "*Synthetic Chemistry*" e segue uma estrutura notoriamente fiel ao *checklist* proposto pelo DCC, apresentando a mesma subdivisão deste, como se pode ver no Anexo C.

O Quadro 9, como segue abaixo, mostra a verificação dos requisitos presentes no *checklist* do DCC, seguindo os mesmos procedimentos realizados na análise dos planos abordados nas seções anteriores.

QUADRO 9 – *Checklist* do PGD do EPSRC

ITEM	P/A*	ITEM	P/A*
ID	✗	Forma de coleta	✓
Entidade Financiadora	✓	Tipo de documentação	✓
Número Concedido	✗	Questões Éticas	✓
Nome do Projeto	✓	Direitos Intelectuais	✓
Descrição do Projeto	✓	Modo de Guarda	✓
Pesquisador Principal	✗	Acesso e Segurança	✓
ORCID	✗	Retenção de Dados	✓
Contatos do Projeto	✗	Plano de Preservação	✓
Data da Primeira Versão	✗	Modo de Compartilhamento	✓
Data da Última Versão	✗	Restrições	✓
Políticas Relacionadas	✗	Responsáveis	✓
Tipo de dados	✓	Recursos Adicionais	✓

Legendas: **P** - Informação presente ✓ | **A** - Informação ausente ✗.
Fonte: Elaboração própria.

Dos três PGDs analisados até o momento, o documento da EPSRC é o que mais se adequa aos padrões estabelecidos pelo DCC, respondendo diretamente às perguntas propostas no *checklist*, excetuando-se àquelas pertencentes à categoria de *Dados Administrativos*, que mais uma vez é ocultado para a privacidade dos pesquisadores envolvidos. Verificou-se uma

estreita semelhança entre a estrutura do PGD do EPSRC e a do BBSRC, discutido na seção antecessora. Procede-se, a seguir, com a análise do PGD relativo ao EPSRC.

- *Dados Administrativos:* assim como no PGD com o qual guarda forte similaridade, o plano apresenta apenas o nome do projeto, uma breve descrição do mesmo, o nome da entidade financiadora e a instituição na qual o projeto está sendo desenvolvido; no caso, a University of Glasgow.

- *Coleta de Dados:* por tratar-se de uma pesquisa experimental, entende-se que a coleta de dados se dará em laboratório, estando os dados previstos para serem registrados e agrupados utilizando-se planilhas do Excel. O texto é extremamente vago e incerto a respeito do tamanho total que os dados poderão assumir no final do processo.

- *Tipo dos dados:*
    - Parâmetros de reação;
    - Dados espectroscópicos e de caracterização geral dos compostos químicos.
- *Formato:* .pdf
- *Volume:* indefinido.
- *Tamanho:*
    - Dados referentes aos parâmetros - “Z MB”
    - Dados espectroscópicos - “entre X-Y GB”.

- *Documentação e Metadados:* o documento não cita nenhum padrão de metadados específico, mencionando apenas que os metadados constarão no *University of Glasgow Research Data Registry* e no DataCite. Em contrapartida, o texto é exaustivo a respeito da documentação dos dados, relatando que serão elaborados documentos de planilha e arquivos de texto para a descrição dos procedimentos realizados, além de mencionar também uma convenção a ser feita para padronizar a atribuição de nomes a pastas e arquivos, o que é de relevância indiscutível para uma boa documentação de dados.

- *Ética e Direitos Autorais:* ao contrário dos outros planos até aqui analisados, que no geral não fazem alusão alguma a aspectos não aplicáveis ao projeto, o PGD da EPSRC deixou claro não envolver pessoas nem amostras na coleta de dados, o que exclui implicações de natureza

ética. Já no que concerne aos direitos de propriedade intelectual, mencionam-se os sujeitos associados e é deixado explícito que há o objetivo de patentear o procedimento final.

- *Armazenamento e* Backup: o PGD também é claro e objetivo a respeito desses dois aspectos, narrando que os dados serão armazenados em discos rígidos pertencentes aos pesquisadores envolvidos no projeto e o *backup* será feito em servidores locais da *School of Chemistry* da instituição. No tocante ao acesso e à segurança, relata-se que os arquivos serão criptografados de modo a permitir o acesso somente do investigador principal e outros pesquisadores participantes, além de que a transferência de dados entre as duas partes se dará mediante dispositivos de armazenamento de memória ao invés de via e-mail, o que assegura ainda mais um alto nível de segurança a esses dados sujeitos a patenteamento.

- *Seleção e Preservação:* em virtude das intenções de patenteamento, consequentemente alguns dados serão selecionados para ficarem retidos durante algum tempo, haja vista as publicações consideradas de valor a longo prazo, sendo os dados liberados para compartilhamento apenas após a proteção da propriedade intelectual gerada, por meio de patente. Quanto ao plano de preservação, o PGD segue mais ou menos as mesmas diretivas descritas para a preservação dos dados da seção antecessora, relativa ao PGD do BBSRC, à exceção dos dados atinentes ao pedido de patenteamento, os quais serão excepcionalmente armazenados em servidor da *School of Chemistry* da universidade.

- *Compartilhamento de Dados:* por ser realizada na University of Glasgow, os dados serão depositados na *Enlighten: Research Data*, após o patenteamento e as publicações dos artigos. Além disso, relata-se que o DOI do conjunto de dados também será vinculado às publicações do *Enlighten: Publication*, repositório de publicações da universidade, no intuito de ampliar a visibilidade dos dados. Somado a isso, informações sobre o conjunto de dados também serão exibidas nos perfis dos pesquisadores nas páginas da University of Glasgow, garantindo visibilidade ainda maior para os dados.

- *Responsabilidades e Recursos:* mais uma vez, o PGD do EPSRC se destaca pela objetividade e obediência aos requisitos do *checklist* do DCC, respondendo a questões relativas às responsabilidades e recursos de uma forma que os outros planos até agora analisados não abordaram de forma clara. Afirma-se que o investigador principal ficará encarregado do gerenciamento de dados do projeto, incluindo sua organização e armazenamento à medida que forem

sendo produzidos, ao passo que a equipe de tecnologia da informação da *School of Chemistry* será responsável pelo gerenciamento do servidor onde os dados serão armazenados. Ademais, também aponta-se que a responsabilidade pela gestão dos repositórios ficará a cargo da equipe da biblioteca da universidade. Já no que tange aos recursos, deixa-se claro que o pesquisador possui o *software* necessário à coleta dos dados e também fala-se a respeito dos fundos aplicados no depósito dos dados.

4.1.4 PGD do Economic and Social Research Council

O *Economic and Social Research Council* é conselho de pesquisa britânico que promove a pesquisa relacionada às Ciências Sociais e Econômicas, apoiando pesquisadores e estudantes de pós-graduação em instituições acadêmicas e institutos de pesquisa independentes (ECONOMIC AND SOCIAL RESEARCH COUNCIL, 2019).

O PGD desta entidade selecionado para análise é intitulado "*Realist Evaluation of Adapted Sex Offender Treatment Programs for Men with Intellectual Disability*" e foi redigido em um texto de três páginas numa estrutura de formulário, dividido e enumerado de acordo com o aspecto a ser abordado.

No Quadro 10, exibe-se a verificação dos itens presentes no PGD do ESRC em cotejo com os requisitos constantes no *checklist* do DCC. Em seguida, discute-se tópico por tópico, ressaltando-se tanto a enfatização e riqueza de detalhes atribuídas a determinados itens quanto a exiguidade de outros tópicos registrados no DCC.

QUADRO 10 – *Checklist* do PGD do ESRC

ITEM	P/A*	ITEM	P/A*
ID	✗	Forma de coleta	✓
Entidade Financiadora	✓	Tipo de documentação	✓
Número Concedido	✗	Questões Éticas	✓
Nome do Projeto	✓	Direitos Intelectuais	✓
Descrição do Projeto	✗	Modo de Guarda	✓
Pesquisador Principal	✓	Acesso e Segurança	✓
ORCID	✗	Retenção de Dados	✗
Contatos do Projeto	✗	Plano de Preservação	✓

Data da Primeira Versão	✗	Modo de Compartilhamento	✓
Data da Última Versão	✗	Restrições	✓
Políticas Relacionadas	✗	Responsáveis	✓
Tipo de dados	✓	Recursos Adicionais	✗

Legendas: **P** - Informação presente ✓ | **A** - Informação ausente ✗.
Fonte: Elaboração própria.

- *Dados Administrativos*: seguindo o exemplo dos planos examinados anteriormente, o PGD do ESRC não expõe muitos detalhes sobre as informações de caráter administrativo do projeto de pesquisa, limitando-se a mencionar as informações mais básicas, como a entidade financiadora e o nome do projeto. No entanto, ao contrário dos outros planos analisados, neste consta o nome do pesquisador principal, havendo no canto superior esquerdo da primeira página um comentário ressaltando que a publicação do PGD recebeu a permissão prévia do pesquisador, além de frisar que o conteúdo do PGD não deve ser replicado, justamente devido ao fato de que um PGD constitui um documento exclusivo de um projeto de pesquisa específico, cabendo a cada investigador elaborar seu próprio PGD adequadamente à sua proposta de pesquisa.

- *Coleta de Dados*: destaca-se, logo de início, que a pesquisa recorrerá a fontes de dados pré-existentes; porém, a abordagem a ser adotada é diferente daquelas praticadas por pesquisas precedentes, o que leva à necessidade de se coletar novos dados de acordo com a proposta da pesquisa em questão, original em relação às outras já realizadas sobre a mesma temática. Aponta-se que, apesar de já existirem pesquisas que avaliam os programas de tratamento a serem investigados, não foram realizadas a partir das perspectivas do grupo focal. O método de coleta de dados a ser adotado é a entrevista, cujos dados colhidos serão analisados utilizando-se o NVivo 9[8].

- *Tipo dos dados:*
  - Dados qualitativos: áudios gravados nas entrevistas;
  - Dados quantitativos: extraídos dos arquivos dos pacientes.
- *Formato:*
  - Áudios: .mp3 ou .wav;

[8] *Software* utilizado para análise de dados qualitativos e de métodos mistos, que proporciona a organização, categorização, análise, visualização, armazenamento e recuperação de dados, seja em formato de texto, áudio, vídeo, e-mail, imagem, planilha, pesquisa na *web*, entre outros. Disponível em: https://www.qsrinternational.com/nvivo/what-is-nvivo. Acesso em: 10 ago. 2019.

    - Imagens: .jpeg;
    - Texto: .pdf;
    - Arquivos SPSS: .sav;
- *Volume:* não previsto.
- *Tamanho:* não previsto.

- *Documentação e Metadados*: o plano esclarece que a formatação dos dados e os metadados estarão em conformidade com os padrões e diretrizes do UKDA[9]. Acrescenta-se que os metadados incluirão uma descrição clara dos dados, como informações contextuais e identificador único para cada transcrição de gravação, além de detalhes das entrevistas, como data, local e entrevistado, respeitando-se, porém, as questões éticas envolvidas. Além disso, menciona-se também que os arquivos serão nomeados seguindo uma estrutura organizada que envolverá a criação de nomes consistentes, significativos e breves para pastas e arquivos, proporcionando uma fácil recuperação.

- *Ética e Direitos Autorais*: por se tratar de uma pesquisa cuja coleta de dados é realizada por meio de entrevista, o PGD deixa claro que os dados oriundos das gravações serão totalmente anonimizados e os grupos focais não serão divulgados, devido ao fato de a população entrevistada ser considerada vulnerável, a não ser que seja dado o devido consentimento por parte de cada um dos participantes. São propostas, ainda, técnicas para eliminar qualquer rastro de identificação, como frases peculiares, eventos exaustivamente detalhados ou histórias pessoais. O autor do projeto ainda descreve alguns procedimentos que poderão ser tomados para convencer os participantes a autorizarem o compartilhamento dos dados gerados nas entrevistas, ainda que de forma parcial, não deixando, claro, de preservar a identidade desses indivíduos. No tocante à questão dos direitos autorais, informa-se que a propriedade intelectual dos dados será de posse da University of Leeds, ressalvando-se, no entanto, que a pesquisa não utilizará nenhum dado coberto pela *Copyright, Designs and Patents Act 1988*[10], tampouco por outra legislação semelhante.

[9] Repartição integrante da infraestrutura do ESRC, responsável por promover a descoberta, o acesso e o suporte a dados da área de Ciências Sociais, desempenhando importante papel no desenvolvimento e na manutenção de metadados e padrões de citação de dados. Disponível em: https://www.data-archive.ac.uk/expertise/metadata-and-data-discovery/. Acesso em: 10 ago. 2019.
[10] Lei que regulamenta os direitos de propriedade intelectual no Reino Unido. Disponível em: http://www.legislation.gov.uk/ukpga/1988/48/contents. Acesso em: 10 ago. 2019.

- *Armazenamento e* Backup: o autor do projeto esclarece que os dados eletrônicos serão armazenados na *Storage Area Network* (SAN) da University of Leeds, a qual fornece servidores de arquivos em *data centers* fisicamente seguros contra incêndios. A descrição acerca dos procedimentos de *backup* dos arquivos é extensa e profundamente detalhada. Afirma-se que serão tirados *snapshots*[11] dos arquivos diariamente às 22h, sendo mantidos por um mês. Em um segundo nível, *snapshots* serão obtidos mensalmente e mantidos por onze meses. Quanto aos *backups* propriamente ditos, determina-se que será feito um *backup* incremental[12] todas as noites, a ser mantido por 28 dias, ao passo que um *backup* completo será feito mensalmente. Trimestralmente, as fitas de *backup* serão transferidas para um armazenamento a longo prazo, a serem mantidas por 12 meses em cofres à prova de fogo, sendo seu acesso controlado pelo *Active Directory Group*, sendo posteriormente deslocadas para lugares seguros fora do *campus* da universidade.

- *Seleção e Preservação*: o PGD deixa claro que quaisquer dados confidenciais (conforme o *Data Protection Act 1998*[13]) armazenados em dispositivos eletrônicos serão protegidos por *software* de criptografia conforme o padrão FIPS 140-2[14]. Salienta-se, ainda, que dados altamente confidenciais não estarão disponíveis fora do *campus* da universidade. Não se menciona nenhuma proposta de depósito dos dados em repositório, mas sim no provedor de serviços de dados do UKDA. No entanto, ressalta-se que os arquivos de dados serão convertidos em formatos abertos adequados à preservação a longo prazo.

---

[11] A diferença entre *snapshot* e *backup* reside no fato de que o primeiro diz respeito ao registro de um estado de determinado sistema, aplicação ou arquivo em determinado momento, análogo a uma imagem, servindo a propósitos de restauração de arquivos e obedecendo à forma como estes estavam exatamente no momento da captura. O *snapshot* não armazena uma cópia exata, mas sim metadados sobre o estado do arquivo ou sistema em determinado ponto do tempo. Sendo assim, *snapshots* não constituem, portanto, cópias de *backup* propriamente ditas, ocupando, inclusive, muito menos espaço em disco. Disponível em: https://www.controle.net/faq/o-que-e-snapshot. Acesso em: 10 ago. 2019.

[12] *Backup* incremental é um tipo de *backup* no qual se realiza uma cópia apenas dos dados que foram modificados desde o último *backup* feito anteriormente. Sua vantagem reside principalmente no fato de que, como é copiada uma quantidade menor de dados, o procedimento é mais rápido e requer menor espaço de armazenamento que um *backup* completo. Disponível em: http://www.aliancatecnologia.com/conteudo/2015/05/quatro-tipos-de-backup/. Acesso em: 10 ago. 2019.

[13] Lei britânica que protege as pessoas dispondo sobre como os dados pessoais podem ser usados por organizações ou órgãos governamentais. Disponível em: http://www.legislation.gov.uk/ukpga/1998/29/contents. Acesso em: 10 ago. 2019.

[14] Norma técnica criada pelo governo estadunidense para especificar requisitos de segurança referentes à criptografia em um sistema de segurança que proteja informações confidenciais em sistemas de computadores e telecomunicações. Disponível em: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.140-2.pdf. Acesso em: 10 ago. 2019.

- *Compartilhamento de Dados*: em virtude da natureza altamente sensível da pesquisa, tendo em vista o perfil dos participantes da pesquisa, o compartilhamento dos dados dificilmente poderia ser realizado livre de grandes restrições e critérios. O autor do projeto assevera que a reutilização dos dados está sujeita ao consentimento dos indivíduos envolvidos na pesquisa, sendo essencial que suas identidades sejam efetivamente preservadas.

- *Responsabilidades e Recursos*: o autor do projeto assume a responsabilidade geral pela implementação do plano de gestão dos dados da pesquisa, cabendo aos gestores de informação e tecnologia da faculdade a incumbência de garantir que as permissões de acesso aos arquivos sejam atribuídas corretamente e de aconselhar sobre aspectos relativos ao armazenamento e à segurança dos dados.

4.1.5 PGD do Horizon 2020

O programa *Horizon 2020*, iniciado em janeiro de 2014, é considerado o maior projeto da UE em matéria de pesquisa e inovação, tendo recebido um investimento de cerca de 75 bilhões de euros a serem aplicados em pesquisas de diferentes campos do conhecimento no período compreendido entre 2014 e 2020. O H2020 possui como objetivos principais: impulsionar o crescimento econômico, gerar empregos, produzir ciência de alta qualidade no âmbito da UE, superar as barreiras à inovação e facilitar o trabalho conjunto entre os setores público e privado (EUROPEAN COMMISSION, 2019).

Ao partir para a análise do PGD do H2020, observou-se com clareza maior exaustividade (13 páginas de texto), detalhismo e completeza em comparação com os outros planos analisados neste estudo. O PGD do H2020 (Anexo E), cujo projeto intitula-se "*Deformable Surface Tracking and Alpha Matting for the Automation of Post-production Workflows*", apresentou uma riqueza de informações não encontrada nos outros planos integrantes da amostra desta pesquisa, sobressaindo-se, portanto, como o PGD que mais fielmente cumpriu os requisitos propostos pelo *checklist* do DCC. O Quadro 11 mostra a verificação simples realizada a partir do cotejo entre os itens listados no documento-exemplo do DCC e aqueles presentes no PGD de pesquisadores financiados pelo H2020.

QUADRO 11 – *Checklist* do PGD do H2020

ITEM	P/A*	ITEM	P/A*
ID	✗	Forma de coleta	✓
Entidade Financiadora	✓	Tipo de documentação	✓
Número Concedido	✓	Questões Éticas	✓
Nome do Projeto	✓	Direitos Intelectuais	✓
Descrição do Projeto	✓	Modo de Guarda	✓
Pesquisador Principal	✓	Acesso e Segurança	✓
ORCID	✗	Retenção de Dados	✓
Contatos do Projeto	✓	Plano de Preservação	✓
Data da Primeira Versão	✓	Modo de Compartilhamento	✓
Data da Última Versão	✓	Restrições	✓
Políticas Relacionadas	✗	Responsáveis	✓
Tipo de dados	✓	Recursos Adicionais	✓

Legendas: **P** - Informação presente ✓ | **A** - Informação ausente ✗.
Fonte: Elaboração própria.

O PGD desse projeto segue uma estrutura e uma ordenação de informações muito similar àquelas recomendadas pelo DCC. Em sua introdução, evidencia-se que o projeto está inserido no plano de ação do H2020 relativo ao acesso aberto a dados de pesquisa. Ao final do documento, há um quadro-resumo organizado em várias colunas que classificam o método de produção dos dados (se serão criados ou coletados), a categoria dos dados e sua respectiva descrição, tipo (imagem, vídeo, texto, *plugin*, etc.), formato (.zip, .cpp, .dpx, .jpg, etc.), tamanho (em MB, GB ou TB), proprietário, modo de armazenamento, frequência de *backups* (se serão diários, mensais ou de periodicidade variável), necessidade ou não de destruição dos dados após o fim do projeto e tempo pelo qual cada tipo de dado deverá ser preservado (duração em anos). A seguir, procede-se com a análise descritiva do PGD de acordo com cada uma das categorias adotadas neste estudo.

- *Dados Administrativos*: ao contrário de todos os outros PGDs examinados neste estudo, o plano do H2020 apresenta vários dados administrativos que identificam o projeto de pesquisa,

tais como a data da primeira e última versão atualizada do PGD, nome dos pesquisadores envolvidos (inclusive e-mail de contato), além de um resumo descrevendo o conteúdo do PGD, e também um número de referência concedido pelo H2020, o qual é "H2020-ICT-18-2014 GA-644628", espécie de informação não divulgada nos outros planos analisados. Outro detalhe interessante é a atribuição de um acrônimo ao projeto, apelidado de "AutoPost", diretamente alusivo ao título do projeto de pesquisa, e que permite ao mesmo ser identificado e referenciado com muito mais simplicidade[15].

- *Coleta de Dados*: o PGD informa que o projeto irá tanto coletar dados já existentes quanto produzir novos dados, os quais poderão enquadrar-se em cinco categorias: dados para avaliação (em forma de imagens e vídeos), *software* de computador, dados e metadados de pesquisa, manuscritos e material de divulgação. Por se tratar de uma pesquisa que prevê a coleta de dados de naturezas e formatos diversos, vários são os métodos de coleta/produção de dados e formatos de arquivos a serem adotados. Abaixo, são listadas as extensões de arquivo previstas para serem adotadas ao longo do projeto, de acordo com o tipo de dado:

- Dados de avaliação: .jpg e .png (para imagens); .dpx, .mov, .mxf e .nk (para vídeos);
- *Softwares* de computador: .zip e .cpp;
- Dados e metadados: .png e .exr (para imagens); .doc, .pdf e .txt (para textos);
- Manuscritos: .docx e .pdf (versões finais); .odt e .tex (versões intermediárias);
- Materiais de divulgação: .mov e .avi (para vídeos); .pdf (para outros).

- *Documentação e Metadados*: o PGD declara que a documentação incluirá informações sobre as metodologias empregadas, procedimentos, suposições levantadas, além de informações referentes ao formato dos arquivos e ao tipo dos dados armazenados. A respeito da padronização da nomenclatura dos arquivos, o PGD determina que estes serão denominados de acordo com seu conteúdo, acrescentando que, no caso de arquivos com mais de uma versão, esta será especificada ao final do nome, como por exemplo "nomedoarquivo_v1".

- *Ética e Direitos Autorais*: quanto às questões éticas, informa-se que serão obtidas autorizações dos atores que aparecerão nos vídeos do projeto para que se possa distribuir essas imagens. Já sobre direitos de propriedade intelectual, o PGD lista no quadro ao final do arquivo

[15] Simplicidade de identificação e referenciação no sentido de que é muito mais descomplicado e conveniente aludir a algo (instituição, programa de governo, projeto de pesquisa, entre outros exemplos) por meio de um acrônimo ou uma sigla do que pronunciando seu nome por extenso.

os proprietários de cada um dos itens de dados existentes no projeto de pesquisa (Figura 9), esclarecendo que direitos de acesso serão concedidos por meio de licenças gratuitas.

FIGURA 9 – Exemplo de informações relevantes sobre os dados de uma pesquisa

Category	Type	Format	Size	Owner	Privacy level	Storage / Storage for public access
Computer Software	Library (SDK)	ZIP	40MB	EC	Consortium	Eurecat's redmine
Computer Software	Library (SDK)	ZIP	<10MB	HHI	Consortium	Eurecat's redmine
Computer Software	Plugin	ZIP	100 MB	IL	Consortium	Eurecat's redmine
Computer Software	Plugin	ZIP	100 MB	IL	Consortium	Eurecat's redmine
Computer Software	Source code (library)	C,CPP		EC	EC	EC
Computer Software	Source code (library)	C,CPP		HHI	HHI	HHI
Computer Software	Source code (plugin)	C,CPP	200 MB	IL	IL	IL
Computer Software	Source code (plugin)	C,CPP	200 MB	IL	IL	IL
Data for evaluation	File	NK	<1M	Producer of the environment	Consortium	owncloud
Data for evaluation	Images	DPX	135GB	DG	Consortium	DG
Data for evaluation	Images	DPX	10GB	MOT	Consortium	MOT
Data for evaluation	Images	4K	1.2TB	HHI	Consortium	HHI
Data for evaluation	Images	JPG	2GB	Hollywood Camera Works	Public	MOT
Data for evaluation	Images	PNG	21MB	alphamatting.com	Public	EC
Data for evaluation	Videos	PNG	2.6GB	videomatting.com	Public	EC
Data for evaluation	Videos	DPX	>10GB	ALL	Consortium	MOT,DG, hard drive at all premises

Fonte: Caballero e Fuenmayor (2015).

- *Armazenamento e* Backup: sobre o armazenamento dos dados, é dito que estes serão armazenados de acordo com sua categoria, nível de privacidade e tamanho, como se pode ter um panorama na figura exposta acima. Os dados de *software* são previstos para serem armazenados em um servidor Redmine[16] hospedado no Eurecat[17]. Quanto ao *backup*, a dinâmica de realização de cópias se dará de acordo com a categoria de dados, podendo ser feita em servidores na nuvem ou em discos rígidos.

- *Seleção e Preservação*: alguns dados poderão ser retidos pelos parceiros do projeto, para fins de comercialização, ao passo que versões anteriores de SDKs serão destruídas ao final do projeto. Os dados de imagem e vídeo gerados ao longo da pesquisa deverão ser mantidos pelos parceiros do projeto para serem utilizados em outros projetos, ou para fins de validação.

[16] Aplicativo da *web* de gerenciamento de projetos que utiliza código aberto. Disponível em: http://www.redmine.org. Acesso em: 15 ago. 2019.

[17] Centro de tecnologia da Catalunha (Espanha) que oferece, entre outros, serviços de consultoria em tecnologia da informação, possuindo equipamentos e instalações de alta qualidade. Disponível em: https://ec.europa.eu/growth/tools-databases/kets-tools/infrastructure/eurecat-technology-centre. Acesso em: 15 ago. 2019.

Quanto à preservação, alguns tipos de dados deverão ser preservados durante pelo menos três anos, ao passo que outros não precisarão ser mantidos devido ao fato de não fornecerem valor agregado para a comunidade de pesquisa ou por já serem preservados por alguma entidade. De um modo geral, os dados serão preservados pelos parceiros do projeto (vide última coluna da Figura 9), assim como sua manutenção nos devidos repositórios.

- *Compartilhamento de Dados*: em consonância com a política de acesso aberto às publicações e aos dados dos projetos financiados pelo H2020, o PGD aponta o Zenodo[18] como repositório oficial dos dados produzidos no decorrer do projeto. Como os dados serão compartilhados apenas para fins de pesquisa e instrução, os interessados em acessá-los terão que explicar o uso que farão dos dados, além de assinar uma licença limitando seu uso e distribuição. No caso de vídeos de grande tamanho, os usuários autorizados poderão baixá-los integralmente do servidor da Eurecat ou acessá-los fisicamente desde que assumam os custos de armazenamento e transporte. Além disso, os dados do AutoPost serão divulgados no *website* do projeto, que permanecerá ativo por pelo menos 4 anos após o término do projeto.

- *Responsabilidades e Recursos*: o documento atribui à Eurecat, uma das entidades coordenadoras do projeto, a responsabilidade pela implementação do PGD. Como o projeto possui vários parceiros, cada um é incumbido de uma tarefa específica, seja pela geração de dados, produção de metadados, armazenamento, *backup* ou compartilhamento. Recursos adicionais, como mídia de armazenamento físico, também são ligeiramente abordados.

A Figura 10, extraída do quadro presente no PGD, mostra a quantidade de parceiros que o AutoPost possui, cada qual com suas responsabilidades próprias, particularmente aquelas relativas ao armazenamento, demonstrando ser um projeto complexo com alto nível de coordenação entre diversos entes.

[18] Tendo seu nome derivado de Zenodotus, bibliotecário da antiga Biblioteca de Alexandria, e tido como autor do primeiro uso registrado de metadados, o Zenodo é um repositório de dados de pesquisa de código aberto que não impõe requisitos de formato, tamanho, restrições de acesso ou licença aos depositários. Além disso, colabora com a recompensa de pesquisadores, ajudando-os a receber seus devidos créditos na medida em que torna citáveis os dados de pesquisa. Disponível em: http://about.zenodo.org/. Acesso em: 15 ago. 2019.

FIGURA 10 – Responsáveis pelo armazenamento dos dados do projeto AutoPost

Storage / Storage for public access
Eurecat's redmine
Eurecat's redmine
Eurecat's redmine
Eurecat's redmine
EC
HHI
IL
IL
owncloud
DG
MOT
HHI
MOT
EC
EC
MOT,DG, hard drive at all premises
MOT,DG, hard drive at all premises
IL, hard drive at all premises
IL, EC, hard drive at all premises / proxy on Zenodo
owncloud / Zenodo
owncloud / Zenodo
owncloud / Zenodo
owncloud / Zenodo
Partners servers / Zenodo

Fonte: Caballero e Fuenmayor (2015).

4.2 Análise comparativa dos documentos

Perseguindo o segundo objetivo específico deste estudo, procede-se com a análise comparativa dos planos de gestão de dados examinados. Nas seções anteriores, levantaram-se ligeiras comparações entre os documentos, apontando-se semelhanças e diferenças entre um e outro.

Percebe-se, primeiramente, que dados pessoais dos pesquisadores, como o ORCID e o ResearcherID, não constaram em nenhum dos planos, o que leva a crer que tais informações provavelmente constavam nos documentos originais, mas foram ocultados para efeitos de

divulgação pública, priorizando-se o conteúdo referente aos processos mais destacados da gestão de dados, como documentação, armazenamento e preservação.

O único PGD a comunicar dados administrativos de modo mais extensivo foi o do H2020, coincidentemente considerado, após a análise dos documentos, o PGD mais "completo" da amostra, respondendo ao maior número de questões levantadas no *checklist* do DCC e destacando-se em relação aos outros PGDs no que tange à cobertura das oito categorias de informações abordadas nesta análise.

Verificou-se que todos os planos versaram sobre questões relativas ao tipo de dados coletados, forma de coleta, documentação, armazenamento, acesso, preservação e responsabilidades. Dos cinco, quatro dissertaram sobre questões éticas, direitos intelectuais, modo de compartilhamento e restrições. Informações relativas à retenção de dados e recursos adicionais foram apresentadas quando aplicáveis.

Observou-se, também, que enquanto um documento conferia maior ênfase a algum tópico, outros PGDs demonstravam maior preocupação com outros aspectos, o que pode ser justificado pelo fato de que cada projeto de pesquisa, inserido em seu domínio disciplinar, possui suas características próprias e unívocas, o que significa que certas questões serão naturalmente consideradas e discutidas, ao passo que outras não serão aplicáveis ou pertinentes.

O Quadro 12, na página seguinte, evidencia de maneira simples essa comparação entre os PGDs, ressaltando os elementos presentes e ausentes em cada um deles.

QUADRO 12 – Quadro comparativo dos PGDs

ITEM	AHRC	BBSRC	EPSRC	ESRC	H2020
ID	✗	✗	✗	✗	✗
Entidade Financiadora	✓	✓	✓	✓	✓
Número Concedido	✗	✗	✗	✗	✓
Nome do Projeto	✓	✓	✓	✓	✓
Descrição do Projeto	✗	✓	✓	✗	✓
Pesquisador Principal	✗	✗	✗	✓	✓
ORCID	✗	✗	✗	✗	✗
Contatos do Projeto	✗	✗	✗	✗	✓
Data da Primeira Versão	✗	✗	✗	✗	✓
Data da Última Versão	✗	✗	✗	✗	✓
Políticas Relacionadas	✗	✗	✗	✗	✗
Tipo de dados	✓	✓	✓	✓	✓
Forma de coleta	✓	✓	✓	✓	✓
Tipo de documentação	✓	✓	✓	✓	✓
Questões Éticas	✓	✗	✓	✓	✓
Direitos Intelectuais	✗	✓	✓	✓	✓
Modo de Guarda	✓	✓	✓	✓	✓
Acesso e Segurança	✓	✓	✓	✓	✓
Retenção de Dados	✗	✓	✓	✗	✓
Plano de Preservação	✓	✓	✓	✓	✓
Modo de Compartilhamento	✗	✓	✓	✓	✓
Restrições	✗	✓	✓	✓	✓
Responsáveis	✓	✓	✓	✓	✓
Recursos Adicionais	✗	✗	✓	✗	✓

Legendas: **P** - Informação presente ✓ | **A** - Informação ausente ✗.
Fonte: Elaboração própria.

5 DISCUSSÃO DOS RESULTADOS

A análise dos planos de gestão de dados dos pesquisadores financiados pelos conselhos europeus escolhidos para este estudo permitiu observar e examinar na prática aquilo que tem sido discutido na literatura dos últimos anos a respeito da gestão de dados científicos.

Este estudo adotou como parâmetro os requisitos apontados pelo DCC, centro de referência de iniciativas relacionadas à *e-Science*, o qual disponibiliza uma série de modelos de planos de gestão de dados procurando atender às imposições das agências de financiamento do Reino Unido, assumindo uma postura vanguardista em relação à gestão de dados de pesquisa (BORGMAN, 2012).

Observou-se que muitas das indagações feitas no *checklist* do DCC não foram respondidas pelos PGDs analisados, o que demonstra duas realidades: a primeira é que o DCC é exaustivo e altamente detalhista em seu roteiro para elaboração de PGDs; segundo, que nem todas as questões relativas à gestão de dados serão necessariamente pertinentes a todos os projetos de pesquisa, pois cada um possui contexto e características próprias de coleta, documentação, armazenamento e compartilhamento.

Um dos pontos que merecem destaque é aquele referente aos dados cujo acesso ou divulgação são alvo de algum tipo de restrição, seja "devido à possibilidade de patentes, por interesses comerciais, por segurança ou por se tratarem de dados sensíveis que precisam de tratamentos específicos, como processos de anonimização" (SAYÃO; SALES, 2016b, p. 68).

Da perspectiva do acesso e uso, sobressaem as licenças associadas aos dados, as quais "definem as responsabilidades e os limites de uso que os usuários devem respeitar, incluindo nesse escopo o reconhecimento da autoria dos dados por meio de citação padronizada" (SAYÃO; SALES, 2016b, p. 69).

Um fato que não pode ser ignorado é a discrepância ainda existente entre as práticas de gestão de dados empreendidas na Europa e aquelas observadas no Brasil, ainda em fase embrionária, visto que ainda existem muitas barreiras culturais e tecnológicas que dificultam a modernização efetiva do sistema de comunicação científica (CASTELLI; MANGHI; THANOS, 2013). Devido a isso é que se optou por trabalhar com documentos estrangeiros, assim obtendo-se resultados mais satisfatórios do que aqueles que provavelmente se obteriam caso este estudo realizasse um recorte brasileiro.

A partir da análise dos PGDs, foi possível enxergar mais de perto aquilo que é discutido na literatura a respeito os dados de pesquisa. Observou-se que os dados podem ser de vários tipos, como números, imagens, textos, vídeos, áudios, *softwares*, animações, modelos,

entre outros. Conforme postula na literatura Sayão e Sales (2015, p. 7), notou-se por meio da comparação entre os documentos analisados que “alguns tipos de dados têm valor imediato e duradouro, enquanto outros adquirem valor ao longo do tempo; alguns dados são capturados num momento específico e irrecuperável, enquanto outros são passíveis de se reproduzir”.

No tocante às questões éticas, foi possível inferir, em conformidade com Borgman (2012), que alguns dados são anonimizáveis com mais facilidade e, portanto, mais fáceis de compartilhar, ao passo que dados derivados de entrevistas gravadas, por exemplo, são mais difíceis de anonimizar e de codificar em formatos consistentes que não distorçam o conteúdo original. Nesse aspecto, surge um impasse entre adaptar os dados para permitir seu compartilhamento (correndo o risco de adulteração), armazenar com acesso terminantemente restrito ou simplesmente destruir esses dados ao final da pesquisa.

As listas de verificação formuladas para cada um dos cinco documentos examinados serviram como uma distinção entre as informações presentes e as informações ausentes em cada um dos planos, tendo-se destacado o plano pertencente ao H2020, o qual preencheu a maioria dos requisitos estabelecidos pelo *checklist* do DCC, ao passo que o plano do AHRC deixou muitos itens em branco.

Traçando um paralelo entre os itens presentes nos planos e o auxílio que os profissionais da informação podem prestar, pode-se corroborar com Corrêa (2016, p. 388), quando este afirma que “a assistência do bibliotecário aos pesquisadores ajuda a avaliar o que realmente necessitam entender, a compreenderem como organizar uma variedade de tipos de dados e tomar as decisões corretas sobre o acesso e preservação de dados para os seus projetos”.

Por fim, constatou-se uma forte consonância entre a discussão presente na teoria e os elementos contemplados pelos PGDs, o que parece demonstrar que as considerações feitas na literatura e as iniciativas tomadas na prática têm caminhado lado a lado.

6 CONSIDERAÇÕES FINAIS

A partir da revisão de literatura e da análise dos planos de gestão de dados selecionados para este estudo, concluiu-se que os dados científicos, apesar de estarem em ascensão na escala de prestígio acadêmico, ainda carecem de maior atenção e tratamento mais adequado por parte de pesquisadores e instituições acadêmicas. Enquanto as agências financiadoras fazem sua parte, instituindo politicas mandatórias de elaboração de planos de gestão de dados, cabe às universidades e institutos de pesquisa esclarecer a comunidade científica acerca dos benefícios que uma efetiva gestão de dados pode trazer a nível pessoal, coletivo e institucional.

Foi possível depreender que os pesquisadores ainda precisam de mais informações e treinamentos referentes ao gerenciamento de dados de pesquisa, incluindo os procedimentos de cada uma das etapas do ciclo de vida dos dados e o manejo das ferramentas tecnológicas necessárias para se realizar um tratamento adequado dos dados brutos de pesquisa, o que engloba processos de coleta, documentação, armazenamento, *backup*, preservação, compartilhamento e retenção de dados, para citar apenas alguns.

A gestão de dados de pesquisa é um processo complexo que requer a ação conjunta de pesquisadores, entidades financiadoras e profissionais da informação, os quais podem promover iniciativas de apoio à gestão de dados e de disseminação dos princípios da Ciência Aberta em suas instituições. A fim de se ajustar à dinâmica do novo paradigma da Ciência (dominado pela *e-Science*) e prestar efetivo auxílio aos pesquisadores como os usuários e produtores de informação que estes sempre foram, os bibliotecários podem empreender uma série de ações de amparo à gestão de dados científicos, desde que esteja ao alcance de suas competências e que sejam compatíveis com seus conhecimentos técnicos, os quais podem ser progressivamente aprimorados de acordo com as novas demandas da tecnologia e da comunidade de usuários que deve ser atendida.

Como sugestões de trabalhos futuros, recomenda-se que sejam realizadas pesquisas focadas na construção de repositórios de dados científicos ou em programas de treinamento para pesquisadores e profissionais da informação, capacitando-os para lidar com mais destreza com as ferramentas e os procedimentos envolvidos na gestão de dados de pesquisa.

REFERÊNCIAS

ALBAGLI, S.; APPEL, A. L.; MACIEL, M. L. E-Science e ciência aberta: questões em debate. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Florianópolis. **Anais [...]**. Florianópolis: ANCIB, 2013. Disponível em: http://ridi.ibict.br/handle/123456789/465. Acesso em: 16 jul. 2019.

ANDRADE, R. M.; MURIEL-TORRADO, E. Declarações de Acesso Aberto e a Lei de Direitos Autorais brasileira. **Reciis: Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, Rio de Janeiro, v. 11, nov. 2017. Suplemento. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1374. Acesso em: 18 jul. 2019.

ANJOS, R. L.; DIAS, G. A.; RODRIGUES, A. A. Dados científicos: as práticas de gestão dos pesquisadores brasileiros na Ciência da Informação. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília (SP). **Anais [...]**. Marília: Unesp, 2017. Disponível em: http://www.brapci.inf.br/index.php/article/download/58903. Acesso em: 7 dez. 2018.

APPOLINÁRIO, F. **Metodologia da ciência**: filosofia e prática da pesquisa. 2. ed. São Paulo: Cengage Learning, 2012. 226 p.

ARTS AND HUMANITIES RESEARCH COUNCIL. **What we do**. 2013. Disponível em: https://ahrc.ukri.org/about/what-we-do/. Acesso em: 22 jun. 2019.

BAPTISTA, A. A.; COSTA, S. M. S.; KURAMOTO, H.; RODRIGUES, E. Comunicação científica: o papel da "Open Archives Initiative" no contexto do Acesso Livre. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, jan./jun. 2007. Edição Especial. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2007v12nesp1p1/435. Acesso em: 16 jul. 2019.

BELL, G. Prefácio. *In*: HEY, T.; TANSLEY, S.; TOLLE, K. (org.) **O quarto paradigma**: descobertas científicas na era da eScience. São Paulo: Oficina de Textos, 2011. p. 11-15.

BERTIN, P. R. B.; MACHADO, C. D.; VISOLI, M. C.; DRUCKER, D. P.; PINTO, D. M.. A construção do Plano de Dados Abertos de uma organização pública de pesquisa e desenvolvimento e o desafio de uma Ciência Agropecuária aberta. **Reciis: Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, Rio de Janeiro, v. 11, nov. 2017. Suplemento. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1411. Acesso em: 18 jul. 2019.

BIBLIOTECA CEPAL. **Gestión de datos de investigación**. Disponível em: https://biblioguias.cepal.org/gestion-de-datos-de-investigacion/PGD. Acesso em: 23 jul. 2019.

BIOTECHNOLOGY AND BIOLOGICAL SCIENCES RESEARCH COUNCIL. **About us**. 2019. Disponível em: https://bbsrc.ukri.org/about/. Acesso em: 23 jun. 2019

BORGMAN, C. L. The conundrum of sharing research data. **Journal of the American Society for Information Science and Technology**, v. 63, n. 6, p. 1059-1078, 2012. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1002/asi.22634. Acesso em: 14 maio 2019.

BOURDIEU, P. O campo científico. *In*: ORTIZ, R. (org.). **Pierre Bourdieu**: sociologia. São Paulo: Ática, 1983. p. 122-155.

BRASIL. **Decreto nº 8.777, de 11 de maio de 2016**. Institui a Política de Dados Abertos do Poder Executivo Federal. Brasília: Presidência da República, 2016. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2016/Decreto/D8777.htm. Acesso em: 25 maio 2019.

BUENO, W. C. Comunicação científica e divulgação científica: aproximações e rupturas conceituais. **Informação & Informação**, Londrina (PR), v. 15, n. esp., p. 1-12, 2010. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/6585. Acesso em: 20 out. 2018.

CABALLERO, M.; FUENMAYOR, M. E. **AutoPost-D6.3**: Data Management Plan. Zenodo, 2015. Disponível em: https://zenodo.org/record/56107#.XXR69ihKiUn. Acesso em: 15 ago. 2019.

CARDANO, M. **Manual de pesquisa qualitativa**: a contribuição da teoria da argumentação. Petrópolis, RJ: Vozes, 2017. 371 p.

CASTELLI, D.; MANGHI, P.; THANOS, C. A vision towards scientific communication infrastructures: on bridging the realms of research digital libraries and scientific data centers. **International Journal on Digital Libraries**, n. 13, p. 155–169, 2013.

CERVO, A. L.; BERVIAN, P. A.; SILVA, R. **Metodologia científica**. 6. ed. São Paulo: Pearson Prentice Hall, 2007. 162 p.

CHIWARE, E.; MATHE, Z. Academic libraries' role in research data management services: a South African perspective. **South African Journal of Libraries and Information Science**, v. 81, n. 2, p. 1-10, 2015.

CORRÊA, F. C. O papel dos bibliotecários na gestão de dados científicos. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas (SP), v. 14, n. 3, p. 387-406, set./dez. 2016. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8646333. Acesso em: 17 maio 2018.

COSTA, M.; LEITE, F. C. L. Princípios e recomendações basilares para a comunicação dos dados de pesquisa. **Em Questão**, Porto Alegre, v. 23, n. 1, p. 87-112, jan/abr. 2017. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/65623. Acesso em: 28 set. 2018.

COX, A. M.; PINFIELD, S.; SMITH, J. Moving a brick building: UK libraries coping with research data management as a 'wicked' problem. **Journal of Librarianship and Information Science**, v. 48, n. 1, p. 3-17, 2016.

CRESWELL, J. W. **Projeto de pesquisa**: métodos qualitativo, quantitativo e misto. 3. ed. Porto Alegre: Artmed, 2010. 296 p.

CURTY, R. G. As diferentes dimensões do reuso de dados científicos. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. **Anais [...]**. Salvador: UFBA, 2016. p. 3299-3321. Disponível em: https://drive.google.com/file/d/0B7rxeg_cwHajMW9ZV0xFZHBhTnc/view. Acesso em: 2 jul. 2019.

CURTY, R. G.; AVENTURIER, P. O paradigma da publicação de dados e suas diferentes abordagens. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília (SP). **Anais [...]**. Marília: Unesp, 2017. Disponível em: http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/viewFile/468/820. Acesso em: 13 maio 2019.

DELFANTI, A.; PITRELLI, N. Ciência aberta: revolução ou continuidade? *In*: ALBAGLI, S.; MACIEL, M. L.; ABDO, A. H. (org.). **Ciência aberta, questões abertas**. Brasília: IBICT; Rio de Janeiro: UNIRIO, 2015. p. 59-69.

DIGITAL CURATION CENTRE. **Checklist for a Data Management Plan**. v.4.0. Edinburgh: Digital Curation Centre, 2013. Disponível em: http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf. Acesso em: 20 mar. 2019.

DUDZIAK, E. **Competências do bibliotecário na gestão de dados de pesquisa, comunicação científica e acesso aberto**. 2016. Disponível em: http://www.sibi.usp.br/noticias/competencias-gestao-dados-pesquisa/. Acesso em: 17 maio 2018.

ECONOMIC AND SOCIAL RESEARCH COUNCIL. **What we do**. 2019. Disponível em: https://esrc.ukri.org/about-us/what-we-do/. Acesso em: 10 ago. 2019.

ENGINEERING AND PHYSICAL SCIENCES RESEARCH COUNCIL. **Delivery Plan 2019**. 2019. Disponível em: https://epsrc.ukri.org/newsevents/pubs/deliveryplan2019/. Acesso em 29 jun. 2019.

EUROPEAN COMMISSION. **Horizon 2020**: Work Programme 2018-2020. Disponível em: https://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-intro_en.pdf. Acesso em: 15 ago. 2019.

FERREIRA, V. B.; VILLALOBOS, A. P. O.; MOURA, M. A. O modelo e-Science nos institutos nacionais de ciência e tecnologia de nanotecnologia: evidências de práticas colaborativas. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. **Anais [...]**. Salvador: UFBA, 2016. p. 3361-3380. Disponível em: https://drive.google.com/file/d/0B7rxeg_cwHajMW9ZV0xFZHBhTnc/view. Acesso em: 11 jul. 2019.

GIBBS, G. **Análise de dados qualitativos**. Porto Alegre: Artmed, 2009. 198 p.

GRANT, R. Recordkeeping and research data management: a review of perspectives. **Records Management Journal**, v. 27, n. 2, p. 159-174, 2017.

GRAY, J. Jim Gray e a eScience: um método científico transformado. Transcrição de palestra ministrada por Jim Gray no Conselho Nacional de Pesquisa (EUA), 11 jan. 2007. *In*: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (org.) **O quarto paradigma**: descobertas científicas na era da eScience. São Paulo: Oficina de Textos, 2011. p. 17-30.

HENNING, P.; RIBEIRO, C. J. S.; SALES, L.; MOREIRA, J.; SANTOS, L. O. B. S. Desmistificando os princípios FAIR: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados FAIR. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19., 2018, Londrina (PR). **Anais [...]**. Londrina: UEL, 2018. Disponível em: http://enancib.marilia.unesp.br/index.php/XIXENANCIB/xixenancib/paper/viewFile/1475/1717. Acesso em: 23 set. 2019.

HENNING, P. C.; RIBEIRO, C. J. S.; SANTOS, L. O. B.; SANTOS, P. X. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, Porto Alegre, v. 25, n. 2, p. 389-412, maio/ago. 2019. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/84753. Acesso em: 10 jul. 2019.

HOURCADE, V. **O movimento ciência aberta no Brasil**. 2015. 135 p. Dissertação (Mestrado em Divulgação Científica e Cultural) – Universidade Estadual de Campinas, Campinas, 2015.

HURD, J. M. The transformation of scientific communication: a model for 2020. **Journal of the American Society for Information Science**, v. 51, n. 14, p. 1279-1283, dez. 2000.

KATZ, J. S.; MARTIN, B. R. What is research collaboration? **Research Policy**, n. 26, p. 1-18, 1997.

KRUSE, F.; THESTRUP, J. B. **Research data management***:* an European perspective. Berlim: De Gruyter Saur, 2018.

LAFUENTE, A. Ciência 2.0. **Revista Madri+d**, 2006. Edição Especial. Disponível em: http://www.madrimasd.org/revista/revistaespecial1/articulos/lafuente.asp. Acesso em: 18 jul. 2019.

LEITE, F. C. L. **Modelo genérico de gestão da informação científica para instituições de pesquisa na perspectiva da comunicação científica e do acesso aberto**. 2011. 262 f. Tese (Doutorado em Ciência da Informação) — Universidade de Brasília, Brasília, 2011.

LÓSCIO, B. F.; VILA NOVA, S. Plano de dados abertos na Universidade Federal de Pernambuco - UFPE. **Reciis: Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, Rio de Janeiro, v. 11, nov. 2017. Suplemento. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1386. Acesso em: 5 jun. 2019.

MACHADO, J. Dados abertos e ciência aberta. *In*: ALBAGLI, S.; MACIEL, M. L.; ABDO, A. H. (org.). **Ciência aberta, questões abertas**. Brasília: IBICT; Rio de Janeiro: UNIRIO, 2015. p. 201-227.

MEADOWS, A. J. **A comunicação científica**. Brasília: Briquet de Lemos, 1999. 268 p.

MINAYO, M. C. S. O desafio da pesquisa social. *In*: MINAYO, M. C. S.; DESLANDES, S. F.; GOMES, R. **Pesquisa social**: teoria, método e criatividade. 28 ed. Petrópolis, RJ: Vozes, 2009. p. 9-29.

MONTEIRO, E. C. S. A.; SANT'ANA, R. C. G. Plano de gerenciamento de dados em repositórios de dados de universidades. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 23, n. 53, p. 160-172, set./dez., 2018. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p160. Acesso em: 24 out. 2018.

MORENO, F. P. Repositórios de dados de pesquisa na Espanha: breve análise. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 23, n. 53, p. 52-63, set./dez., 2018. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p52. Acesso em: 24 out. 2018.

MUELLER, S. P. M. A comunicação científica e o movimento de acesso livre ao conhecimento. **Ciência da Informação**, Brasília, v. 35, n. 2, p. 27-38, maio/ago. 2006.

OLIVEIRA, A. C. S. **Desvendando a autoralidade colaborativa na e-Science sob a ótica dos direitos de propriedade intelectual**. 2016. 297 f. Tese (Doutorado em Ciência da Informação) – Centro de Ciências Sociais Aplicadas, Universidade Federal da Paraíba, João Pessoa, 2016.

OPEN KNOWLEDGE FOUNDATION. **Manual dos dados abertos**: governo. Tradução e adaptação de Comunidade Transparência Hacker. 2010. Disponível em: http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf. Acesso em: 25 maio 2019.

PAVÃO, C. G.; COSTA, J. S. B.; HOROWITZ, Z.; FERREIRA, M. K.; CAREGNATO, S. E. Contribución del acceso abierto a la visibilidad de la literatura científica en una institución de educación superior. **E-colabora**, v. 2, n. 3, p. 48-66, jan./jun. 2012.

PRÍNCIPE, P.; SARAIVA, R. Serviços para suporte à gestão de dados científicos na UMINHO: plano de intervenção dos SDUM. *In*: CONGRESSO NACIONAL DE BIBLIOTECÁRIOS, ARQUIVISTAS E DOCUMENTALISTAS, 12., 2015, Évora. **Anais [...]**. Évora, 2015.

REBIUN. **Cita tus datos de investigación**. 2016. Disponível em: https://www.rebiun.org/node/184. Acesso em: 20 jun 2019.

ROCHA, L. L.; SALES, L. F.; SAYÃO, L. F. Uso de cadernos eletrônicos de laboratório para as práticas de ciência aberta e preservação de dados de pesquisa. **PontodeAcesso**, Salvador, v. 11, n. 3, p. 2-16, dez. 2017. Disponível em: https://portalseer.ufba.br/index.php/revistaici/article/view/24945. Acesso em: 17 jul. 2019.

ROCHA, R. P.; CAREGNATO, S.; GABRIEL JUNIOR, R. F. Aspectos de inovação na implantação de um centro de digitalização e gestão de dados de pesquisa. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 23, n. esp., p.

1-15, 2018. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23nespp1. Acesso em: 5 jun. 2019.

SAYÃO, L. F. Interoperabilidade das bibliotecas digitais: o papel dos sistemas de identificadores persistentes - URN, PURL, DOI, Handle System, CrossRef e OpenURL. **TransInformação**, Campinas (SP), v. 19, n. 1, p. 65-82, jan./abr. 2007. Disponível em: http://www.scielo.br/pdf/tinf/v19n1/06.pdf. Acesso em: 2 ago. 2019.

SALES, L. F.; SAYÃO, L. F. A ciência invisível: revelando os dados da cauda longa da pesquisa. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19., 2018, Londrina (PR). **Anais [...]**. Londrina: UEL, 2018. p. 4180-4199. Disponível em: http://enancib.marilia.unesp.br/index.php/XIXENANCIB/xixenancib/paper/view/1538. Acesso em: 2 jul. 2019.

SAYÃO, L. F.; SALES, L. F. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, Londrina (PR), v. 21, n. 2, p. 90-115, maio/ago. 2016a. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27939. Acesso em: 17 maio 2018.

SAYÃO, L. F.; SALES, L. F. Curadoria digital e dados de pesquisa. **AtoZ: Novas Práticas em Informação e Conhecimento**, Curitiba, v. 5, n. 2, p. 67-71, jul./dez. 2016b. Disponível em: https://revistas.ufpr.br/atoz/article/view/49708. Acesso em: 17 maio 2018.

SAYÃO, L. F.; SALES, L. F. Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 8, n. 2, 2013. Disponível em: http://eprints.rclis.org/22562/. Acesso em: 17 maio 2018.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: Comissão Nacional de Energia Nuclear, 2015. Disponível em: http://www.cnen.gov.br/images/CIN/PDFs/GUIA_DE_DADOS_DE_PESQUISA.pdf. Acesso em: 14 nov. 2018.

SILVA, D. M.; PINTO, E. M.; CARVALHO, E. R. S.; PEREIRA, P. R.; LEITE, F. C. L. Comunicação científica sob o espectro da Ciência Aberta: um modelo conceitual contemporâneo. **Reciis: Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, Rio de Janeiro, v. 11, nov. 2017. Suplemento. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1414. Acesso em: 5 jun. 2019.

TARGINO, M. G. Comunicação Científica: uma revisão de seus elementos básicos. **Informação & Sociedade**, João Pessoa, v. 10, n. 2, 2000. Disponível em: http://www.periodicos.ufpb.br/ojs2/index.php/ies/article/view/326. Acesso em: 3 dez. 2018.

TENOPIR, C.; TALJA, S.; HORSTMANN, W.; LATE, E.; HUGHES, D.; POLLOCK, D.; SCHMIDT, B.; BAIRD, L.; SANDUSKY, R. J.; ALLARD, S. Research data services in European academic research libraries. **Liber Quarterly**, v. 27, n. 1, p. 23-44, 2017. Disponível em: https://www.liberquarterly.eu/article/10.18352/lq.10180/. Acesso em: 24 out. 2018.

UK DATA SERVICE. **Research data lifecycle**. [2015?]. Disponível em: https://www.ukdataservice.ac.uk/manage-data/lifecycle. Acesso em: 29 nov. 2018.

VALOIS, E. C.; RAMOS, M. G.; RODRIGUES, N. S. S.; ESTEVÃO, S. N. M. Comunicação científica e usuários: elementos de discussão. **Ciência da Informação**, Brasília, v. 18, n. 1, p. 28-34, jan./jun. 1989. Disponível em: http://revista.ibict.br/ciinf/article/view/320. Acesso em: 20 out. 2018.

VERHAAR, P.; SCHOOTS, F.; SESINK, L.; FREDERIKS, F. Fostering effective data management practices at Leiden University. **Liber Quarterly**, v. 27, n. 1, p. 1-22, 2017. Disponível em: https://www.liberquarterly.eu/articles/10.18352/lq.10185/. Acesso em: 23 jul. 2019.

WILEY, C. Metadata use in research data management. **Bulletin of the Association for Information Science and Technology**, v. 40, n. 6, p. 38-40, ago./set. 2014.

ANEXO A – PGD do AHRC

This DMP, made public with the kind permission of the PI for the research project, represents a real example of a proposal submitted to a research funder. The content of this DMP should not be directly replicated, as a DMP is a document unique to your research proposal. Please take advice from your Faculty IT Manager and/or the Research Data Leeds team (researchdataenquiries@leeds.ac.uk) on your requirements, prior to making a submission to your research funder.

Example AHRC Technical Plan

Virtual Holocaust Memory: from Testimony to Holography

Section 1. Summary of Digital Outputs and Digital Technologies

The digital outputs associated with this bid are:

1. A website built in WordPress and hosted on a University of Leeds server. A basic website has already been built: http://arts.leeds.ac.uk/transnationalholocaustmemory/. Through this Fellowship this site will undergo further development to improve:

- **design** - with a focus on the homepage and the optimal presentation of new video content
- **functionality** - ensuring content in all areas of the site can easily be updated through the WordPress dashboard
- **usability** - creating a responsive site to ensure content is adaptive and viewable on mobile and tablet devices

2. A series of videos (around 15) will be commissioned from Leeds Media Services and uploaded to the Transnational Holocaust Memory YouTube channel and embedded in the WordPress site (see 1).

3. Video conferencing technology such as Skype and Adobe Connect (for which the University holds licenses which it will make available to this project) will be used at a seminar series taking place at the universities of Leeds, Sheffield and York in 2015/16 and at a series of public engagement events at the Johannesburg Holocaust and Genocide Centre in April/May 2016. Presentations made at video conferences will be recorded and added to a Resources section of the website.

Section 2.a. Technical Methodology: Standards and Formats

A series of approximately 15 videos along with 6-8 recorded video conference presentations will be in MP4 video format which is optimised for YouTube. The videos will be between 5 and 30 minutes long, ranging from approximately 50 to 500 MB in size. The video conference presentations will be approximately 30 minutes long. The length will vary according to the nature of the content (e.g. short vlog or longer video interview). The videos will be processed by Leeds Media Services. The total volume of the 15 videos is likely to be in the region of 3.75 GB. The recorded video conference presentations are likely to be in the region of 3.5 GB.

Section 2.b. Technical Methodology: Hardware and Software

The website is built in WordPress.

All video filming, processing and editing associated with this project will be done by Leeds Media Services, a trusted and respected local corporate filming company and approved supplier for the University who have the necessary hardware and software.

Section 2.c. Technical Methodology: Data Acquisition, Processing, Analysis and Use

WordPress and YouTube are highly compatible and easily interoperable.

Leeds Media Services will be responsible for filming, processing and editing all the video content. They will supply us with the edited films 14 days after filming which the PI or research assistant will upload onto YouTube and save in line with the protocol outlined below.

The mp4 files will be stored free of charge on a new University of Leeds media server called MediaStore which is currently in development and due to be launched in January 2015. Backup copies will be stored on the University of Leeds SAN (Storage Area Network) which is presented to the user as the M: and N: drives. The MediaStore, SAN and associated server infrastructure are configured for resilience and data protection to a high standard, including daily backup. Copies of raw and processed footage will also be kept by Leeds Media Services.

Permission will be obtained from all external interviewees before filming to enable sharing of all videos.

Section 3. Technical Support and Relevant Experience

It has been agreed that Steve Honeyman from the University's Blended Learning Educational Support Team will be responsible for the development of the project website. Steve has previous experience as a freelance web-designer and developer, with clients ranging from SMEs to the NHS and assorted design agencies. He has longstanding experience of design and development projects involving research, client meetings, design, development, testing, deployment and delivery of the completed site. Other responsibilities have included coding Photoshop files and creating animation. His core skills are advanced HTML, CSS (including HTML5 and CSS3), Responsive Design and CMS integration (usually using WordPress and occasionally using Perch), all of which he has been doing since 2000. He has an MA in Interactive Multimedia (2004) which focussed on design for the web, interaction and aesthetics. In addition, the Blended Learning Educational Support team can provide backup technical support for WordPress.

Leeds Media Services (http://www.leedsmediaservices.co.uk/) will be responsible for all filming associated with this project. They have a long history of collaboration with the PI and the University to produce high-quality media content. See, for example, the videos for the Art of Risk project: http://arts.leeds.ac.uk/artofrisk/ (please note that the actual website is also being redeveloped by Honeyman to optimise design and usability). Leeds Media Services is staffed by an award winning team with 40 years combined experience in video production.

They use the latest digital technology and can film and deliver on any video format, DVD or streaming video file. In 2010 they had success with Ethics in Computing: Real Ethics and Virtual Reality when it was nominated in the British Universities Film and Video Council (BUFVC) Learning on Screen Awards for the Courseware and Curriculum category. In 2012 the University of Leeds Flying Start resource was awarded the Special Jury Prize 'in recognition of the excellence of the underlying design' at the 2012 MEDEA Awards. Leeds Media Services provided all the video assets.

The PI and research assistant will be responsible for uploading content to the website and saving files in line with the University's best practice (see 2.c.) In his current role at the University of Leeds the PI has acted in a similar capacity for numerous projects, including the Art of Risk. The PI has extensive experience of managing web development projects. In his previous role as the Research Strategy Manager at the University of Salford he was responsible for managing the redevelopment of all the University's Research Institute websites, which included overseeing the development of a bespoke Content Management System by a Salford-based company called ED Interactive.

Research Data Leeds (a University of Leeds central support team) have provided guidance for this document and will provide advice and support for the implementation of this technical plan on an ongoing basis throughout the project and after.

Section 4: Preservation, Sustainability and Use

The Transnational Holocaust Memory website serves as a hub for a number of interlinked projects about Holocaust memory at the University of Leeds (currently 8 funded projects linked to 3 different PIs) and the site, along with the content developed through this project, will continue to be used for as long as some or all of these PIs remain at the University of Leeds and active in this broad research area. The lifespan of the site and its content is therefore estimated to be a minimum period of 5 years. Transnational Holocaust Memory is currently a burgeoning area of research at the University of Leeds and there is every likelihood that this site and new content produced through this Fellowship will remain relevant and accessible to the public and researchers for a decade or more.

In terms of the overall sustainability of the website, a theme will be created and coded from scratch so that it will be future-proof and maintainable should the current developer leave. There will be no reliance on external widgets or plug-ins to create layout and manage content; rather, we will have the simplicity of working with the WordPress dashboard.

Section 4.a. Preserving Your Data

All video files will be stored safely in the Leeds MediaStore and SAN for a minimum of five years at no cost. The PI has full discretion over the length of time that the website and its associated content are preserved on the University server and storage system.

Section 4.b. Ensuring Continued Accessibility and Use of Your Digital Outputs

All video files will be tagged with relevant metadata in both YouTube and on the relevant WordPress page. The videos will remain on YouTube and the project website for a minimum of 5 years. It is anticipated that the website will remain an important online archive and research resource for many years beyond that. This project will lead to a substantial upgrade of the design and content of the existing site but as new research projects are initiated the technical and aesthetic dimensions of the site will be reviewed to ensure that they are fit for purpose and new developments will be factored into new research bids accordingly.

A note on this plan

The Peer Review comments from the AHRC on this plan asked for further clarification about how digital outputs from the project would be maintained and preserved. The Research Data Leeds team submitted a letter to the AHRC with further details of the Research Data Leeds repository, including a commitment to keep the deposited data for a minimum of 10 years after the end of the project. This satisfied the AHRC reviewers.

This example DMP illustrates the importance the AHRC places on long term access to and preservation of digital outputs.

ANEXO B – PGD do BBSRC

DMP title

Project Name Drosophila Genetics - BBSRC Example

Description This project will investigate the role of Polo kinase in metaphase to anaphase transition in Drosophila melanogaster.

Funder Biotechnology and Biological Sciences Research Council

Institution University of Glasgow

Data areas and data types

Outline the volume, type and content of data that will be generated e.g. experimental measurements, models, records and images

This project will generate three main types of raw data.

1. Images from transmitted-light microscopy of giemsa-stained squashed larval brains.
2. Images from confocal microscopy of immunostained whole-mounted larval brains.
3. Western blot data.

Measurements and quantification of the images will then be recorded in spreadsheets.

Micrograph data is expected to total between 100GB and 1TB over the course of the project.

Scanned images of western blots are expected to total around 1GB over the course of the project.

Other derived data (measurements and quantifications) are not expected to exceed 10MB.

Standards and metadata

Outline the standards and methodologies that will be adopted for data collection and management, and why these have been selected

All samples on which data are collected will be prepared according to published standard protocols in the field. All microscopes used for sample examination are serviced and recalibrated regularly. All Drosophila lines used in experiments are checked periodically for phenotypic markers. Drosophila are maintained in live culture according to standard methods in the field.

Files will be named according to a pre-agreed convention. The dataset will be accompanied by a README file which will describe the directory hierarchy and filenaming convention.

Each directory will contain an INFO.txt file describing the experimental protocol used in that experiment. It will also record any deviations from the protocol and other useful contextual information.

Microscope images capture and store a range of metadata (field size, magnification, lens phase, zoom, gain, pinhole diameter etc) with each image.

This should allow the data to be understood by other members of our research group and add contextual value to the dataset should it be reused in the future.

Relationship to other data

State the relationship to other data available in public repositories

This dataset will provide a novel characterisation of Drosophila Polo kinase mutants documented in the Flybase database. To the best of my knowledge, no other study has perturbed the metaphase to anaphase transition in these mutants, then examined the

phenotypes seen in mitosis.

Secondary Use

Outline the further intended and/or foreseeable research uses for the completed dataset(s)

The confocal and transmitted light images generated in this work may well be of use in the future. It is entirely possible that another study would want to measure a different aspect of mitosis in Drosophila (both the wild-type controls and the mutants) treated as per the protocols in this study.

I cannot see the western blot data being of future use.

Methods for data sharing

Outline the planned mechanisms for making these data available, e.g. through deposition in existing public databases or on request, including access mechanisms where appropriate

Datasets from this work which underpin a publication will be deposited in Enlighten: Research Data, the University of Glasgow's institutional data repository, and made public at the time of publication. Data in the repository will be stored in accordance with funder and University data policies. Files deposited in Enlighten: Research Data will be given a Digital Object Identifier (DOI) and the associated metadata will be listed in the University of Glasgow Research Data Registry and the DataCite metadata store. The retention schedule for data in Enlighten: Research Data will be 10 years from date of deposition in the first instance, with extensions applied to datasets which are subsequently accessed. This complies with both University of Glasgow guidance and funder policies.

Enlighten: Research Data is backed by commercial digital storage with is audited on a twice-yearly basis for compliance with the ISO27001 Information Security Management standard.

The DOI issued to datasets in the repository can be included as part of a data citation in publications, allowing the datasets underpinning a publication to be identified and accessed. DOIs will also be linked with appropriate records in Enlighten: Publications, the University's publication repository, to enhance visibility of datasets.

Metadata about datasets held in the University Registry will be publicly searchable and discoverable and will indicate how and on what terms the dataset can be accessed.

Information about datasets from the Registry will be displayed on researcher profile pages on the University of Glasgow webpages which will also increase the visibility of the datasets.

Proprietary data

Outline any restrictions on data sharing due to the need to protect proprietary or patentable data

It is not anticipated that this study will generate any patentable data or proprietary data which would have to be protected.

Timeframes

State the timescales for public release of data

Data will be made available at the point of publication of the associated paper or publication.

Formats

State the format of the final dataset

Images will be stored as .tif

Data in spreadsheets will be stored as .csv

Data in freetext documents will be stored as .txt.

These formats are platform agnostic and should support future access and reuse.

Any data which has to be stored in a proprietary format will have the necessary software (including version number) noted in the associated INFO.txt file.

ANEXO C – PGD do EPSRC

DMP title

Project Name Synthetic Chemistry / EPSRC example

Description This research project involves the development of a new chemical reaction for incorporating [your atom of choice] into [your molecule / compound of interest]. An experimental procedure will be developed that will allow the preparation of a range of compounds.

Funder Engineering and Physical Sciences Research Council

Institution University of Glasgow

Data Collection

What data will you collect or create?

The data produced from this work will fall into two categories:

1. The various reaction parameters required for optimisation of the chemical transformation.

2. The spectroscopic and general characterisation data of all compounds produced during the work.

I anticipate that the data produced in category 1 will amount to approximately Z MB and the data produced in category 2 will be in the range of X - Y GB.

How will the data be collected or created?

The reaction conditions will be recorded and collated using Excel spreadsheets and named according to each generation of reaction.

The various experimental procedures and associated compound characterisation will be written up using the Royal Society of Chemistry standard formatting in a Word document. The associated NMR spectra will be collated in chronological order in a .pdf document.

These are standard practices for synthetic methodology projects.

Documentation and Metadata

What documentation and metadata will accompany the data?

The data will be accompanied by the following contextual documentation, according to standard practice for synthetic methodology projects:

1. spreadsheet documents which detail the reaction conditions.

2. text files which detail the experimental procedures and compound characterisation.

Files and folders will be named according to a pre-agreed convention.

The final dataset as deposited in the institutional data repository will also be accompanied by a README file listing the contents of the other files and outlining the file-naming convention used.

Ethics and Legal Compliance

How will you manage any ethical issues?

There are no ethical issues in the generation of results from a synthetic methodology project. There are no human subject or samples involved.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

This project is being carried out in collaboration with an industrial partner. The intellectual property rights are set out in the collaboration agreement. The intellectual property generated from this project will be fully exploited with help from the University of Glasgow's IP and Commercialisation Office.

The aim is to patent the final procedure and then publish the work in a research journal.

Storage and Backup

How will the data be stored and backed up during the research?

The data will be stored on hard-drives belonging to the researchers involved in the work. These hard-drives are backed up onto the School of Chemistry's local servers.

How will you manage access and security?

Files created during this project will be encrypted so that only the PI and researcher will be able to access them. Data will be transferred between the PI and researcher on memory storage devices rather than by email.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Any data from this research which underpin or contribute to our patent application or subsequent research publications will be considered to be of long-term value and will be retained and preserved. These data would be suitable for sharing only once the intellectual property is protected by a patent.

What is the long-term preservation plan for the dataset?

Data which underpin our patent application and research publications will be stored on the School of Chemistry's server.

The dataset will also be deposited in Enlighten: Research Data, the University of Glasgow's institutional data repository. Data in the repository will be stored in accordance with funder and University data policies. Files deposited in Enlighten: Research Data will be given a Digital Object Identifier (DOI) and the associated metadata will be listed in the University of Glasgow Research Data Registry and the DataCite metadata store. The retention schedule for data in Enlighten: Research Data will be 10 years from date of deposition in the first instance, with extensions applied to datasets which are subsequently accessed. This complies with both University of Glasgow guidance and EPSRC policy.
Enlighten: Research Data is backed by commercial digital storage with is audited on a twice-yearly basis for compliance with the ISO27001 Information Security Management standard.

Data Sharing

How will you share the data?

If the research is sucessful, the research will be protected by the filing of a patent. Following this, the research will be diseminated by the publication of an open-access manuscript in a chemical journal. The manuscript will be deposited in our institutional publication repository, Enlighten: Publications. The manuscript will contain a data citation indicating where and on what terms the data can be accessed.

The data which underpins the publication and patent will be made available for sharing via Enlighten: Research Data, the University of Glasgow's Data Repository. This will be available at the time of publication of the corresponding manuscript.

Data in the repository will be issued with a Digital Object Identifier (DOI). This can be included as part of a data citation in publications, allowing the datasets underpinning a publication to be identified and accessed. DOIs will also be linked with appropriate records in Enlighten: Publications, the University's publication repository, to enhance visibility of datasets.

Metadata about datasets held in the Institutional repository will be publicly searchable and discoverable and will indicate how and on what terms the dataset can be accessed.

Information about datasets from the repository will be displayed on researcher profile pages on the University of Glasgow webpages which will also increase the visibility of the datasets.

Are any restrictions on data sharing required?

After the patent is filed to protect the intellectual property, there will be no restrictions on data sharing. The PI will actively disseminate the results as widely as possible.

Responsibilities and Resources

Who will be responsible for data management?

The PI will be ultimately responsible for data management of this project. The researcher will be responsible for organisation and storage of the data as it is produced during the project. The School of Chemistry's IT staff will manage teh school's server, where the data will be stored. The University's library staff will be reponsible for management of the institutional repositories, Enlighten: Publications and Enlighten: Research Data.

What resources will you require to deliver your plan?

The researcher already has the required software to implement the data collection plan. Funds to cover final deposit of the dataset in the institutional repository have been costed into the grant application as advised by the Research Data Management Service. Funds to support open access publication of the research are available from the institutional RCUK fund for this purpose.

ANEXO D – PGD do ESRC

This DMP, made public with the kind permission of the PI Andrea Holomotz, represents a real example of a funded proposal from the University of Leeds that forms part of a wider suite of documentation submitted to the research funder. The contents of this DMP should not be directly replicated as a DMP is a document unique to your research proposal. Please take advice from your Faculty IT Manager and/or the Research Data Leeds Team (researchdataenquiries@leeds.ac.uk) on your requirements prior to making a DMP submission to your research funder.

Realist Evaluation of Adapted Sex Offender Treatment Programs for Men with Intellectual Disability

Project Stage: Basic
RCUK Research Councils: Economic and Social Research Council
Lead organisation: University of Leeds

1	**An explanation of the existing data sources that will be used by the research project (with references).**
	There is some research that evaluates Adapted Sex Offender Treatment Programs (ASOTPs), but not necessarily by drawing on staff and especially user perspectives. A review of existing datasets, including searches on the UK Data Services website, found no published datasets on sex offenders with intellectual disability (ID) and/ or treatment programs. The British Crime Survey can contextualise the research to some extent by exposing the prevalence of sexually violent crimes, but this is not disaggregated by whether or not the offender had an ID.
2	**An analysis of the gaps identified between the currently available and required data for the research.**
	ASOTPs have not previously been examined in the manner proposed by this research and there are no existing datasets in the archives which are adequate for answering the proposed research questions. There are no available datasets, which evaluate ASOTPs.
3	**Information on the data that will be produced by the research project**
3.1	**Data volume and data type, e.g. qualitative or quantitative data**
	Qualitative data will be generated in audio format from interviews with practitioners (n=24 (phase 1&3) + up to 20 (phase 2)). There will be interviews (up to 20) and focus groups (4) with sex offenders with ID, which will be fully transcribed and anonymised. Fieldnotes of focus groups and also any drawing created by respondents with ID in focus groups and interviews will also be kept. This data will be anonymised and then processed and analysed using NVivo9. Quantitative data will be extracted from approximately 80 patient files and each case will be anonymised and then processed and analysed using SPSS.
3.2	**Data quality, formats, standards documentation and metadata**
	Audio files will be stored in MP3 or WAV format. Digital images will be stored as JPEGs. (Note that neither of these will be made available for data sharing, see section 7.) Microsoft Word 2007/2010 will be used for text based documents. .sav will be used for SPSS files. These file formats have been chosen because they are accepted standards and in widespread use. At the end of the project, the Word documents will be converted to both plain text and PDF-A and long term preservation of the data from statistical analysis packages such as Stata will be carried out in accordance with the advice from the Council of European Social Science Data Archives (http://www.cessda.org/project/doc/D10.4_Data_Formats.pdf section 4.3 pp33). I am committed to providing high standard quality data and research excellence. To ensure the integrity and quality of the research data and increase the potential for data sharing, the transcriptions of the audio files will be checked and anonymised to make them ready for archiving. The formatting of data and the provision of metadata will conform to the UKDA standards and guidelines. This will also include clear data description, annotation, contextual information and documentation, e.g. unique identifier for each transcript, uniform and consistent

	layout throughout data collection, cover sheet with interview details such as date, place and interviewee details. (Note that details will be kept vague to preserve anonymity, see section 7.)
3.3	**Methodologies for data collection**
	Semi-structured interviews and focus groups will be digitally recorded and subsequently transcribed into Microsoft Office Word 2010/2007. Fieldnotes on focus groups will be kept in Microsoft Office Word 2010/2007. Quantitative data will be extracted from patient files and collated into SPSS. A consistent system of file naming and an organised folder structure will ensure easy retrieval. This will involve creating meaningful but brief names and using file names to classify types of files.
4	**Planned quality assurance and back-up procedures (security/storage)**
	Electronic data will be stored on the University of Leeds SAN (Storage Area Network), which comprises enterprise level file servers in physically secure data centres with appropriate fire suppression equipment. Snapshots are taken every day at 10pm (and accessible for 1 month). A second level of snapshots is taken every month and are kept for 11 months. Snaphots are user recoverable from the desktop. An incremental copy to backup tape is taken every night (and kept for 28 days) and a full copy is taken every month. Every quarter, the full dump tapes are moved to a long term storage facility where they are kept for 12 months. Tapes are initially stored in on-campus fireproof safes and then moved to off-campus secure locations. Access to electronic data is controlled by Active Directory (AD) Group membership. The Faculty IT Manager will set up a dedicated folder for this research project and create read-only and read-write AD groups. I will decide which users require read-only and read-write access. Off-campus access is via the Citrix portal. External users who need access to the data will apply for a University username and then be assigned to the appropriate AD group. Any sensitive data (as defined by the Data Protection Act) that is stored on portable electronic devices will be protected by encryption software to FIPS 140-2 standard. Any sensitive data that needs to be transmitted electronically will first be encrypted to FIPS 140-2. If any highly sensitive data needs to be stored, then a research data folder on the SAN will be encrypted, so it can only be accessed by authorised members of the project with the appropriate encryption software installed on their desktop PCs. Highly sensitive data is not available from off-campus.
5	**Plans for management and archiving of collected data**
	As required by ESRC, this data management plan seeks to prepare the project data for future sharing and potential secondary analysis. This is particularly important in this project, as sex offenders with ID have thus far often been the objects and rarely the subjects of research. Therefore, the data (including anonymised transcripts of interviews where I have explicit permission for these to be used, but not including any audio or visual files) will be deposited for archiving and re-use with the ESRC data service provider, UKDA, at the end of the project and within three months of the end of the award. The data management plan will be reviewed during the life of the project to ensure the success of the long-term strategy. Prior to archiving, the data files will be converted to suitable open formats for long term preservation as described in section 3.2.
6	**Expected difficulties in data sharing, along with causes and possible measures to overcome these difficulties.**
	As much of the data is generated through interviews with human participants, the ability to make it available for reuse will be subject to receiving the necessary level of consent from the individuals involved. Due to the highly sensitive nature of this research it is essential that the identities of participants remain concealed. However, ethically it is preferable to re-use data on

	vulnerable populations, such as sex offenders with ID, rather than having to re-visit these for each individual research project. I will therefore keep a close eye on the data that is being produced during this research and seek to find ways of making it available for other researchers. One way could be to hand-pick a number of sample transcripts in partnership with respondents and edit these further for data sharing. I have consulted with UK Data Service on possible strategies as part of my early data management planning. Appropriate restrictions to data sharing offered by the UK data archive will be put in place. I will furthermore attend a "Managing Sensitive Data in an Open Access Age" workshop and any other relevant training run by the UK data archive to find out more.
7	**Explicit mention of consent, confidentiality, anonymisation and other ethical considerations**
	Fully anonymised data arising from interviews and focus groups will not be shared, unless explicit consent was given by respondents. As sex offenders with intellectual disability (ID) are such a vulnerable group, particular care will be taken. Only transcripts that were especially prepared for this will be shared. As discussed in section 6, one way of handling this sensitive data could be to hand-pick a number of sample transcripts in partnership with respondents and edit these further, to remove any identifies, such as very peculiar phrases used by a person that makes them easy to identify or in fact mis-identify or lengthy descriptions of events or personal stories that may give a person away. Focus group data may not be shared, unless all participants consented to data sharing. However, editing may be applied to remove respondents who did not consent to data sharing from a focus group transcript, to allow at least some of the data to be shared. Enabling people with ID to give their informed consent to something as complex as data sharing poses considerable challenges. I have experience of working with this group professionally and have also conducted past research with this group. I am able to break down complex concepts to this population and will put considerable effort into developing an accessible way to explain “data sharing”. Consent for data sharing will be sought from the respondent in the presence of their key worker, who will sign the consent form as a witness. If they are in doubt about the respondent having understood “data sharing”, the key worker will make a decision about this on their behalf, unless the respondent opted out. In that case their choice will be respected, even if they do not appear to have understood “data sharing” fully. NHS and “Kantonale Ethikkommission” (Switzerland) approval will be sought regarding the sharing of fully anonymised and quantified data extracted from patient files in .sav format.
8	**Copyright and intellectual property ownership of the data**
	The intellectual property of the data generated will remain with the University of Leeds. However, the University policy of the management of research data requires all data arising from research projects to be made openly available where possible. The research will not use any data which is covered by the Copyright, Designs and Patents Act 1988 or any other similar legislation.
9	**Responsibilities for data management and curation within research teams at all participating institutions**
	I will have overall responsibility for implementing the data management plan. The Faculty IT Manager will be responsible for ensuring that electronic file permissions have been correctly assigned and for advising on other aspects of data storage and security. Staff involved in the project at participating organisations will be responsible for following data management procedures. The data management plan will be monitored in meetings with my experienced mentor and during RA supervision.

ANEXO E – PGD do H2020

European Commission | Horizon 2020 European Union Funding for Research & Innovation

AutoPost

Deformable Surface Tracking and Alpha Matting for the Automation of Post-production Workflows

D6.3: Data management plan

Project ref. no.	H2020-ICT-18-2014 GA-644628
Project Acronym	AUTOPOST
Start date of project (dur.)	1 January 2015 (18 months)
Document due Date:	30/06/2015 (M6)
Actual date of delivery	31/07/2015 (M7)
Leader of this deliverable	Eurecat (EUT)
Reply to	monica.caballero@eurecat.org
Document status	Final ready for submission

Version	Date	Description
1	03/06/2015	Draft version circulated to partners
2	26/07/2015	Revised version with updates
3	31/07/2015	Final version

Deliverable Identification Sheet

Project ref. no.	H2020-ICT-18-2014 GA-644628
Project acronym	AUTOPOST
Project full title	Deformable Surface Tracking and Alpha Matting for the Automation of Postproduction Workflows
Document name	Autopost_D6 3_20150630
Security (dissemination level)	PU
Contractual date of delivery	Month 6, 30.06.2015
Actual date of delivery	Month 7, 31.07.2015
Deliverable number	D6.3
Deliverable name	Data management plan
Type	RE
Status & version	Final, v7
Number of pages	13
WP / Task responsible	Eurecat (EUT)
Author(s)	Monica Caballero, Mº Eugenia Fuenmayor
Other contributors	All partners
Project Officer	Philippe Gelin
Abstract	This deliverable describes all the data that will be collected and generated during the AutoPost project, how it will be created, stored and backed-up, who owns it and who is responsible for the different data and which data will be preserved and shared according to the participation of the project in the Open Research Data Pilot.
Keywords	DMP, Data management plan, Open Research Data Pilot
Sent to peer reviewer	26/07/2015
Peer review completed	30/07/2015
Circulated to partners	30/07/2015
Read by partners	Via project's repository
Mgt. Board approval	pending

Table of contents

Executive summary **4**
1. Data collection **5**
1.1 AutoPost collected and/or created data 5
1.2 Standards and methodologies 6
2. Storage and backup **8**
2.1 Data storage and back up during the research 8
2.2 Selection and preservation 9
2.3 Data sharing 10
2.4 Responsibilities and resources 11
3. Ethics and legal compliance **11**
3.1 Ethical issues 11
3.2 Copyright and Intellectual Property Rights (IPR) issues 11

Executive summary

The AUTOPOST project participates in H2020 pilot action on open access to research data. This deliverable, the AUTOPOST data management plan, describes the research data that will be collected and generated during the project and explains how it will be exploited or if it will be shared for verification and re-use.

AUTOPOST is an industry-driven innovation action that will deliver ICT-based solutions to enhance established post-production workflows. As most of the project outcomes are susceptible of being protected for exploitation, this data management plan will clearly identify which data will be kept confidential and which will be made openly available.

This document describes the data, how it will be created, how it will be stored and backed-up, who owns it and who is responsible for the different data.

The AutoPost management Plan will be updated as the project progresses.

1. Data collection

1.1 AutoPost collected and/or created data

AutoPost will both collect existing data from partners and third parties, and will create new data within the project. AutoPost will collect and produce five broad categories of data: data for evaluation, computer software, research data and metadata, manuscripts and dissemination material. In the following we describe the types of data and the formats used.

A complete list of all data to be collected and created is shown in Table 1. The additional information about each item will explained in the following sections.

Data for evaluation

This data will consist mainly of image and video datasets to be used as a test data for development and evaluation of the AutoPost tools, as well as project files for post-production platforms such as Nuke. This data is to be used within the Consortium throughout the duration of the project.

AutoPost will take advantage of already existing data that can be used in the project. Datasets will include existing material collected by partners in the Consortium and existing public images and video matting and tracking test datasets.

AutoPost will also acquire new footage and will generate processed video datasets based on that: edited video test data to be used as input for the VFX tools and the output obtained after its usage, as well as a subset of the acquired data to be shared with the research community.

Image and video datasets will use common file formats. Images will be JPG and PNG files, while video datasets will use DPX (Digital Picture Exchange), a format used in the industry for digital intermediate and visual effects work, and multimedia container files such as MOV or MXF. Nuke[1] project files will use the native NK format of this platform.

Computer software

AutoPost will produce different kinds of software: tracking and matting libraries in the form of SDKs, tracking and matting plugins for post-production platforms, and source code for both libraries and plugins.

Plugins and libraries in binary form will be stored and shared as ZIP archives which are commonly used for this purpose. Source code will use standard programming language format files such as C or CPP, depending on the chosen language for the development.

Research data and metadata

This category comprehends on the one hand, data generated by user interaction with the AutoPost tools such as uv maps, shading maps, region maps for tracking and propagated trimaps, alpha mattes or composite foregrounds for matting. All this data will be stored and shared as images which will use the open standards PNG and EXR.

On the other hand, research data will also consist on the feedback gathered during the project developments: the bug logs and feedback generated by partners when integrating SDKs into plugins, and the feedback and opinions (in the form of questionnaires or interviews) of end-users when using the AutoPost tools. For the former, standard file formats such as log files from ticketing platforms will be used. Regarding the latter, opinions will be collected in text files using DOC, PDF or TXT formats depending on its stage, purpose and audience.

[1] https://www.thefoundry.co.uk/products/nuke/

Manuscripts

Manuscripts will consist of all the reports generated during the project, including the description of matting and tracking algorithms, all deliverables, publications and internal documents. Microsoft Word (DOCX) and PDF will be used for final versions, while intermediate versions can consider the usage of ODT or TEX (LateX) files.

Dissemination material

AutoPost will produce dissemination material in a diversity of forms: flyers, public presentations, videos demonstrating the performance of algorithms in SDKs and plugins, and a short film produced with the AutoPost tools based on the newly acquired footage.

For video dissemination data, widely used video file formats for distribution, such as MOV or AVI will be used. All other dissemination material will be shared in PDF format.

1.2 Standards and methodologies

Collection and creation of data

In the following, details on the collection or creation of the data of the different categories/types will be provided:

- **Data for evaluation.** Already recorded audiovisual material and pieces of commercial films will be provided by AutoPost partners, as well as public datasets such as hollywoodcamerawork.com, selecting those shoots where matting and tracking tasks were especially costly to carry out, taking into account the use cases defined in D2.1. Existing public research-purpose images and video matting and tracking datasets will be downloaded by the RTD partners (HHI, EC) for validating the coverage and performance of their algorithms from the corresponding websites: alphamatting.com and video.matting.com.
  Regarding new footage, the consortium will acquire new video data for testing purposes. The scenes of the shooting will be designed to cover interesting instances of the use cases defined. The shooting will be organized following the standard procedures (generation of the script, storyboard, selection of actors, scheduling, etc...), where all partners will have different responsibilities and will assist the shooting.
- **Software.** SDKs and plugins will be provided by the RTD partners (HHI, EC, IL) as soon as new versions are released. RTD partners will implement SDKs and plugins following a user centered-approach involving end user partners in all stages of the project. Developments will be based on use cases defined by the users and will follow an interactive and iterative process, where prototypes and early versions of both SDKs and plugins will be produced for integration tests and user tests, respectively.
- **Research data and metadata** will be generated during the R&D process by the RTD partners (HHI and EC) as intermediate results of the application of algorithms (uv maps, mattes). Feedback regarding the software development will be generated by the RTD partners (IL, HHI and EC) when carrying out the integration of SDKs into the plugins by means of generating tickets, logs and documenting bugs in a dedicated software platform. Finally, the feedback from end-users will be obtained through questionnaires and interviews with end-users who will test the AutoPost tools.
- **Manuscripts and dissemination material** will be produced using the Microsoft Office Suite. Whenever possible and needed, collaborative workflows and tools will be used.

Structure, name and versioning of files

Regarding the structure, all data will be stored using a folder structure following WPs and tasks organization whenever possible and depending on the chosen storage system (see next section). Thus, every file will be stored under its corresponding WP and task folder. There will be cases

when task folders will not make sense (e.g. WP4 tasks). In that cases, WP folders and folders differentiating between matting and tracking will be used.

In the case of video data, the project will keep the raw data separate from the processed data using different folders, separating input from output data. SDKs and plugins will be stored in different folders and different folder structures will be used for matting and tracking versions. Source code, as well as research data from the matting and tracking algorithms will follow partner's local folder structure conventions, since this data will be solely managed by one partner of the Consortium during its generation (some will be shared afterwards).

Processed data files, SDKs, plugins and research data will be accompanied by a readme file including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed. Documentation will also include details on the methodology used, analytical and procedural information, any assumptions made, and the format and file type of the data. In the case of software it may also include installation instructions and usage examples. All this information will be inside the manuscripts as well, unless structure of the document inhibits it (e.g. a journal/conference paper).

As to the naming of files, in all cases, files will be named according to their content to ease their identification. Versioning of the files will be handled by specifying its version after the filename "filename_vx". In the case of processed data, SDKs and plugin versions, the name of each version will have the date it was created next to the original title (filename_date). During development of software data version control will be done through specific software versioning and revision control systems (SCM) such as svn[2] or git[3].

In the case of manuscripts, the owner of the document will be the one controlling the version of the document, while files created by partners adding contributions to the original will be named by attaching "_initials" to the filename. Other aspects concerning document and version numbering of reports and deliverables are described in the AutoPost Handbook and Quality Plan (D1.1).

Quality procedures

The Consortium has set up quality procedures for internal documents, deliverables and software. Publications are not considered in the procedure as they already go through an external refereed process.

The quality control process for internal documents and deliverables is described in the AutoPost Handbook and Quality Plan (D1.1).

Images and videos to be used and those acquired in the project will go through a natural quality control by the RTD partners as they will monitor that a minimum quality requisites are obtained in the shootings to be able to run their algorithms. Quality of images and videos produced during the project will be assessed by end-user partners which will control the obtained material is compliant with standards in the industry.

In the case of the software produced the quality is guaranteed by several means:

1. Continuous integration performed by partners
2. Plugins and SDK tests
3. Integration tests
4. User tests

[2] *https://subversion.apache.org/*

[3] *https://git-scm.com/*

2. Storage and backup

2.1 Data storage and back up during the research

Storage and maintenance of AutoPost data will be handled according to the data category, privacy level, need to be shared among the consortium, and its size. This section covers the storage selections for data independently of if the data is to be shared externally. For that purpose, specific storage systems allowing public access will be selected. These will be detailed in Section "2.3.Data sharing".

Software data (except source code) will be stored on a Redmine[4] server hosted at Eurecat. Redmine is an open-source project management web application offering multiple project support, version control (svn and git), issue tracking, files management, activity feeds, wiki and forums. Allowing installation on a partner's server is an important feature as it is a project requisite for internal sharing of software. Source code will be archived locally at partner's servers because of its privacy level. For this purpose, local version control and software development platforms at partner's servers will be used. All software data will be backed up in a daily basis. The Redmine server, as well as partner's servers will provide the means for that.

All electronic data generated during research activities (tracking and matting results) will be also redundantly stored locally at partners' workstations and servers. Locally, research partners have secure servers on which all information will be stored. The server drives are backed up periodically. A back-up copy once these results are generated or changed is considered sufficient for that type of data.

The AutoPost Consortium has chosen the open source self-hosted file sharing platform owncloud[5], which is hosted by Eurecat, to be the official repository of non-software data, containing data meant to be shared (created and/or generated) among the Consortium which size is smaller than approximately 2GB. Dissemination material, reports and deliverables, and data such as Nuke project environments and feedback from plugin evaluation by end-users will be stored in owncloud. The owncloud server is backed-up periodically, ensuring needed back-up frequencies required for this data (see Table 1).

Image and video datasets are the remaining data typically exceeding 2GB. These datasets will be stored using different means. Existing datasets provided and collected by partners for internal use (software validation) will be stored in their own server to ensure its privacy if needed. Designated shots will be transferred to other partners on request via external hard drive. Datasets acquired in the project will be saved primary by IL, responsible for the storage of the raw data. Moto and DG, who will edit raw footage to create the data to be used as input for VFX and will generate the processed data using the AutoPost tools will save the new datasets. All datasets will be available for all partners through external hard drives to be sent to all partners.

Maintenance of datasets stored in partners' servers will be carried out according to the partners' backup policy. Backup of publicly shared datasets is considered unnecessary. The back-up of the newly acquired dataset will be done by IL and as final responsible by project coordinator, Eurecat.

[4] *www.redmine.org/*

[5] https://owncloud.org

2.2 Selection and preservation

2.2.1 Data to be retained, shared, and/or preserved

The Consortium has identified some data that may be retained by project partners in accordance to the grant and consortium agreement

- The final versions of the SDKs may be retained by IL for further plugin developments and commercialization under a licensing agreement. A royalty-free license for the SDKs will be granted to end-user partners (DG, MOT) for their own use as part of the final plugins. All previous versions of the SDKs will be destroyed at the end of the project from each computer at each partner that was using it.
- A limited number of the final plugins may be retained by partners for their own use during the project and also for commercial use after the project, allowing SMEs (DG, MOT) to use them for their own post-production work.
- Already existing image and video datasets provided by partners DG, MOT and HHI should only be preserved for joint commercialization purposes. Otherwise they must be destroyed at the end of the project due to privacy issues.

The Consortium has also identified some of the data to be preserved since they can be further used by partners and because this data can be of interest of the research community for different reasons.

Image and video dataset acquired during the project, including raw and processed versions will be kept by all partners since this dataset can be used in other projects beyond AutoPost for validation purposes. A subset containing interesting shots, as well as intermediate research data and research results (to be used as ground truth for example) will be preserved and shared with the research community, so to provide useful data for enabling performance analysis and comparison of matting and tracking algorithms in the field of computer vision, improving the availability of current public datasets and ground-truth.

Preserving this dataset and making it publicly available will require in the first place, the selection of the shots to be shared and the preparation of the research data associated with the selected sequences. Since the video dataset may be quite big, proxies in lower quality and resolution will be created to ease the sharing and allow interested researcher to evaluate their interest before requesting the whole set (see Section 2.3). In a second place, it will also require the maintenance of the data and the management of their access.

Similarly, in order to let other researchers know about AutoPost advances and compare their research, public reports, including public deliverables and open-access papers on journals or conferences will be also preserved and shared whenever possible (when they do not limit future exploitation plans), along with the research data necessary for validating the published results.

All dissemination material produced during the project, including demonstration videos of AutoPost algorithms and plugins, short-films produced using AutoPost tools using acquired footage, and project public presentations, will be preserved and made it public as soon as possible to let the research community know about AutoPost solutions and results in a more graphical way.

For the public reports and dissemination material, no much extra effort is considered for its preservation beyond the act of publishing them in public repositories (see Section 2.3).

It is agreed that this data has to be preserved a minimum of 3 years after the project end.

As for the rest of the data, preservation is not considered necessary, meaning it is already preserved by other institution (public video datasets) or that they do not provide an added value for the research community (internal data generated during the project development). This does not

avoid that partners preserve it for themselves in their archives if they consider it useful for their research and innovation activities.

2.2.2 Preservation plan for the datasets

All data to be preserved or that can be preserved by partners to be used by the Consortium beyond the end of the project will be kept in partner's servers, and maintenance will be carried out by each partner responsible for the data.

Data that will be made public will be held in different repositories, as it will be explained in next section.

2.3 Data sharing

The Consortium is aware of the mandate for open access of publications in the H2020 projects and the participation of the project in the Open Research Data Pilot. The Consortium has chosen ZENODO[6] as the scientific publication and data repository for the project outcomes. The Consortium, through WP6, will ensure that scientific results that will not be protected and can be useful for the research community will be duly and timely deposited in the scientific results repository Zenodo, free of charge to any user. As detailed in previous sections, these will be:

1. Machine-readable electronic copies of the final version or final peer-reviewed manuscript accepted for publication; made available immediately with open access publishing (gold open access) or with a certain delay to get past the embargo period of green open access.
2. Public project deliverables and public summaries of confidential project deliverables.
3. Teasers, flyers, project public presentations and any other kind of dissemination material.
4. Video dataset composed of a selection of shots from the acquired video material in the project.
5. Research data needed to validate the results presented in the deposited publications and associated with the public video dataset acquired during the project. Once there is a collection of data worthy to be shared, a set will be build and shared on Zenodo.

Regarding the video dataset, as the maximum file size allowed in Zenodo is 2GB and the videos will be much larger than that, the Consortium has agreed on using proxies to overcome this limitation for sharing video datasets. The proxies will be a compressed version of the original videos stored on Zenodo. Versions in their original resolution, format and length will be stored primary at IL, and a back-up will be help at coordinator's (Eurecat) repository as explained in section 2.1. Through proxies, interested researchers can look at the contents and choose whether or not they want a complete full-resolution version, in which case they would have to send an on-line petition to the project-appointed data manager/administrative to be authorised to obtain them. This applies during at least 3 years after the duration of the project.

Autopost's shared data is to be shared for research and training purposes only, therefore, requesters will be asked to explain the usage they will give to them (for internal information on post-production-related research activities), and will be asked to sign a dataset license limiting its usage and distribution.

Upon petition, authorized users can download complete video datasets from Eurecat's server or receive them on physical storage provided that they assume storage and shipping expenses.

Dissemination of available data for research will be done through the project's website and at AutoPost's dissemination activities. As stated in the Contract Agreement, Autopost's website will remain active for at least 4 years after the project ends.

[6] http://zenodo.org

Also, external users of AutoPost data will be asked to make a visible acknowledgement to the project adding also the address in Zenodo where they can be viewed and requested.

Datasets for internal project activities are available to all partners. No extra-agreement other than the Consortium Agreement is needed as it covers all the appropriate limitations.

Finally, as explained in previous sections, AutoPost software data (SDKs and plugins) will not be shared with external parties given the exploitation plans derived from them. However, intermediate versions will be made accessible for some selected end-user partners to test them and get feedback. In these cases, strategies such as time-limited licenses and watermarked versions will be used for this purpose.

2.4 Responsibilities and resources

Responsibilities

Eurecat, as coordinator, is responsible for implementing the data management plan (DMP).

In principle, all partners are responsible for data generation, metadata production and data quality. Specific responsibilities are to be assigned depending on the data and the internal organization in the WPs and tasks where data is created. Thus, for example, HHI and EC partners are responsible for the creation of SDKs, IL is responsible for the creation of the plugins, and end users are responsible for the creation of the feedback data after its evaluation and the short demo video produced with AutoPost tools. In the case of the acquisition of new test data, task leader will organize the responsibilities for all the partners which will participate jointly in the shooting.

Dataset storage and backup, data set archiving & sharing will be in the majority of cases the responsibility of the partners who owns the data and/or the servers in which they will be stored. Beyond data to be stored at some of the partner's repositories (mostly video datasets), Eurecat, will be responsible for storage and back-up of computer software data and all data stored on owncloud, since Eurecat hosts both services. Regarding the set of video data to be made public, which will be jointly owned by the Consortium, it will be stored at IL premises, and Eurecat as a coordinator will be also responsible for its back-up.

Resources for delivering the DMP

Extra resources, as physical storage media and redmine and owncloud special features, are needed to accomplish the storage and maintenance activities described above.

3. Ethics and legal compliance

3.1 Ethical issues

AutoPost does not handle personal data except for actors appearing in shoots and pictures, in which case the partners follow the standard procedure of getting authorization from the actors to show and distribute the videos and images to the public.

3.2 Copyright and Intellectual Property Rights (IPR) issues

Table 1 provides the details of the owners of each of the data to be collected and produced by AutoPost project.

As a general principle, for collected data, the owner of the data will remain the same. For produced data, the producer of the data will own the data (e.g. SDKs, plugins, algorithms, reports, etc.) SDK access rights will be granted to IL through licensing, and plugin and SDK access rights to user partners will be granted through free licenses, as stated in Section 2 of AutoPost Grant Agreement.

All data not available for reuse has been identified as to be destroyed at the end of the project. Reuse of other kind of data not to be destroyed at the end of the project by project partners will not require any licensing policy.

Datasets produced within the project as part of the project's testshoot belong to the Consortium (joint ownership), as set out in the Consortium agreement in relation to joint RTD activities. Usage of full-version of shared datasets will be restricted to research-only activities and no distribution to others will be allowed. Users requesting downloading full-version of datasets will be asked to sign a free license agreement (see previous section).

Collected/Created	Title	Description	Category	Type	Format	Size	Owner	Privacy level	Storage / Storage for public access	Back-up frequency	Destroyed at the end of the project?	Duration of preservation (in years)
Created	Matting SDK (Binary)	Binary releases of the matting library	Computer Software	Library (SDK)	ZIP	40MB	EC	Consortium	Eurecat's redmine	Daily	No (1)	0
Created	Tracking SDK (Binary)	Binary releases of the tracking library	Computer Software	Library (SDK)	ZIP	<10MB	HHI	Consortium	Eurecat's redmine	Daily	No (1)	0
Created	Matting Plugin (Binary)	Matting plugin	Computer Software	Plugin	ZIP	100 MB	IL	Consortium	Eurecat's redmine	Daily	No (2)	0
Created	Tracking Plugin (Binary)	Tracking plugin	Computer Software	Plugin	ZIP	100 MB	IL	Consortium	Eurecat's redmine	Daily	No (2)	0
Created	Matting SDK (Source)	Source code of the matting library	Computer Software	Source code (library)	C,CPP		EC	EC	EC	Daily	-	0
Created	Tracking SDK (Source)	Source code of the tracking library	Computer Software	Source code (library)	C,CPP		HHI	HHI	HHI	Daily	-	0
Created	Matting Plugin (Source)	Source code of the matting plugin	Computer Software	Source code (plugin)	C,CPP	200 MB	IL	IL	IL	Daily	-	0
Created	Tracking Plugin (Source)	Source code of the tracking plugin	Computer Software	Source code (plugin)	C,CPP	200 MB	IL	IL	IL	Daily	-	0
Created	Environments/ project files (Nuke)	Project files of Nuke used or to be used for evaluation of the plugins in different VFX tasks	Data for evaluation	File	NK	<1M	Producer of the environment	Consortium	owncloud	Daily	Unnecessary	0
Collected	Existing VFX Dataset (DG)	Collected test data for visual effects from DG.	Data for evaluation	Images	DPX	135GB	DG	Consortium	DG	Once	No (3)	3
Collected	Existing VFX Dataset (MOT)	Collected test data for visual effects from MOT.	Data for evaluation	Images	DPX	10GB	MOT	Consortium	MOT	Once	No (3)	3
Collected	HHI Natalie Facial Expression Database		Data for evaluation	Images	4K	1.2TB	HHI	Consortium	HHI	Once	No (3)	0
Collected	hollywoodcamerawork.com examples		Data for evaluation	Images	JPG	2GB	Hollywood Camera Works	Public	MOT	Once	Unnecessary	0
Collected	alphamatting.com dataset	Public image matting dataset	Data for evaluation	Images	PNG	21MB	alphamatting.com	Public	EC	Once	Unnecessary	0
Collected	videomatting.com dataset	Public video matting dataset	Data for evaluation	Videos	PNG	2.6GB	videomatting.com	Public	EC	Once	Unnecessary	0
Created	New Test Data (VFX Input/Edited)	Edited footage used as input to VFX.	Data for evaluation	Videos	DPX	>10GB	ALL	Consortium	MOT,DG, hard drive at all premises	Change	Unnecessary	0
Created	New Test Data (VFX Output)	Output from the visual effects.	Data for evaluation	Videos	DPX	>10GB	ALL	Consortium	MOT,DG, hard drive at all premises	Change	Unnecessary	0
Created	New Test Data (Acquired/Raw)	Raw footage acquired during the project.	Data for evaluation	Videos	MXF,MOV	3TB	ALL	Consortium	IL, hard drive at all premises	Once	Unnecessary	0
Created	Selected subset of acquired data	Subset of the footage acquired during the project.	Data for evaluation	Videos	DPX	>2GB	ALL	Public	IL, EC, hard drive at all premises / proxy on Zenodo	Once	No	3
Created	Flyers, project public presentations	Flyers, project public presentations	Dissemination material	Documents	PDF	<2GB	ALL	Public	owncloud / Zenodo	Once	No	3
Created	New Test Data (Short Film)	Finished short film based on the new footage.	Dissemination material	Videos	MOV	<2GB	ALL	Public	owncloud / Zenodo	Once	No	3
Created	Demo videos of SDKs/plugins	Videos demonstrating the performance of algorithms in SDKs and plugins	Dissemination material	Videos	MOV, AVI	<2GB	ALL	Public	owncloud / Zenodo	Change	No	3
Created	Internal reports (deliverables)	Internal reports covering project activities	Manuscript	Report	DOC	<20M	Producer of the report	Nature of the deliverable	owncloud / Zenodo	Monthly	No (for shared reports)	0
Created	Public reports (papers, articles, etc..)	Papers describing results from the project	Manuscript	Report	DOC, PDF, .TEX	<20M	Producer of the report	OpenAccess when not exploitable	Partners servers / Zenodo	Monthly	No	3
Created	Tracking results	uv maps, shading maps, region maps	Research data	Image	EXR	<2GB	HHI	Public	HHI/Zenodo	Change	No	3
Created	Matting results	Propagated trimaps, alpha mattes, composite foregrounds	Research data	Images.	PNG	<2GB	EC	Public	EC/Zenodo	Change	No	3
Created	Feedback plugin evaluation	Interviews collecting user experience and feedback	Research data	Report	DOC,PDF,TXT	<10MB	ALL	Consortium	owncloud	Change	Unnecessary	0
Created	Feedback SW development	Feedback from integrating SDKs into plugins	Research data	Text/bugs	tickets,bugs, log	<10MB	ALL	Consortium	redmine	Daily	Unnecessary	0

No(1)	Last version can be kept for exploitation purposes under license agreement, and a royaltie free version will be provided to end-user partners for their own use
No(2)	Dongle-protected versions can be kept for own use
No (3)	Data can be kept only for joint exploitation

Table 1. AutoPost collected and produced data