



UNIVERSIDADE FEDERAL DO RIO DE JANEIRO – UFRJ
ESCOLA DE COMUNICAÇÃO – ECO
INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA –
IBICT
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO – PPGCI

KARINA VERAS PRAXEDES

**CONTRIBUIÇÃO EPISTÊMICA E EMPÍRICA DA ARQUIVOLOGIA NA
GESTÃO E PRESERVAÇÃO DIGITAL QUALITATIVA DOS DADOS DE
PESQUISA EM CENÁRIOS DE CIÊNCIA ABERTA**

Tese de Doutorado
Dezembro de 2025



KARINA VERAS PRAXEDES

**CONTRIBUIÇÃO EPISTÊMICA E EMPÍRICA DA ARQUIVOLOGIA NA GESTÃO
E PRESERVAÇÃO DIGITAL QUALITATIVA DOS DADOS DE PESQUISA EM
CENÁRIOS DE CIÊNCIA ABERTA**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Informação do convênio entre o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) e a Escola de Comunicação da Universidade Federal do Rio de Janeiro (Eco-UFRJ), como requisito parcial à obtenção do título de doutor em Ciência da Informação.

Orientadora: Profa. Dra. Luana Farias Sales Marques.
Coorientador: Prof. Dr. Luis Fernando Sayão.

Rio de Janeiro

2025

CIP - Catalogação na Publicação

P919c Praxedes, Karina Veras
Contribuição epistêmica e empírica da Arquivologia
na Gestão e Preservação digital qualitativa dos
dados de pesquisa em cenários de Ciência Aberta /
Karina Veras Praxedes. -- Rio de Janeiro, 2025.
334 f.

Orientadora: Luana Farias Sales.
Coorientador: Luis Fernando Sayão.
Tese (doutorado) - Universidade Federal do Rio
de Janeiro, Escola da Comunicação, Instituto
Brasileiro de Informação em Ciência e Tecnologia,
Programa de Pós-Graduação em Ciência da Informação,
2025.

1. Ciência Aberta. 2. Preservação Digital. 3.
Arquivologia. 4. Dados de pesquisa. I. Sales, Luana
Farias, orient. II. Sayão, Luis Fernando, coorient.
III. Título.

KARINA VERAS PRAXEDES

**CONTRIBUIÇÃO EPISTÊMICA E EMPÍRICA DA ARQUIVOLOGIA NA
GESTÃO E PRESERVAÇÃO DIGITAL QUALITATIVA DOS DADOS DE
PESQUISA EM CENÁRIOS DE CIÊNCIA ABERTA**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Informação do convênio entre o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) e a Escola de Comunicação da Universidade Federal do Rio de Janeiro (Eco-UFRJ), como requisito parcial à obtenção do título de doutor em Ciência da Informação.

Rio de Janeiro, 12 de dezembro de 2025.



Documento assinado digitalmente
LUANA FARIAS SALES MARQUES
Data: 17/03/2026 11:52:55-0300
Verifique em <https://validar.iti.gov.br>

Profa. Dra. Luana Farias Sales (Orientadora) PPGCI Ibict/Eco-UFRJ



Documento assinado digitalmente
LUIS FERNANDO SAYAO
Data: 19/03/2026 11:40:36-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Luis Fernando Sayão (Coorientador) PPGCI Ibict/Eco-UFRJ



Documento assinado digitalmente
SARITA ALBAGLI
Data: 20/03/2026 11:54:28-0300
Verifique em <https://validar.iti.gov.br>

Profa. Dra. Sarita Albagli (Membro interno) PPGCI Ibict/Eco-UFRJ



Documento assinado digitalmente
FABIO CASTRO GOUVEIA
Data: 11/05/2026 10:50:25-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Fabio Castro Gouveia (Membro interno) PPGCI Ibict/Eco-UFRJ



Documento assinado digitalmente
VANESSA DE ARRUDA JORGE
Data: 07/05/2026 12:57:41-0300
Verifique em <https://validar.iti.gov.br>

Dra. Vanessa de Arruda Jorge (Membro externo) Fiocruz



Documento assinado digitalmente
PAULO ROBERTO ELIAN DOS SANTOS
Data: 11/05/2026 08:26:58-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Paulo Roberto Elian dos Santos (Membro externo) PPGPAT/COC/Fiocruz

Aos meus amores, em especial ao meu vô
Zezinho (*in memoriam*).

AGRADECIMENTOS

A Deus e ao Universo, por absolutamente tudo!

À minha orientadora, Luana Sales, pelo acompanhamento paciente e empático, com trocas generosas, positivas e engrandecedoras. Ao meu coorientador, Luis Sayão, pelo compartilhamento de seu conhecimento e aprimoramento do meu saber.

Aos membros da Banca Examinadora, Profa. Sarita Albagli, Prof. Fábio Castro Gouveia, Dra. Vanessa Arruda Jorge, Prof. Paulo Roberto Elian dos Santos, Profa. Lillian Maria Araújo de Rezende Alvares e Prof. Marcus Vinícius Pereira da Silva. Muito obrigada por aceitarem fazer parte desse processo e compartilharem comigo tão valiosos conhecimentos e experiências! Me sinto privilegiada! Sem dúvida, farão toda a diferença para esta tese.

Aos colegas do PPGCI IBICT/UFRJ que, ao longo desses anos, me transmitiram força e incentivo!

Aos colegas da Fiocruz e da Casa de Oswaldo Cruz pela paciência em minhas ausências e pela torcida diária! Às mulheres doces e fortes da sala 007, pela parceria e emanção de carinho, especialmente, Ivone Sá, Érica Loureiro, Renata Perez, Denise Nogueira, Zemira Lisboa, Luciana Lima, Rafaella Serafim e Zenaide Santos. Às minhas chefes, Ana Roberta e Renata Lourenço que foram muito queridas e compreensivas comigo.

Aos queridos amigos Eliane Pontes e Marcus Vinícius pelas infinitas vezes que me ouviram falar sobre esta pesquisa e, de um jeito muito fraterno, caloroso e generoso, me acompanharam em cada momento e ajudaram demais. Vocês foram essenciais!

À minha amiga Kíssila pelo nosso encontro de alma! Pela sua escuta ativa, seus conselhos, orientações, revisões e puxões de orelha carregados com muito afeto. Você mudou a forma como enxerguei esse processo e me ajudou a passar por ele com mais leveza, resiliência e entendimento. Obrigada pelo gás final que tanto precisei. Um salve ao Universo por ter você ao meu lado, de tantas maneiras, há tantos anos! Te amo, amiga!

À minha amada filha Manuela por ser tudo de melhor que há na minha vida! Desde o mestrado me acompanhando nesses desafios acadêmicos... Na defesa da dissertação, você estava agitada e vibrando, dentro de mim, pela nossa conquista! Agora, com seus dez aninhos, você está ao meu lado, me dando força e coragem ainda mais intensas! Obrigada, meu amor! Sei que não foi fácil para nós duas... Mas seu combo de muito amor, muita paciência, companhia, compreensão e torcida foi tudo o que precisei para conseguir

concluir esta etapa. Você é minha inspiração e o motivo de tudo! Você foi, e é, minha grande parceirinha! Te amo infinitamente, minha pequena!

Aos meus pais, como agradecer tudo o que fizeram por mim até hoje? Ao passarem junto comigo cada uma das etapas desse longo ciclo, vocês me demonstraram, ainda mais, o verdadeiro significado do amor de pai e mãe! Não há palavras para descrever o quanto reconheço e sou grata por tudo o que fizeram e fazem por mim e por Manu! Também sei que não foi fácil para vocês nossa força-tarefa, haja amor! Rs... Assim como todas as outras, essa conquista também devo a vocês! Vocês fizeram com que tudo isso, até hoje, fosse possível! Obrigada por tudo em cada detalhe! Amo muito vocês!

A toda minha família, minha base, meu refúgio! Em especial à minha vó Dadá e ao vô Zezinho! Amor eterno e incondicional! Graças aos senhores estou aqui e sou o que sou! Aos meus amados: irmão Raphael, meus sobrinhos Biel e Teteco, minha cunhada Lelê e minha prima Kkq que participaram mais de perto de todo esse período e sempre tinham algo a falar ou a fazer para me tirar do caos.

Ao meu parceiro de vida, de trabalho, de estudos, de paixões compartilhadas... João Guilherme Machado. Obrigada, meu amor! Te agradeço por ter estado ao meu lado em todos os altos e baixos que vivi nesses anos desafiadores. Não dá para contar quantas vezes conversamos sobre esta pesquisa, mas dá para falar que cada uma dessas conversas enriqueceu demais o que foi escrito aqui. Obrigada pelo apoio, incentivo, força, acolhimento e amor gigantes que me fizeram chegar forte e inteira neste momento. Agradeço também ao meu lindo enteado Gueguel que, por tantas vezes, me fez companhia enquanto eu estudava, escrevia, apagava, escrevia de novo... com muito carinho e compreensão, como sempre. Amo vocês, meus meninos, meus amores! À família Machado que torceu muito junto comigo por esse momento!

Às meninas SMZ: Gi, Cela e Kika, minhas grandes amigas, que entenderam meus sumiços e ausências com muito humor, um bocado de ironia e, acima de tudo, amor! Como foi ao longo desses quase 30 anos de amizade.

À Lilian, que me acompanha e, com muita atenção, presença e acolhimento me ajudou a não deixar meu mental se perder e meu emocional me boicotar. Quantos processos... Gratidão!

Foram tantas as pessoas que estiveram ao meu lado nesses anos... e cada uma delas contribuiu de uma maneira muito particular e especial. Hoje, mais do que nunca, reconheço que tenho uma rede de apoio maravilhosa que não me deixa esmorecer. Só de lembrar tudo o que se passou, a emoção já bate forte! E os aprendizados? Foram muitos!

E para muito além do viés acadêmico, com certeza! O doutorado foi se mostrando um reflexo da minha vida e ao enxergá-lo com mais leveza ao contar com o carinho de pessoas tão especiais, tudo mudou... Hoje, meu coração se enche de alegria, fé, amor e muita gratidão!!

A todos que fizeram ou fazem parte da minha vida e que, direta ou indiretamente, colaboraram para o desenvolvimento desta tese e torceram por esse momento, meu eterno agradecimento!

“A sabedoria está em não se acreditar
saber aquilo que não se sabe” (Sócrates)

RESUMO

PRAXEDES, Karina Veras. **Contribuição epistêmica e empírica da Arquivologia na Gestão e Preservação digital qualitativa dos dados de pesquisa em cenários de Ciência Aberta**. 2025. 334 f. Tese (Doutorado em Ciência da Informação) — Instituto Brasileiro de Informação em Ciência e Tecnologia; Universidade Federal do Rio de Janeiro, 2025.

Esta tese investiga como os fundamentos teórico-metodológicos da Arquivologia podem contribuir para a qualificação da preservação digital de dados de pesquisa, ao articular-se à gestão de dados no contexto da Ciência Aberta. Parte-se do reconhecimento que, embora o movimento da Ciência Aberta promova a democratização do conhecimento e incentive o compartilhamento e o (re)uso de dados de pesquisa, a efetividade desses princípios depende de sua confiabilidade. Assume-se, assim, a hipótese que registros abertos, documentados e integrados desde os estágios iniciais do ciclo de vida dos dados de pesquisa são determinantes para sua preservação digital qualitativa. A pesquisa, de natureza aplicada e abordagem quanti-qualitativa, adotou metodologia exploratória baseada em revisão de literatura e análise empírica de repositórios institucionais brasileiros que utilizam o software Dataverse Project. Foram definidas categorias analíticas (proveniência, contexto e fixidez) com base no Modelo de Referência OAIS (2024) e elaborados instrumentos de coleta e análise que permitiram verificar o grau de transparência e maturidade da gestão de dados de pesquisa antes do depósito nos repositórios. Os resultados revelaram lacunas informacionais que comprometem a confiabilidade e o potencial de (re)uso dos dados compartilhados, especialmente pela ausência de metadados e documentação capazes de evidenciar sua proveniência digital. A partir desses achados, a tese propõe o Modelo de Fluxo para Gestão Inicial de Dados de Pesquisa, fundamentado na interlocução entre a gestão arquivística de documentos e a gestão de dados de pesquisa, com vistas a subsidiar a preservação digital qualitativa e fortalecer o (re)uso confiável desses dados. A principal contribuição da pesquisa reside na consolidação de um referencial teórico-metodológico que reafirma a relevância da Arquivologia, especialmente em cenários de Ciência Aberta, oferecendo subsídios para o estabelecimento de políticas que promovam práticas de gestão capazes de garantir transparência, confiabilidade e acesso seguro aos dados de pesquisa.

Palavras-chave: Ciência Aberta; Preservação Digital; Arquivologia; Dados de Pesquisa.

ABSTRACT

PRAXEDES, Karina Veras. **Epistemic and Empirical Contribution of Archival Science to the Management and Qualitative Digital Preservation of Research Data in Open Science Contexts**. 2025. 334 f. PhD Thesis (Doctorate in Information Science) — Instituto Brasileiro de Informação em Ciência e Tecnologia; Universidade Federal do Rio de Janeiro, 2025.

This thesis investigates how the theoretical and methodological foundations of Archival Science can contribute to enhancing the digital preservation of research data by articulating them with data management practices within Open Science context. It is based on the premise that, although the Open Science movement promotes the democratization of knowledge and encourages the sharing and (re)use of research data, the effectiveness of these principles depends on the reliability of the data produced and shared. It is therefore assumed that open, documented, and integrated records from the early stages of the research data lifecycle are essential for ensuring their qualitative digital preservation. The research, applied in nature and employing a mixed quanti-qualitative approach, adopted an exploratory methodology based on a literature review and empirical analysis of Brazilian institutional repositories using the Dataverse Project software. Analytical categories: provenance, context, and fixity, were defined based on the OAIS Reference Model (2024), and data collection and analysis instruments were developed to assess the degree of transparency and maturity in research data management prior to repository deposit. The results revealed informational gaps that compromise the reliability and potential (re)use of shared data, particularly due to the absence of metadata and documentation capable of demonstrating their digital provenance. Based on these findings, the thesis proposes the Flow Model for Initial Research Data Management, grounded in the dialogue between archival records management and research data management, aiming to support qualitative digital preservation and strengthen the trustworthy (re)use of research data. The main contribution of this research lies in the consolidation of a theoretical and methodological framework that reaffirms the relevance of Archival Science, especially in Open Science environments, by offering guidelines for establishing institutional and national policies that foster management practices capable of ensuring transparency, reliability, and secure access to research data.

Palavras-chave: Open Science; Digital Preservation; Archival Science; Research data.

LISTA DE FIGURAS

Figura 1 - Escolas de pensamento sobre Ciência Aberta e suas características.	41
Figura 2 – Categorização dos termos relacionados ao conceito de reuso de dados de pesquisa	50
Figura 3 - Modelo de comportamento do (re)uso recorrente de dados	54
Figura 4 – Estrutura conceitual da qualidade de dados	56
Figura 5 - Classificação dos critérios de metadados de qualidade da informação	57
Figura 6 – Modelo do ciclo de vida dos dados de pesquisa da JISC.....	73
Figura 7 – Modelo do ciclo de vida da curadoria - DCC	77
Figura 8 – Modelo de Objetos Digitais FAIR - FDO.....	84
Figura 9 – Estrutura de gestão de dados – DAMA-DMDBOK	92
Figura 10 - Ciclo de vida da pesquisa integrada ao ciclo de vida de seus dados	93
Figura 11 – Extrato do guia de suporte ao gerenciamento de dados de pesquisa	99
Figura 12 – Banquinho de três pernas da preservação digital.....	106
Figura 13 - Conceitos e relacionamentos de um Objeto de Informação em um Pacote de Informação.....	109
Figura 14 – Entidades funcionais do OAIS.....	111
Figura 15 - Visão detalhada das PDI na estrutura de um AIP.....	114
Figura 16 – Composição das Informações de Descrição de Preservação (PDI)	121
Figura 17 – Modelo de dados PREMIS.....	126
Figura 18 - Relações entre os Padrões CCSDS	130
Figura 19 - Camadas de preservação ao longo do ciclo de vida dos dados de pesquisa	132
Figura 20 – Extrato de um documento METS com exemplo da proveniência digital registrada	140
Figura 21 – Extrato de um documento METS com exemplo da proveniência digital registrada – Eventos e Agentes.....	141
Figura 22 – Boas práticas de gestão de documentos	150
Figura 23 – Representação dos elementos epistêmicos da Arquivologia	164
Figura 24 – Outras informações registradas na atividade de pesquisa	166
Figura 25 - Relação da informação orgânica registrada da pesquisa	167
Figura 26 - Serviços de gestão de dados de pesquisa.....	186
Figura 27 – Correlação entre etapas metodológicas, objetivos específicos e objetivo geral da pesquisa.....	198

Figura 28 – Fluxo de ciclo de vida dos dados de pesquisa com ênfase no uso de ferramentas (pela abordagem da Cadeia de Preservação)	199
Figura 29 – Passos para identificação de repositórios brasileiros para gestão de dados cadastrados no FAIRSharing	201
Figura 30 - Repositórios brasileiros para gestão de dados cadastrados no FAIRSharing	202
Figura 31 – Metodologia aplicada para escolha do campo empírico	203
Figura 32 – Diagrama esquemático do dataverse	204
Figura 33 – Passo a passo dos critérios para amostragem.....	206
Figura 34 – Hierarquização das informações a serem buscadas	208
Figura 35 – Tela Dataverse Project (versão demo) para editar opções do arquivo depositado.....	211
Figura 36 – Registro de proveniência no Dataverse.....	213
Figura 37 – Tela de cadastro para criação de um dataverse	214
Figura 38 – Tela ilustrativa da seleção de metadados para dataverse	215
Figura 39 – Tela ilustrativa da escolha de facetas do Dataverse.....	215
Figura 40 – Tela de edição de metadados de citação no dataset	221
Figura 41 – Metadados de Exportação no Dataverse	229
Figura 42 – Edição de tags no Dataverse	231
Figura 43 – Telas com ilustração das etapas da análise no nível dataset	232
Figura 44 – Telas com ilustração das etapas da análise no nível Arquivo de dados... ..	233
Figura 45 - Exemplo de arquivo Read.me no dataset selecionado do Repositório 3 ..	251
Figura 46 – Ilustração da conexão sistemática entre os temas abordados na pesquisa	258
Figura 47 - A gestão inicial dos dados de pesquisa (etapa pré-ingestão).....	269
Figura 48 - Iceberg do (re)uso sob a perspectiva das ações de gestão de dados de pesquisa	273
Figura 49 – Ciclo de vida dos dados de pesquisa e seus ambientes	275
Figura 50 – A gestão dos dados de pesquisa e sua integração com a preservação digital à luz da perspectiva arquivística.....	278
Figura 51 – Modelo de fluxo para gestão inicial de dados de pesquisa	282
Figura 52 – O Modelo de fluxo para gestão inicial de dados de pesquisa no registro da informação orgânica da pesquisa.....	290
Figura 53 - Articulação entre a gestão inicial dos dados de pesquisa e o modelo de Curadoria Digital do DCC.....	291

LISTA DE QUADROS

Quadro 1 – Metodologia da pesquisa.....	26
Quadro 2 – Descrição, critérios e parâmetros no desenvolvimento da revisão de literatura	27
Quadro 3 – Dimensões da qualidade de dados e suas relações internas e externas	57
Quadro 4 – Estratégias de preservação digital	115
Quadro 5 – Relações do FAIR com OAIS/ISO 16.363.....	125
Quadro 6 – Extrato do quadro comparativo de competências dos bibliotecários acadêmicos de dados, administradores de dados e gestores de dados.....	193
Quadro 7 – Visualização dos dataverses do recorte por ano.....	205
Quadro 8 – Quadro-base para quantificação de tipos informacionais	209
Quadro 9 – Modelo de instrumento de coleta para registro de itens de informação identificados por dataset	210
Quadro 10 – Níveis de identificação para aposição de status	210
Quadro 11 - Campos e definições da tela de cadastro de um dataverse.....	214
Quadro 12 – Campos de preenchimento inicial para cadastro de dataset	216
Quadro 13 – Campos e definições da tela de cadastro de um dataset.....	219
Quadro 14 - Campos adicionais para preenchimento de metadados no dataset.....	221
Quadro 15 – Definições dos campos adicionais	223
Quadro 16 – Categorização dos metadados dos datasets à luz das PDI do Modelo OAIS	225
Quadro 17 – Listagem de metadados encontrados nos datasets	235
Quadro 18 – Campos e definições dos metadados por domínio	236
Quadro 19 – Categorização de informações encontradas nos datasets	238
Quadro 20 – Correlação hierárquica dos itens de informação selecionados.....	241
Quadro 21 - Identificação de registros informacionais nos datasets selecionados	249
Quadro 22 - Correspondência entre as lacunas informacionais encontradas e etapas do fluxo de gestão inicial dos dados de pesquisa	270
Quadro 23 – Correlações entre as etapas da gestão de dados e as funções inerentes à gestão de documentos.....	280
Quadro 24 – Formatos de arquivo recomendados e aceitos para a fase de gestão inicial dos dados de pesquisa.....	284
Quadro 25 - Correspondência entre as lacunas informacionais encontradas e etapas do fluxo de gestão inicial dos dados de pesquisa	290

LISTA DE GRÁFICOS

Gráfico 1 – Distribuição de categorias de informações encontradas nos <i>datasets</i>	239
Gráfico 2 – Tipos informacionais da Categoria Proveniência encontrados nos datasets selecionados.....	242
Gráfico 3 – Itens de informação relativos ao Tipo Fonte ou origem.....	243
Gráfico 4 – Itens informacionais relativos ao Tipo Histórico	243
Gráfico 5 – Status de identificação de itens para Categoria Proveniência	244
Gráfico 6 – Tipos informacionais da Categoria Contexto encontrados nos datasets selecionados.....	245
Gráfico 7 – Itens informacionais relativos ao Tipo Relacionamentos.....	246
Gráfico 8 – Itens informacionais relativos ao Tipo Relações com ambiente	246
Gráfico 9 – Status de identificação de itens para Categoria Contexto	247
Gráfico 10 – Status de identificação de itens para Categoria Fixidez	248
Gráfico 11 – Disponibilização de documento Read.me nos datasets selecionados.....	251
Gráfico 12 – Fonte de identificação de itens de informação do dado	254
Gráfico 13 – Uso de tags nos arquivos de dados dos datasets selecionados	256

SUMÁRIO

1 INTRODUÇÃO	18
1.1 JUSTIFICATIVA E RELEVÂNCIA DA PESQUISA	21
1.2 OBJETIVO GERAL E ESPECÍFICOS.....	25
1.3 METODOLOGIA DA PESQUISA.....	26
1.4 ESTRUTURA DA TESE	31
2 A CIÊNCIA ABERTA EM CONSTRUÇÃO: FUNDAMENTOS EPISTEMOLÓGICOS, DESAFIOS E A QUALIDADE DOS DADOS PARA O (RE)USO.....	34
2.1 A CIÊNCIA ABERTA E SEUS DESDOBRAMENTOS PARA A PRODUÇÃO CIENTÍFICA.....	35
2.1.1 Capitalismo Cognitivo e a apropriação do conhecimento	37
2.1.2 Surgimento, marcos históricos e definições da Ciência Aberta	40
2.1.3 A Ciência Aberta diante de um campo com contradições e complexidades ..	43
2.2 A EMERGÊNCIA DA CULTURA DO COMPARTILHAMENTO E DO (RE)USO DE DADOS	48
2.2.1 Compartilhar e (re)usar: relações e desafios imbricados	49
2.3 DIMENSÕES, COMPLEXIDADES E PERSPECTIVAS DA QUALIDADE DOS DADOS	55
2.3.1 A centralidade da confiança como atributo essencial da qualidade de dados na Ciência Aberta: metadados de proveniência e a percepção do usuário	55
2.3.2 Confiança para (re)uso de dados de pesquisa: o valor de evidência e a autenticidade na perspectiva arquivística.....	64
3 O ESCOPO DA CURADORIA DIGITAL PARA O (RE)USO CONFIÁVEL DE DADOS DE PESQUISA.....	69
3.1 A CURADORIA DIGITAL DADOS DE PESQUISA: MODELOS E <i>FRAMEWORKS</i>	69
3.1.1 Modelo do ciclo de vida dos dados de pesquisa – JISC.....	71
3.1.2 Modelo do Ciclo de Vida da Curadoria Digital (DCC).....	77
3.1.3 Os Princípios FAIR	81
3.2 A GESTÃO DE DADOS DE PESQUISA COMO FUNDAMENTO PARA A PRESERVAÇÃO DIGITAL QUALITATIVA.....	89
3.2.1 A gestão, a abertura e a FAIRificação de dados de pesquisa	94
3.2.2 Planos de Gestão de Dados – PGD (<i>Data Management Plan</i> – DMP).....	97
3.2.3 Repositórios de gestão de dados de pesquisa	100
4 PRESERVAÇÃO DIGITAL QUALITATIVA: O REGISTRO DE METADADOS PARA O FORTALECIMENTO DO (RE)USO DOS DADOS DE PESQUISA...	104

4.1 O MODELO DE REFERÊNCIA OAIS COMO BASE CONCEITUAL PARA A PRESERVAÇÃO DIGITAL QUALITATIVA.....	104
4.1.1 A relevância dos metadados na preservação digital qualitativa	115
4.1.2 Referenciais internacionais complementares ao OAIS	122
4.2 PREPARAÇÃO DOS DADOS DE PESQUISA: PAVIMENTANDO O CAMINHO PARA A PRESERVAÇÃO DIGITAL QUALITATIVA	129
4.2.1 O registro da proveniência dos dados de pesquisa	136
5 A GESTÃO DE DOCUMENTOS: CONTRIBUTO DA ARQUIVOLOGIA PARA A QUALIFICAÇÃO DA PRESERVAÇÃO DIGITAL DE DADOS DE PESQUISA	143
5.1 A ARQUIVOLOGIA E UM DE SEUS CONCEITOS BASILARES: A GESTÃO DE DOCUMENTOS.	143
5.1.1 Correntes teóricas da gestão de documentos	147
5.1.2 A gestão de documentos no cenário brasileiro	156
5.2 A APLICAÇÃO PRÁTICA DO CONCEITO DE GESTÃO DE DOCUMENTOS NO BRASIL E SUAS CORRELAÇÕES PARA A GESTÃO DE DADOS DE PESQUISA	163
5.2.1 As funções arquivísticas na gestão de dados de pesquisa: Classificação, avaliação e descrição	171
5.2.2 A relevância do acompanhamento especializado para a efetivação da gestão de dados de pesquisa	186
6 PERCURSO DA PESQUISA DE CAMPO.....	198
6.1 DELIMITAÇÃO DO RECORTE EMPÍRICO	199
6.1.1 Categorias analíticas e Níveis de registro informacional	207
6.1.2 Critérios de análise	211
6.1.3 Limitação da pesquisa	233
6.2 ANÁLISE DO CAMPO EMPÍRICO: IDENTIFICAÇÃO E CATEGORIZAÇÃO DAS INFORMAÇÕES ENCONTRADAS.....	237
6.2.1 À luz da análise do conjunto de dados (<i>dataset</i>).....	237
6.2.2 À luz da análise de cada um dos arquivos de dados	253
6.3 GRAU DE TRANSPARÊNCIA E MATURIDADE: ANALISANDO A GESTÃO DE DADOS DE PESQUISA A PARTIR DAS INFORMAÇÕES ENCONTRADAS NA PESQUISA DE CAMPO	257
6.4 A ANÁLISE DO POTENCIAL DE CONFIABILIDADE PARA PRESERVAÇÃO DIGITAL QUALITATIVA DOS DADOS DE PESQUISA	263
7 A ARTICULAÇÃO DOS FUNDAMENTOS ARQUIVÍSTICOS NO CONTEXTO ATUAL DE GESTÃO DE DADOS DE PESQUISA.....	267
7.1 LACUNAS EMPÍRICAS E O PAPEL DA ARQUIVOLOGIA.....	268

7.2 MODELO DE FLUXO PARA GESTÃO INICIAL DE DADOS DE PESQUISA.	277
8 CONCLUSÃO E PERSPECTIVAS FUTURAS.....	294
REFERÊNCIAS	302
APÊNDICE A – STATUS DE IDENTIFICAÇÃO	322
APÊNDICE B – TABELA DE VERIFICAÇÃO DE ARQUIVOS DE DADOS DOS <i>DATASETS</i>	325
ANEXO A - PROPOSTA DE TEMPLATE DE PLANO DE GESTÃO DE DADOS DO ICICT/FIOCRUZ	330
ANEXO B – METODOLOGIA PARA IDENTIFICAÇÃO DA TIPOLOGIA DE DADOS DA COC/FIOCRUZ.....	332

1 INTRODUÇÃO

Esta pesquisa tem como ponto de partida inquietações profissionais e técnico-científicas despontadas no desenvolvimento de estudos voltados à análise de possíveis estratégias de preservação digital de dados de pesquisa na Fundação Oswaldo Cruz (Fiocruz). A Fiocruz é uma instituição pública e estratégica de Estado que atua nos campos da saúde, da educação e do desenvolvimento científico e tecnológico, com vistas à promoção de melhorias na qualidade de vida da população brasileira e ao fortalecimento do Sistema Único de Saúde – SUS e da saúde pública como direito.

O contexto em que essas inquietações surgem advém, inicialmente, da pesquisa de mestrado em Gestão de Documentos e Arquivos, realizada na Universidade Federal do Estado do Rio de Janeiro (UNIRIO). A pesquisadora desta tese, à época 1º Tenente Arquivista do Centro de Documentação da Aeronáutica (CENDOC), concluiu em sua dissertação que a preservação dos documentos arquivísticos do Comando da Aeronáutica tem por condição *sine qua non* a execução da gestão desses documentos. Esse achado foi possível mediante estudo quanti-qualitativo baseado em análises empíricas e a observação de que muitos dos documentos com temporalidade de guarda permanente, armazenados nos Arquivos Intermediários daquela Instituição, já se encontravam em significativo estágio de deterioração, o que inviabilizava ou prejudicava seu acesso, mesmo que siguessem a um futuro recolhimento ao Arquivo Permanente da Força Aérea Brasileira.

Em adição, posteriormente, a pesquisadora passou a atuar na coordenação do Sistema de Gestão de Documentos e Arquivos da Fiocruz (Sigda/Fiocruz), enquanto Tecnologista em Saúde Pública da Casa de Oswaldo Cruz¹. O Sigda/Fiocruz, apesar de ter sido constituído em meados de 1990, foi formalizado por Portaria da Presidência nº 353/2009 e funciona enquanto uma “Rede de serviços de arquivos” responsável por estabelecer políticas e programas de gestão de documentos e arquivos da Fiocruz em consonância com as diretrizes do Conselho Nacional de Arquivos (CONARQ), promover a gestão dos documentos arquivísticos da Fundação, entendendo-os como elementos de prova e instrumentos de apoio tanto à administração quanto à preservação de seu

¹ Unidade técnico-Científica da Fiocruz que tem por missão: “Produzir e disseminar o conhecimento histórico da saúde e das ciências biomédicas; **preservar e valorizar o patrimônio cultural da saúde**; educar em seus campos de atuação e divulgar ciência e tecnologia em saúde, de forma a contribuir para o desenvolvimento científico, cultural e social” (Fiocruz, [s.d], grifo nosso). Disponível em: <https://www.coc.fiocruz.br/index.php/pt/institucional/missao-visao-e-valores>. Acesso em: 23 set. 2025.

patrimônio científico e cultural; e assegurar a implementação e harmonização das operações técnicas e procedimentos da gestão documental nas fases corrente, intermediária e permanente.

A atuação do Sigda/Fiocruz permeia diversas esferas institucionais, por intermédio do desenvolvimento de reflexões e diálogos multidisciplinares internos e externos a respeito da interface entre gestão de documentos, preservação digital, curadoria de dados de pesquisa, memória administrativa, qualidade, dentre outras questões que reforçam a abrangência transversal da área em todos os setores da Fiocruz. Por essa razão, a pesquisadora, na condição de membro e coordenadora do Sigda/Fiocruz, tem constantemente atuado junto a diversas instâncias institucionais que são responsáveis não somente pela gestão de documentos arquivísticos, mas, também, pelo planejamento e execução de ações relativas tanto ao tratamento técnico dos dados de pesquisa quanto à preservação digital no âmbito da Fiocruz.

Essa atuação conduziu a percepção da necessidade de que os conhecimentos arquivísticos fossem ainda mais aprofundados para que pudessem contribuir com questões atuais e ainda incógnitas impostas tanto pelo uso de novos formatos digitais na produção e tramitação de documentos quanto pela demanda de preservar dados de pesquisa oriundos das atividades científicas da Fundação. A partir disso, reconheceu-se ainda mais a importância de uma abordagem interdisciplinar que possibilitasse a interlocução da Arquivologia com outras áreas que atuam na reflexão e nas práticas envolvidas no tratamento técnico desses objetos digitais.

Ao realizar uma busca por referências e/ou relatos de experiências disponíveis na literatura brasileira, ao passo em que se identificou a incipiência de uma abordagem pragmática que correlacionasse os temas preservação digital, gestão de dados de pesquisa e gestão de documentos, observou-se, ainda, que múltiplas instituições de pesquisa e desenvolvimento, nacionais e internacionais, têm direcionado esforços significativos para o desenvolvimento de políticas institucionais e para a reformulação de práticas científicas, buscando alinhamento aos princípios da Ciência Aberta.

A Ciência Aberta, é um movimento global que propõe paradigmas inovadores para a produção, organização, compartilhamento e reutilização do conhecimento científico (Fiocruz, 2020), que se caracteriza como um "construtor inclusivo" fundamentado na democratização do acesso ao conhecimento e no compartilhamento aberto de recursos científicos (Unesco, 2022). Tendo como objetivo potencializar

colaborações e gerar benefícios que transcendem o âmbito científico, alcançando a sociedade como um todo.

Longe de se restringir ao acesso aberto a publicações e dados, a Ciência Aberta preconiza a abertura de todo o processo de pesquisa, o que demanda métodos, ferramentas e *workflows* que facilitem a disponibilização, o compartilhamento e a reutilização dos dados coletados, gerados e utilizados ao longo do percurso de pesquisa.

Nesse sentido, conforme indica Michael Nielsen (2011, *apud* Gezelter, 2009, tradução nossa) “a Ciência Aberta é a ideia de que o conhecimento científico de todos os tipos deve ser compartilhado abertamente o mais cedo possível no processo de descoberta”². Segundo Gezelter (2009), a Ciência Aberta se manifesta a partir do que o autor considera como seus quatro objetivos fundamentais:

- 1) transparência na metodologia experimental, observação e coleta de dados;
- 2) disponibilidade pública e reutilização de dados científicos;
- 3) acessibilidade pública e transparência da comunicação científica; e
- 4) utilização de ferramentas *web* para facilitar a colaboração científica.

Diante da compreensão dos objetivos que Gezelter (2009) considera fundamentais para a Ciência Aberta, destaca-se a necessidade de incremento tecnológico para o seu alcance. Dessa forma, é importante também apontar nesse cenário o papel da *eScience*, que, conforme apontam Hey *et al.* (2009), Jim Gray caracteriza como o quarto paradigma científico³. De acordo com os autores, Gray em sua apresentação proferida no *National Research Council*, em 2007, expôs que a tecnologia da informação está mudando quase tudo na ciência, o que fez emergir, a partir da utilização intensiva de dados, esse novo paradigma, cujo objetivo é “ter um mundo em que toda a literatura científica esteja *online*, todos os dados científicos estejam *online* e interoperem entre si” (Hey *et al.*; 2009, p.30, tradução nossa).

² Em 2011, Gezelter publicou em seu blog (<https://openscience.org/blog>) o que Michael Nielsen, na lista de discussão do site da *Open Knowledge Foundation* (<https://okfn.org/>) chamou de uma “definição informal” da Ciência Aberta:

³ Em um workshop onde apresenta o livro intitulado “*The Fourth Paradigm: Data-Intensive Scientific Discovery*”, Hey expõe um compilado sobre o entendimento de Gray quanto aos quatro paradigmas científicos. São eles 1º paradigma como Ciência Experimental (baseado na descrição de fenômenos naturais), 2º paradigma como Ciência Teórica (com desenvolvimento das Leis de Newton, Equações de Maxwell...), 3º paradigma como Ciência Computacional (baseado em simulações de fenômenos complexos) e o 4º paradigma como Ciência Intensiva em dados (cientistas sobrecarregados com conjuntos de dados de várias fontes diferentes - dados capturados por instrumentos, dados gerados por simulações e dados gerados por redes de sensores) (Tansley *et al.*, 2009, slide 7, tradução nossa).

De acordo com Sales e Sayão (2018), as novas tecnologias digitais, impulsionadas pelo aumento da capacidade computacional e de redes, pelo armazenamento digital e pela simulação por meio de programas de computador, têm um potencial transformador e de reconfiguração na comunicação da ciência e no próprio método científico. Em consequência, nas duas últimas décadas, observou-se um aumento substancial em termos de complexidade, volume e importância dos dados no fluxo das pesquisas científicas.

No entanto, embora essas ações possam, inegavelmente, gerar inúmeras benesses para a condução de uma ciência mais transparente, democrática e inclusiva, elas também apresentam um contraponto que precisa ser observado: apesar do fluxo abundante de dados possibilitar novas formas de exploração colaborativa e de descobertas que conectam pesquisadores e aceleram as taxas de entendimento científico, por vezes, ocorre de maneira caótica (Sayão; Sales, 2020).

Assim, tem-se testemunhado um crescente volume de produção e uso de dados, o que não pressupõe que esse processo esteja acompanhado de uma metodologia que estabeleça sua organização a partir de procedimentos e fluxos padronizados, com vistas a presumir sua autenticidade. Tal circunstância fez perceber que, a ampla geração de dados e a expansão de sua capacidade de compartilhamento e (re)uso, à medida que são fatores que influenciam de forma positiva e direta o progresso das pesquisas científicas e, portanto, a sustentabilidade e inovação da ciência, também são fatores que emergem uma necessidade proporcionalmente desafiadora: a gestão e a preservação digital desses dados de pesquisa.

Diante do cenário apresentado, esta pesquisa parte do **questionamento: De que modo a articulação entre os fundamentos teórico-metodológicos da Arquivologia e a gestão de dados de pesquisa pode subsidiar a preservação digital qualitativa, visando à qualificação das práticas científicas em conformidade com os princípios da Ciência Aberta?**

Como **premissa** para a busca dessa resposta, apresenta-se a seguinte **hipótese: registros abertos, documentados e integrados desde as fases iniciais do ciclo de vida dos dados de pesquisa são não apenas necessários, mas também determinantes para a promoção de sua preservação digital qualitativa, a fim de fortalecer sua confiabilidade e seu potencial de (re)uso.**

1.1 JUSTIFICATIVA E RELEVÂNCIA DA PESQUISA

Os benefícios que a Ciência Aberta pode propiciar tanto para a ciência, em virtude da aceleração de inovações, quanto para a sociedade, à medida em que possibilita transparência das pesquisas, confirmação de resultados obtidos, uso de dados na evolução de questões socioeconômicas e maior participação da sociedade nos próprios processos de produção do conhecimento, são evidenciados pela Unesco no documento “Recomendações para Ciência Aberta”:

[...] o acesso mais amplo aos processos e resultados científicos pode melhorar a efetividade e a produtividade dos sistemas científicos, reduzindo os custos de duplicação na coleta, criação, transferência e reutilização de dados e materiais científicos, permitindo outras pesquisas com os mesmos dados, e aumentando o impacto social da ciência, multiplicando as oportunidades de participação local, nacional, regional e global no processo de pesquisa, assim como as oportunidades para a circulação mais ampla dos resultados científicos (Unesco, 2022, grifo nosso).

A progressão de ações em direção à efetivação desse movimento requer que estudos sobre metodologias sejam desenvolvidos para além do estofo teórico, no intuito de amparar a prática das operações que precisam ser realizadas em direção à consolidação dos objetivos da Ciência Aberta. Dessa forma, a relevância do compartilhamento e do (re)uso dos dados de pesquisa é evidente, contudo, a efetivação do uso, do compartilhamento e do (re)uso está intrinsecamente condicionada à manutenção da vida útil desses dados.

Observa-se que dados de pesquisa estão sendo produzidos em escalas cada vez mais volumosas e abundantes, em termos de quantidade, natureza, categorias, gêneros e formatos. Junto a isso, o estímulo por uma maior abertura e compartilhamento dos dados de pesquisa intensifica ainda mais as complexidades inerentes à sua produção, manutenção e, conseqüentemente, ao seu (re)uso. Sendo assim, discussões e estudos sobre a preservação desses dados estão se tornando ainda mais urgentes.

Sem que sejam subtraídas as vantagens vinculadas ao meio digital, Innarelli (2012, p.12) destaca que “assim como a Era da informação nos traz uma facilidade imensa na geração de dados, informações e documentos, o mesmo acontece com a perda destas informações [...]”. Isso pode ser percebido diante da afirmação de que “muitos dos dados de pesquisa científica já foram irremediavelmente perdidos”, disseram Sayão e Sales, em 2012, quando complementaram com a constatação de que este problema “estava longe de ser superado” àquela época. Mais de uma década depois, percebe-se que este cenário ainda pode ser constatado nos dias de hoje.

Recursos para criar e consumir informação digital têm progredido de forma constante e, com esses avanços, surge a necessidade de desenvolver capacidade suficiente para assegurar a gestão adequada desta informação a longo prazo (Lavoie, 2014, p.4). Portanto, destaca-se que, enquanto não são analisadas e estabelecidas metodologias que busquem na prática a efetivação da gestão e preservação desses dados, eles permanecem em condição de vulnerabilidade e, conseqüentemente, suscetíveis à perda.

Nesse sentido, a perda se refere a dois vieses. Pode-se perder o dado em si, quando relacionado a uma possível incapacidade de acessá-lo, seja porque o dado não foi devidamente identificado, descrito e/ou organizado, ao ponto de não ser mais localizável, seja por alguma corrupção de seu conteúdo ou por estar ilegível, por exemplo.

Contudo, essa perda não se limita apenas aos aspectos mencionados, pois mesmo nos casos em que os dados ainda possam ser localizados, é possível supor que, se não possuírem uma qualidade informacional mínima que propicie sua contextualização e proveniência, sua confiabilidade pode ficar ameaçada e assim, eles podem ser considerados “perdidos” à nível de capacidade de (re)uso.

Isto porque, diante da vasta gama de possibilidades de categorização de um dado, o seu contexto torna-se ainda mais relevante numa perspectiva de análise para uso ou reuso posterior, afinal, “números e caracteres transmitem pouco ou até mesmo nenhum significado, a menos que sejam documentadas de alguma forma” (Fonseca, 2018, p.10). Sendo assim, assegurar que os dados “sejam compreensíveis agora e no futuro” perpassa pela ação de documentá-los, pois “se os dados são bem documentados antes e durante a formação da coleção de dados, eles serão mais facilmente entendidos e reutilizados” (Sayão; Sales, 2015, p. 16).

Portanto, é primordial ter

Uma visão integrada de todo o decurso da pesquisa científica, lançando mão de saberes e fazeres consolidados da arquivologia, abrangendo aspectos inerentes ao tratamento do dado ao longo do seu ciclo de vida, bem como processos e documentos produzidos na sua efetivação, **podem consubstanciar** a disponibilização qualificada, o (re)uso efetivo, a **confiança** e a preservação desses dados pelo tempo necessário (Sá *et al.*, 2021, p. 6, grifo nosso).

Ao reconhecer a particular e essencial contribuição dos fundamentos teórico-metodológicos da Arquivologia nessa conjuntura, é proporcionalmente relevante frisar que a gestão e preservação digital é reconhecidamente uma tarefa multidisciplinar, que envolve tanto as áreas de informação (Arquivologia, Biblioteconomia, Ciência da

Informação), quanto as de Tecnologia de Informação e Comunicação e de Direito, e que no cenário aplicado aos dados de pesquisa, as disciplinas específicas dos pesquisadores produtores e/ou usuários também devem ser associadas.

Além disso, a literatura sobre a abertura dos dados nem sempre considera os aspectos qualitativos necessários ao (re)uso e mesmo quando os metadados são apontados como essenciais à preservação digital dos dados de pesquisa, seu fim tende a se restringir à função de acesso, voltada à recuperação da informação, desconsiderando seu papel essencial no registro do contexto e da proveniência.

Sendo assim, a pesquisa se justifica em virtude da necessidade do desenvolvimento de metodologias de gestão e preservação digital para dados de pesquisa e, para tanto, propõe-se a articulação teórica entre a Gestão de Dados e a Arquivologia, mais especificamente ao que tange à gestão de documentos, a fim de estabelecer diretrizes para promover a manutenção da vida útil dos dados de pesquisa com vistas ao seu efetivo uso, compartilhamento e reuso.

Quanto à relevância da pesquisa, julga-se oportuno apontar que, ultimamente, organizações internacionais (como a Unesco, a OCDE e a Comissão Europeia, por exemplo), editoras e revistas científicas, instituições de fomento à pesquisa, além de parcelas significativas de profissionais das comunidades científica e acadêmica e da sociedade, tem incentivado a abertura, o compartilhamento e o (re)uso dos dados de pesquisa, sob argumentos que destacam os benefícios científicos e sociais.

No entanto, nota-se que muitas dessas iniciativas se orientam mais por diretrizes gerais do que por uma conscientização efetiva sobre a importância da transparência, reprodutibilidade e colaboração científica. Nesse sentido, também se entende que empreender a gestão de dados de pesquisa em uma instituição com pretensões de adequar suas pesquisas a essas diretrizes é uma condição para isso se estabeleça.

Diante do exposto, este trabalho propõe uma aproximação entre os campos da Arquivologia e da Ciência da Informação, reconhecendo que o saber arquivístico pode contribuir significativamente para qualificar a preservação digital dos dados de pesquisa, sobretudo quanto à autenticidade e a confiabilidade requeridas para seu (re)uso responsável. Akmon *et al.* (2011), embora reconheçam que até bem pouco tempo a preservação de dados científicos tenha sido uma preocupação secundária dos arquivistas⁴,

⁴ Os autores indicam como motivos: o não reconhecimento dos dados científicos como com algum valor histórico, ou seja, fora do seu escopo de atuação profissional; o entendimento de que os próprios cientistas tinham maior conhecimento para avaliar se seus dados seriam úteis no futuro e, portanto, passíveis ou não

destacam a importância desses profissionais nas ações relativas à preservação de dados de pesquisa e o papel que podem desempenhar na curadoria de dados.

Considerando que um dos vieses da Ciência Aberta é a sustentabilidade da ciência, compreende-se que este estudo possui potencial de amplificar a efetivação de uma ciência mais sustentável, em razão da gestão de dados de pesquisa, desde sua criação/coleta, propiciar a execução de estratégias de preservação digital dos seus dados com qualidade e confiabilidade, e, conseqüentemente, viabilizar seu (re)uso.

Por fim, pode-se destacar que o estudo pretende promover a discussão sobre o tema e ampliar o reconhecimento de sua importância. Além disso, visa demonstrar o potencial de sucesso e robustez da costura de conhecimentos interdisciplinares da Ciência da Informação em prol de um objetivo em comum.

Espera-se que os resultados desta tese contribuam para o aprimoramento das práticas de gestão e preservação de dados de pesquisa, subsidiando a elaboração de políticas institucionais, o desenvolvimento de ferramentas tecnológicas e a consolidação de um tratamento aos dados de pesquisa voltado à confiança, transparência e (re)uso no âmbito da Ciência Aberta.

1.2 OBJETIVO GERAL E ESPECÍFICOS

O objetivo geral desta tese é propor a gestão arquivística⁵ de documentos como fundamentação teórico-metodológica para o aprimoramento da gestão de dados de pesquisa, desde seus estágios iniciais, com vistas a viabilizar a preservação digital qualitativa.

Para o alcance do objetivo geral da pesquisa, elenca-se os seguintes objetivos específicos:

a) Estabelecer o referencial teórico que subsidia a compreensão necessária para a proposição;

de serem retidos; a avaliação simplória a partir da distinção dos dados como experimentais ou observacionais, onde estes últimos, por serem caros e complexos de serem reproduzidos, deveriam ser mantidos a longo prazo, enquanto dados oriundos de experimentação, pela facilidade de serem novamente coletados ou gerados, ao invés de serem reutilizados, teriam um prazo menor de guarda; e o favorecimento da preservação em repositórios de dados específicos da disciplina, sob os quais os arquivistas não se sentiam atuantes.

⁵ Embora entenda-se que a “gestão de documentos” possua um conceito consolidado na área arquivística, optou-se por somar ao termo a palavra “arquivística” para frisar a qualificação dessa gestão de documentos, considerando a possível variedade de áreas de atuação do público-alvo dessa tese.

b) Verificar o grau de maturidade da gestão aplicada antes do depósito dos dados de pesquisa nos repositórios do recorte a partir da análise da transparência de seus metadados;

c) Evidenciar a importância dos estágios iniciais da gestão de dados de pesquisa para viabilizar a efetivação da preservação digital qualitativa à luz do potencial de confiabilidade dos dados de pesquisa; e

d) Apresentar um modelo de fluxo para a gestão inicial de dados de pesquisa.

1.3 METODOLOGIA DA PESQUISA

O valor da presente tese reside no potencial de utilidade da interlocução proposta, em razão de seu desenvolvimento ser direcionado para a solução prática de uma questão específica. Dessa forma, quanto à natureza, a metodologia adotada para esta pesquisa classifica-se como aplicada, cuja característica fundamental é o interesse na aplicação, utilização e consequências práticas dos conhecimentos (Gil, 1994).

Adicionalmente, em consonância com a perspectiva de que investigações científicas combinam múltiplos métodos, esta tese adota a abordagem quanti-qualitativa de Minayo (1993). Essa escolha considera que o estudo quantitativo pode gerar questões para aprofundamento qualitativo e vice-versa. Complementarmente, a pesquisa possui objetivo exploratório, visando proporcionar uma visão geral e aproximativa sobre um tema pouco explorado (Gil, 1994). Para tanto, a consolidação de um arcabouço teórico multidisciplinar foi essencial, englobando temas como Ciência Aberta, Curadoria digital, Preservação digital e Gestão de documentos.

Portanto, o percurso metodológico incluiu procedimentos, que segundo Fonseca (2002) se enquadram em bibliográficos, documentais e pesquisa de campo. Pode-se dizer, então, que a pesquisa seguiu a seguinte metodologia:

Quadro 1 – Metodologia da pesquisa

METODOLOGIA DA PESQUISA	
Quanto à abordagem	Quanti-qualitativa
Quanto à natureza	Aplicada
Quanto aos objetivos	Exploratório
Quanto aos procedimentos	Bibliográfico
	Documental
	Pesquisa de campo

Fonte: Elaborado pela autora (2025).

Assim, para atingir o primeiro objetivo específico “estabelecer o referencial teórico que subsidia a compreensão necessária para a proposição do modelo” foram adotados os procedimentos bibliográficos e documentais (Fonseca, 2002).

Primeiramente, foi realizada a revisão de literatura, com aprofundamento em literaturas científica e técnica. O escopo dessa revisão foi composto por teses, dissertações e artigos, de autores nacionais e internacionais, além de produtos institucionais advindos de estudos que abordassem reflexões e experiências práticas. A inclusão de produtos institucionais ocorreu com intuito de identificar possíveis propostas já existentes sobre curadoria digital e preservação digital de dados de pesquisa, no cenário da Ciência Aberta, que pudessem embasar a estruturação da proposição apresentada nesta tese.

As seguintes bases de dados foram consultadas: Bases de Dados: Scopus, Web of Science e Brapci, para as quais foi utilizada como estratégias de busca a query: (“research data” AND (“digital curation” OR “data curation” OR “digital preservation” OR “data stewardship” OR “data conservancy”)), retornando 1.666, 1.232 e 91 resultados, respectivamente. Esses números demandaram a definição de critérios de seleção baseados na leitura de títulos e resumos dos artigos, buscando identificar potenciais contribuições para esta pesquisa.

Diante da intenção de conhecer experiências práticas que, porventura, não estejam publicizadas em comunicações formais, optou-se por adicionar à busca o Google Scholar. Essa ampliação permitiu o acesso a materiais complementares, como apresentações, instrumentos técnicos e documentos institucionais.

Quadro 2 – Descrição, critérios e parâmetros no desenvolvimento da revisão de literatura

Etapas	Descrição	Crítérios	Parâmetros Utilizados
Revisão de Literatura	Busca em bases de dados (Scopus, Web of Science, Brapci e Google Scholar) para identificar propostas sobre Ciência Aberta, gestão de dados de pesquisa e preservação digital.	Seleção por título e resumo com análise de relevância para o tema.	Query: “research data” AND (“digital curation” OR “data curation” OR “digital preservation” OR “data stewardship” OR “data conservancy”). Retorno: 1.666 (Scopus), 1.232 (WoS), 91 (Brapci).
Crítérios de Inclusão	Estudos sobre Ciência Aberta, curadoria digital ou preservação digital, com enfoque em modelos	Artigos revisados por pares, teses, dissertações e	Aderência às categorias OAIS (proveniência, contexto e fixidez).

	metodológicos aplicáveis.	documentos técnicos institucionais.	
Crerérios de Exclusão	Trabalhos sem relação direta com curadoria de dados de pesquisa ou com escopo restrito a aspectos puramente tecnológicos.	Exclusão por leitura de título, resumo e ausência de alinhamento conceitual.	Artigos duplicados removidos.

Fonte: Elaborado pela autora (2025).

Foram utilizados livros, políticas, modelos e artigos provenientes de diferentes nacionalidades, parte dos quais integra a biblioteca pessoal da pesquisadora, construída ao longo de sua formação acadêmica e profissional. Também foram consultados títulos sobre diversos temas da área arquivística, epistemológicos, técnicos e práticos, principalmente, àqueles relativos à gestão e preservação de documentos arquivísticos digitais.

Além disso, pelo fato de os dados de pesquisa constituírem um objeto de análise multidisciplinar, foram analisados títulos com reflexões e orientações oriundas de outras disciplinas como Biblioteconomia, Tecnologia da Informação e Ciência da Informação.

O produto da revisão em questão (quadro 2) foi apresentado nesta tese no decorrer das seções primárias 2 “A Ciência Aberta em construção: fundamentos epistemológicos, desafios e a qualidade dos dados para (re)uso”, 3 “O escopo da curadoria digital para o (re)uso confiável de dados de pesquisa” e 4 “Preservação qualitativa: o registro de metadados para o fortalecimento do (re)uso dos dados de pesquisa”. A seção primária 5 foi desenvolvida com base no arcabouço teórico derivado das práticas profissionais e acadêmicas da autora.

Cabe ressaltar que, de maneira contraditória, até mesmo curiosa, a despeito do que foi estudado sobre as premissas da Ciência Aberta e as vantagens do acesso aberto, alguns entraves foram encontrados para acessar determinados artigos que possivelmente contribuiriam para esta construção teórica. Observou-se a necessidade de criação de *login*, assinatura ou pagamento, bem como a ocorrência de *links* inativos e páginas indisponíveis.

Para chegar ao segundo objetivo específico “verificar o grau de maturidade da gestão aplicada antes do depósito dos dados de pesquisa nos repositórios do recorte à luz da transparência de seus metadados”, foi adotado o procedimento de pesquisa de campo (Fonseca, 2002).

Para analisar os dados empíricos oriundos da pesquisa de campo, foi adotada a abordagem quanti-qualitativa de Minayo (1993) como caminho metodológico. Esse caminho permitiu não apenas identificar e quantificar a presença de metadados nos repositórios de gestão de dados de pesquisa, mas também qualificar as informações neles contidas, conforme está detalhadamente disposto na seção primária 6 “Percurso da pesquisa de campo”. Tal escolha visou propiciar uma avaliação de sua adequação como base para uma preservação digital de longo prazo, orientada pela confiabilidade e pelo potencial de (re)uso futuro.

A dimensão quantitativa envolveu, primeiramente, a delimitação do recorte empírico (subseção 6.1) que se concentrou em repositórios para gestão de dados de pesquisa cadastrados no FAIRSharing. A partir de filtros aplicados, foram identificados sete repositórios, dos quais seis apresentavam *status* de prontos.

Com vistas a possibilitar uma análise aprofundada e comparável, estabeleceu-se como critério adicional a utilização de um mesmo *software*, resultando na seleção de três repositórios que utilizam o Dataverse Project. A escolha foi reforçada tanto pela ampla adoção internacional do sistema quanto pela familiaridade prévia desta pesquisadora com sua estrutura e funcionalidades. Esses repositórios, em conjunto, abrangem diferentes áreas do conhecimento, no entanto o assunto “Medicina, saúde e ciências da vida” foi selecionado para compor o recorte temático da análise.

A partir desse universo, definiu-se uma amostragem composta por 24 *datasets*, distribuídos entre os três repositórios, selecionados de forma aleatória a partir de critérios temporais e proporcionais. Com o intuito de preservar o foco analítico no conteúdo informacional, os repositórios e *datasets* não foram nominalmente identificados, sendo apresentados por codificação.

A análise, conduzida sob o olhar desta pesquisadora, por intermédio de um perfil de usuário comum do sistema, fundamentou-se no referencial do modelo OAIS⁶, especificamente nas categorias de Informação de Proveniência, Informação de Contexto e Informação de Fixidez, entendidas como centrais para a aferição da autenticidade e da confiabilidade dos dados de pesquisa digitais. Em virtude da baixa adoção de padrões específicos de proveniência nos repositórios brasileiros, conforme evidenciado por

⁶ O Modelo de Referência *Open Archival Information System* – OAIS se caracteriza como uma “Prática Recomendada” que foi desenvolvida pelo *Consultative Committee for Space Data Systems* (CCSDS) em 2002 e atualizada em 2024. Tal modelo é amplamente utilizado em nível mundial como base para o estabelecimento de práticas de preservação digital.

estudos recentes, a investigação concentrou-se na identificação de registros informacionais que, ainda que não explicitamente rotulados como tal, pudessem refletir essas categorias. Para isso, foram definidas na subseção 6.1.1 as categorias analíticas (proveniência, contexto e fixidez) e os níveis de registro de informações.

Como forma de guiar e sistematizar a exploração dos repositórios selecionados, foram desenvolvidos instrumentos de coleta hierarquizados (tabelas e planilhas) utilizados para examinar os *datasets* e os arquivos dos dados de pesquisa compartilhados. Para operacionalizar a análise quanti-qualitativa, foi desenvolvida uma hierarquização analítica em três níveis: categoria, tipo e item de informação, associada a um esquema de classificação dos registros em ausentes, parciais ou integrais.

A coleta e exame dos dados foram realizadas a partir do estabelecimento de critérios de análise (subseção 6.1.2) e contemplaram tanto os metadados de citação e domínio disponibilizados nos *datasets* quanto os arquivos de dados associados, incluindo metadados de exportação, arquivos *Read.me*, configuração do documento e o uso de *tags*.

Esse delineamento metodológico permitiu observar, de forma sistemática, como esses repositórios de gestão de dados do recorte têm registrado informações relacionadas à proveniência, ao contexto e à fixidez, concentrando-se na verificação do grau de transparência de seus metadados.

Com base nessa verificação entendeu-se possível analisar o potencial de confiabilidade dos dados de pesquisa compartilhados, com vistas evidenciar a importância dos estágios iniciais da gestão de dados de pesquisa na efetivação da preservação digital qualitativa, para alcance do terceiro objetivo específico “evidenciar a importância dos estágios iniciais da gestão de dados de pesquisa para viabilizar a efetivação da preservação digital qualitativa à luz do potencial de confiabilidade dos dados de pesquisa”.

Dessa forma, foi realizada a articulação entre a gestão arquivística de documentos e a gestão de dados de pesquisa (seção primária 7) e apresentado um modelo de fluxo para a gestão inicial de dados de pesquisa, em consonância ao quarto objetivo específico.

Em resumo, a metodologia foi construída a fim de integrar o referencial teórico estabelecido sobre a temática da presente pesquisa aos resultados obtidos por meio da análise dos repositórios. Os procedimentos adotados tiveram como intuito a consolidação do conhecimento adquirido e a identificação de deficiências operacionais que acarretam lacunas informacionais de impacto à qualidade da preservação digital dos dados de pesquisa.

O percurso metodológico apresentado culminou na proposição da gestão arquivística de documentos como fundamentação teórico-metodológica para o aprimoramento da gestão de dados de pesquisa, desde seus estágios iniciais, com vistas a viabilizar a preservação digital qualitativa, no contexto da Ciência Aberta.

1.4 ESTRUTURA DA TESE

Além da introdução, nesta **seção primária 1**, a tese organiza-se da seguinte forma:

A **seção primária 2**, “A Ciência Aberta em construção: fundamentos epistemológicos, desafios e a qualidade dos dados para o (re)uso”, revisita a base teórica e o contexto da Ciência Aberta, no qual esta pesquisa se insere. Parte das relações de poder no campo científico, influenciadas por Bourdieu, Foucault e Deleuze, para discutir o “capitalismo cognitivo” e as tensões entre a socialização e a privatização do conhecimento. Em seguida, apresenta o surgimento, alguns marcos históricos e as diferentes escolas de pensamento da Ciência Aberta, explorando suas contradições e desafios. Por fim, aborda a complexidade do compartilhamento e (re)uso de dados de pesquisa, destacando as dimensões e perspectivas da qualidade desses dados, com ênfase na “centralidade da confiança” como atributo essencial para seu (re)uso efetivo.

Esclarece-se que, embora sejam apresentadas diversas nuances e frentes concernentes à Ciência Aberta, trata-se de um campo de amplo escopo e em constante expansão, não havendo, portanto, a pretensão de desenvolvê-las em profundidade. A intenção ao expô-las é evidenciar granularidades envolvidas em um tema que, por vezes, é tratado de maneira homogênea. Cabe demarcar que os apontamentos feitos ao longo da referida seção visam contextualizar esta pesquisa no cenário complexo e diverso da Ciência Aberta. No entanto, esta tese volta-se especificamente a explorar os aspectos implicados diretamente ao (re)uso de dados de pesquisa.

A **seção primária 3**, “O escopo da Curadoria Digital para o (re)uso confiável de dados de pesquisa”, apresenta a curadoria digital como um processo holístico que abrange todo o ciclo de vida dos dados. Analisa os modelos e *frameworks* que orientam essas práticas, como o Modelo do Ciclo de Vida dos Dados de Pesquisa da JISC, o Modelo do Ciclo de Vida da Curadoria Digital, da DCC, e os princípios FAIR. Discute também a gestão de dados de pesquisa como fundamento para a preservação digital qualitativa, destacando aspectos como abertura e FAIRificação dos dados, elaboração de Planos de

Gestão de Dados (PGD) e uso de repositórios de gestão de dados de pesquisa enquanto estratégias essenciais para garantir o acesso contínuo a dados significativos e autênticos.

A **seção primária 4**, “Preservação qualitativa: o registro de metadados para o fortalecimento do (re)uso dos dados de pesquisa”, apresenta o entendimento sobre a preservação digital, tendo o Modelo de Referência OAIS (2024) como base conceitual. Demonstra a relevância dos metadados para a efetivação desse modelo e a conexão com outros padrões internacionais correlatos. Além disso, evidencia a necessidade de que os dados de pesquisa sejam preparados com registros de proveniência capazes de conferir-lhes a confiabilidade que fortalece seu (re)uso em diferentes contextos de pesquisa, no presente e no futuro.

A **seção primária 5**, “Gestão de documentos: contributo da Arquivologia para a qualificação da preservação digital dos dados de pesquisa”, discute um de seus conceitos basilares: a gestão de documentos. Apresenta um breve histórico, duas correntes teóricas atuais, sua adoção no cenário brasileiro e a aplicação prática das funções arquivísticas, em especial: classificação, avaliação e descrição, precedidas pela técnica de identificação, demonstrando como essas funções se correlacionam com o contexto da gestão de dados de pesquisa. Adicionalmente, destaca-se a relevância do acompanhamento especializado para a realização dessas ações.

Com esse referencial teórico, desenvolve-se a **seção primária 6**, “Percurso da pesquisa de campo”, em que são apresentadas a delimitação do estudo, as categorias analíticas e os níveis de registro informacional, os critérios utilizados na análise empírica e as limitações observadas. Em seguida, expõe-se a análise do campo empírico a partir da verificação dos conjuntos de dados (*datasets*) e de cada um dos arquivos que compõem o recorte selecionado. Essa etapa permitiu avaliar a maturidade da gestão de dados aplicada anteriormente ao depósito desses dados de pesquisa nos repositórios analisados a partir da verificação do grau de transparência dos metadados dos dados de pesquisa compartilhados, evidenciando, assim, seu potencial de confiabilidade.

Assim, a análise empírica revelou lacunas informacionais significativas que, na **seção primária 7**, “A articulação dos fundamentos arquivísticos no contexto atual da gestão de dados de pesquisa”, são identificadas e tratadas por meio da proposição de um modelo de fluxo para a gestão inicial de dados de pesquisa, integrando gestão arquivística de documentos e gestão de dados de pesquisa, concluindo, então, a materialização do objetivo final desta tese.

Cumprе esclarecer que, conforme será percebido ao longo da leitura desta tese, as primeiras seções primárias (2 e 3) são fruto da revisão de literatura que fundamenta esta investigação. Embora possam parecer extensas, foram intencionalmente desenvolvidas para refletir o percurso de construção do conhecimento que esta pesquisadora precisou trilhar para se inserir, e se encontrar, de forma aprofundada no contexto que envolve a Ciência Aberta e a curadoria digital de dados de pesquisa.

A partir desse desencadeamento teórico, a estrutura da tese foi organizada de modo a acompanhar os questionamentos emergentes ao longo do desenvolvimento da pesquisa, permitindo que cada seção respondesse e aprofundasse questões que surgiram no percurso investigativo.

2 A CIÊNCIA ABERTA EM CONSTRUÇÃO: FUNDAMENTOS EPISTEMOLÓGICOS, DESAFIOS E A QUALIDADE DOS DADOS PARA O (RE)USO

Para fundamentar o desdobramento deste estudo, é preciso cumprir o objetivo específico de aprofundar o conhecimento sobre Ciência Aberta, curadoria digital, preservação digital e gestão de documentos. Assim, nesta seção, é importante trazer como ponto de partida o entendimento contextual ao qual o objeto macro de análise, a pesquisa científica (em constante modificação e disputas, inclusive, e sob escrutínio de diversos sujeitos, áreas, vieses e interesses) está situado: a Ciência Aberta.

Embasando essa afirmação, traz-se Bourdieu (1983). O autor introduz o conceito de “capital científico” como um conjunto de recursos simbólicos e materiais que um cientista possui e que os coloca em posições mais ou menos dominantes, dependendo do grau de reconhecimento e autoridade que possuem (como sua formação acadêmica, prestígio, acesso a recursos financeiros e posição nas redes acadêmicas e institucionais). Essa perspectiva revela a complexidade do desenvolvimento do campo científico, a partir da visão sobre uma não neutralidade.

Bourdieu reflete sobre como esse campo, para além de um espaço de produção de conhecimento, também se caracteriza como um palco de conflitos que abarca o conteúdo de ideias entre diferentes agentes (cientistas, pesquisadores, instituições) que buscam produzir e legitimar suas visões, métodos e teorias dentro das estruturas de poder e autoridades existentes no campo.

Dessa forma, para subsidiar a proposição de práticas que compõem e apoiam a nova forma de fazer ciência, esta seção explora os fundamentos epistemológicos e multidisciplinares que sustentam a discussão sobre a Ciência Aberta. Inicialmente, são analisadas as relações de poder transversais ao conhecimento, com base nas apreensões teóricas de Foucault e Deleuze, para compreender como as atuais dinâmicas do capitalismo cognitivo influenciam sua produção e apropriação.

Também são apresentadas definições, marcos históricos e diferentes escolas de pensamento que moldam a Ciência Aberta. Em seguida, ao abordar a emergência da cultura do compartilhamento e do (re)uso de dados, identifica-se suas correlações e desafios. Por fim, são examinadas as dimensões, complexidades e perspectivas da

qualidade dos dados na Ciência da Informação tendo por centralidade a confiabilidade e a sua importância para maximizar o potencial de (re)uso dos dados de pesquisa.

Destaca-se que a Ciência Aberta constitui um campo heterogêneo, atravessado por múltiplas frentes conceituais, políticas e epistemológicas. Nesta seção, tais frentes são apresentadas de forma panorâmica, sem a pretensão de esgotá-las ou desenvolvê-las individualmente. A exposição dessas discussões tem como finalidade situar a pesquisa em um campo complexo e multifacetado, sendo o aprofundamento analítico restrito, ao longo da tese, aos aspectos diretamente relacionados ao (re)uso de dados de pesquisa, que constituem o foco central da investigação.

2.1 A CIÊNCIA ABERTA E SEUS DESDOBRAMENTOS PARA A PRODUÇÃO CIENTÍFICA

Essa subseção fornece um panorama abrangente da Ciência Aberta, explorando seus fundamentos teóricos e conceituais, origens, definições, marcos históricos e diferentes perspectivas, com intuito de consubstanciar as discussões que serão desenvolvidas ao longo das próximas subseções. Para tanto, tem-se por embasamento teórico a tríade de Buckland (1991) que vincula dado, informação e conhecimento⁷, o que permite avançar na condução desta pesquisa no intuito de esclarecer a conexão direta da produção de informação e, consequentemente, de seu registro, com a geração de conhecimento e as relações que se instalam entre o saber e o poder, já que “o poder não é somente uma questão teórica, mas algo que faz parte da nossa existência” (Foucault, 2014, p. 120).

Com base nessa relação entre dado, informação e conhecimento, é possível compreender que a produção do saber envolve também disputas e relações de poder que determinam o modo como o conhecimento é construído e compartilhado. Foucault (2002) indica que compreender o conhecimento implica reconhecer as relações de poder que o atravessam e o produzem.

Se quisermos realmente conhecer o conhecimento, saber o que ele é, aprendê-lo em sua raiz, em sua fabricação, devemos nos aproximar não dos filósofos mas dos políticos, **devemos compreender quais são as relações de luta e poder** (Foucault, 2002, p.23, grifo nosso).

⁷ Buckland (1991) entende que o conhecimento é a assimilação de uma informação que foi compreendida pela interpretação de um dado.

Nesse sentido, ao estudar os mecanismos de poder, Foucault (2002) entende que a análise de como ele é exercido deve se dar mediante a observação de uma sociedade e um período determinado. No século XVIII, surge o que o autor denomina como “sociedade disciplinar” que se configurou num momento de ascensão do capitalismo e se firmava a partir de técnicas de poder centradas no controle sobre os corpos dos indivíduos como meio de obter subjugação e controle por parte do Estado. Contudo, na segunda metade daquele mesmo século, Foucault (2005) observa que, subsequente às disciplinas, tem-se o nascimento de uma nova tecnologia de poder: o biopoder.

Logo, depois de uma primeira tomada de poder sobre o corpo que se fez consoante o modo da individualização, **temos uma segunda tomada de poder** que, por sua vez, não é individualizante mas que é massificante, se vocês quiserem, **que se faz em direção não do homem-corpo, mas do homem-espécie** (Foucault, 2005, p.289, grifo nosso).

Nesse cenário, não mais os corpos individuais e sim a vida passou a ser objeto de regulamentação dos estados modernos ocidentais, tanto no âmbito individual, pela subjugação do cidadão aos moldes estabelecidos, quanto no coletivo, com mecanismos de controle da população a partir de análises e decisões sobre, por exemplo, a natalidade e a mortalidade.

Enquanto Foucault (2005) analisa o poder a partir das formas disciplinares e biopolíticas, Deleuze (1992) amplia esse olhar para as transformações tecnológicas e informacionais que caracterizam a sociedade contemporânea. Cabe destacar que as tecnologias de poder não são substituídas umas pelas outras e mantêm conexão entre si.

Em meados do século XX, após a Segunda Guerra Mundial, surge uma nova forma de poder em uma sociedade regida por senhas e *login*, denominada por Deleuze (1992) como “sociedade do controle” na qual os indivíduos passaram a ser vigiados constantemente, embora tenham tido uma pseudo sensação de liberdade, vivenciando regimes de controle que “rivalizam com os mais duros confinamentos” (Deleuze, 1992, p.1).

Essa nova configuração de poder se sustenta em transformações tecnológicas e econômicas que reconfiguram as formas de vigilância e dominação, como o próprio autor detalha.

[...] as sociedades de controle operam por máquinas de uma terceira espécie, máquinas de informática e computadores, cujo perigo passivo é a interferência, e o ativo a pirataria e a introdução de vírus. Não é uma evolução

tecnológica sem ser, mais profundamente, uma mutação do capitalismo (Deleuze, 1992, p.3, grifo nosso).

Essa fase é denominada por Deleuze (1992) como capitalismo de “sobre produção”, onde o que se vende são serviços e o que se compra são ações. Não se trata apenas de “um capitalismo dirigido para a produção, mas para o produto, isto é, para a venda ou para o mercado” (Deleuze, 1992, p.3). Segundo o autor, isso tornou o marketing como o mais novo instrumento de controle social e deslocou o foco da produção material para imaterial.

Observa-se que a instituição da propriedade é intrínseca ao capitalismo, intensificando a concentração de renda e as desigualdades entre indivíduos e comunidades.

Fundar a liberdade individual sobre a exclusão dos outros, que é o principal significado social da instituição da propriedade, é também o vínculo que esta mantém com a **própria forma da sociedade capitalista** (Dardot; Laval, 2015, p. 263, grifo nosso).

Assim, ao reconhecer que a instituição da propriedade está na base das desigualdades estruturais do capitalismo, compreende-se que a apropriação não se restringe aos bens materiais, mas se estende também às dimensões simbólicas e intelectuais.

Nesse movimento de expansão da lógica capitalista para as esferas imateriais, o conhecimento, o saber e a cultura se tornam cada vez mais importantes e originam o que Lazzarato (2010, p. 178) chama de “capitalismo cognitivo”. Conforme destacam Albagli e Maciel (2011, p.20), a geração de valor no capitalismo cognitivo baseia-se na exploração do “trabalho imaterial – aquele que mobiliza informação, conhecimento, imagens, criatividade, afetos, relações” em vez da produção de bens tangíveis.

Se o capitalismo se baseia num sistema econômico que se firma através da apropriação e exploração para gerar lucros e riquezas, o capitalismo cognitivo, que descreve a atual fase do capitalismo, tem o conhecimento como um ativo fundamental e, portanto, objeto de apropriação e disputa. Dessa forma, discussões sobre o acesso e o compartilhamento do conhecimento científico abrem caminho para a emergência de tensões entre a socialização e a privatização do conhecimento.

2.1.1 Capitalismo Cognitivo e a apropriação do conhecimento

Compreende-se que, no momento atual, onde a capacidade cognitiva e seu potencial inovativo também possuem valor, “justifica-se” que o conhecimento tenha se tornado importante fonte de geração de riqueza e, por isso, é o que se busca reter e acumular. “Nele [capitalismo cognitivo], importa menos a quantidade de trabalho e mais a qualidade do trabalho” (Albagli, 2014, p. 220, grifo nosso).

A inovação se dinamiza pela produção de conhecimento a partir do conhecimento, impulsionada pelas interações em redes e territórios. Albagli e Maciel (2011) argumentam que, nesse contexto, além do deslocamento de poder para o controle de elementos imateriais e intangíveis, ele também se expande aos meios utilizados para comunicá-lo.

A partir da década de 1980, diante desse cenário do capitalismo cognitivo, em que o conhecimento é explorado como mercadoria, nota-se uma série de restrições significativas aos periódicos científicos. Ainda que essas revistas tivessem, nos séculos XVII e XVIII, proporcionado o crescimento exponencial da produção e disseminação do conhecimento científico, elas buscaram maximizar seus lucros. Para isso, instituíram assinaturas pagas para acesso ao conteúdo e passaram a exigir elevadas taxas dos pesquisadores como condição para a publicação de seus trabalhos.

Ao passo em que há a proposta de impulsionar a disseminação da informação e, assim, propiciar a promoção de uma possível socialização do conhecimento, da informação e da cultura, identifica-se o aumento dos mecanismos de controle sobre a informação compartilhada, cerceando ainda mais a possibilidade de acesso. Consequentemente, observamos atualmente “uma crescente privatização, mercantilização e retirada de informações que costumavam ser acessíveis” (Hess; Ostrom, 2007, p.12). Ou seja, todo o esforço direcionado à produção, disseminação e compartilhamento do conhecimento, por pesquisadores, acadêmicos e instituições, passou a ser objeto de

[...] um movimento na direção do fechamento do conhecimento levadas a cabo por grandes corporações no sentido de privar parte das pessoas do acesso ao conhecimento como forma de gerar receita financeira (Sayão; Sales, 2024, p. 5).

Esse processo reflete o deslocamento da lógica de abertura, originalmente voltada à democratização do saber, para um modelo orientado por interesses econômicos. A informação e o conhecimento passam a ser tratados como recursos estratégicos, sujeitos à apropriação e ao controle, o que reforça desigualdades e consolida novas formas de concentração e exclusão.

Amplia-se, assim, a concentração (espacial, social, organizacional), o controle e a privatização da informação e do conhecimento considerados estratégicos (expressando-se também na maior concentração de capitais), seja pelo recrudescimento dos mecanismos de proteção de direitos de propriedade intelectual (DPI), desde a década de 1980, seja por outros meios de apropriação do conhecimento público e socialmente produzido (Albagli, 2015, p.23).

Identifica-se então um dos principais pontos de luta e conflito na sociedade informacional: as tensões geradas entre uma vertente que advoga sobre os benefícios, as necessidades e a obrigação da socialização da informação, do conhecimento e da cultura *versus* a que defende a sua privatização.

Nesse contexto, vemos o endurecimento do Direito de Propriedade Intelectual (DPI) beneficiando sobretudo agentes intermediários, em detrimento dos próprios cidadãos.

A propriedade intelectual consiste em mecanismo de escassez artificial (e, logo, rentista) de algo que não se esgota – informações, conhecimento, cultura, subjetividades -, mas que ao contrário, se fertiliza e se reproduz na livre troca e nas interações (Albagli, 2014, p.221).

Em contrapartida, por uma demanda oriunda do próprio campo científico, Hess e Ostrom (2007) identificam que cada vez mais pesquisadores começam a reconhecer que “combinar disciplinas e reunir conhecimento era a única maneira de chegar a um conhecimento mais profundo para uma gestão eficaz de bens comuns (Hess; Ostrom, 2007, p.6, tradução nossa). Isso passou a incrementar um movimento a favor da ampla difusão e do compartilhamento.

Na segunda metade do século XX, surge uma dinâmica global para tornar a pesquisa e seus produtos mais acessíveis, com práticas transparentes e colaborativas, aliadas a um conhecimento científico verificável. O objetivo é aprimorar a qualidade na reprodutibilidade e no impacto da ciência e, conseqüentemente, na confiabilidade das evidências que fundamentam a tomada de decisões e o desenvolvimento de políticas públicas robustas.

Diante das tensões inerentes ao capitalismo cognitivo e à crescente privatização do conhecimento, surge um movimento, conhecido como Ciência Aberta. Esse movimento contrapõe a lógica do capitalismo cognitivo e propõe novos modelos para a produção e disseminação do conhecimento científico. Definições e marcos históricos que delineiam seu desenvolvimento, bem como seus princípios fundamentais, são explorados a seguir.

2.1.2 Surgimento, marcos históricos e definições da Ciência Aberta

A Ciência Aberta envolve múltiplos agentes, como pesquisadores, editores, formuladores de políticas, programadores de sistemas e cidadãos interessados em ciência, a partir de diversos pontos de interesse (e conflitos).

Na tentativa de mensurar sua escala de abertura e de comprometimento social, alguns conceitos têm sido incorporados aos discursos sobre sua prática como: colaboração, equidade, engajamentos, transparência, revisão aberta, acesso democrático, participação da sociedade, multilinguismo e transparência (Sayão; Sales, 2024). Todavia, é preciso destacar que seus possíveis alcances e proposições ainda estão sob contínuos debates e, ademais, encontram alguns obstáculos. Dentre eles, visões que são conflitantes por consequência de suas variadas possibilidades de interpretações.

Nesse sentido, a apresentação desses elementos tem caráter contextual e não exaustivo, sendo mobilizada com o objetivo de evidenciar a complexidade e a heterogeneidade do campo da Ciência Aberta, sem a pretensão de desenvolver analiticamente cada uma dessas frentes.

Dessa forma, a despeito de existir uma vasta e significativa relação de material literário, acadêmico e técnico-científico desenvolvida sobre o tema, traz-se, para fins desta tese, os seguintes marcos de uma das perspectivas da Ciência Aberta, o Acesso Aberto, em consonância com o foco desta tese:

- 1999: “Declaração sobre Ciência e o Uso do Conhecimento Científico”⁸. O documento coloca a ciência à serviço da humanidade como fonte para compreensão aprofundada da natureza e da sociedade, em vista à melhor qualidade de vida e um ambiente sustentável e saudável às gerações presentes e futuras, reconhece o quanto as inovações oriundas do conhecimento científico tem gerado impactos à humanidade, sejam benéficos ou com consequências negativas e constata que a maioria dos benefícios é distribuída de forma desigual, como resultado de assimetrias estruturais entre países, regiões e grupos sociais, e entre os sexos;

⁸ Disponível em: <https://epedagogia.com.br/materialbibliotecaonline/1934Declaracao-sobre-Ciencia-e-o-Uso-do-Conhecimento-Cientifico-Budapeste.pdf>. Acesso em: 20 out. 2025.

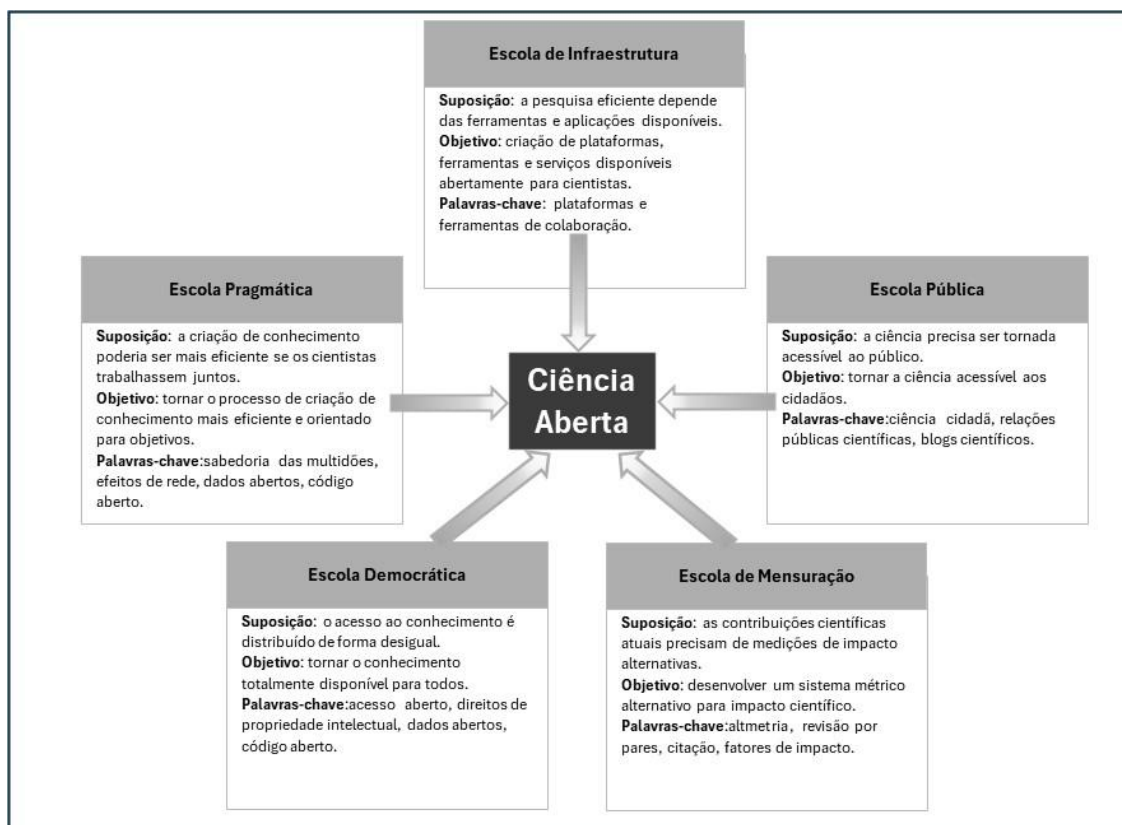
- 2002: “Iniciativa de Acesso Aberto de Budapeste (BOAI)”⁹, que convida governos, universidades, bibliotecas, editores de periódicos, fundações, sociedades acadêmicas, associações profissionais e acadêmicos individuais a se juntarem na tarefa de remover as barreiras ao acesso aberto e “construir um futuro em que a pesquisa e a educação em todas as partes do mundo sejam muito mais livres para florescer” (BOAI, 2002). Esta pode ser considerada a “primeira expressão, da sociedade científica, em prol do acesso aberto e que culminou na declaração que o acesso à literatura científica deveria ser sem barreiras de acesso, livre de custos, imediato e disponível de forma *on-line*” (Henning *et al*, 2019); e
- 2003: as Declarações de Bethesda e de Berlim que buscam incentivar e compartilhar práticas e discussões sobre acesso aberto. Sendo que na Declaração de Berlim, vale destacar, o conceito de acesso aberto teve uma ampliação ao considerar a necessidade de que também fossem tratados para serem acessíveis e utilizáveis os “[...] dados brutos e metadados, fontes originais, representações digitais de materiais pictóricos e gráficos, além de material multimídia acadêmico” (Pampel; Dallmeier-Tiessen, 2014, p. 214, *apud* Henning *et al.*, 2019, p. 391).

Assim, mesmo que a Ciência Aberta tenha tido em seu escopo inicial a reflexão sobre a abertura dos produtos finais de uma pesquisa, ela não se limitou a esse quesito. Esse movimento passou a estimular e promover a análise sobre a importância de que todo o processo científico, que conduz para determinado resultado, também fosse aberto.

Nesse sentido, Fecher e Friesike (2013, p.17) consideram a Ciência Aberta “um termo guarda-chuva que abrange uma infinidade de suposições sobre o futuro da criação e disseminação de conhecimento”. Cabe destacar que, com objetivo de estruturar o discurso geral sobre esse tema, esses autores realizaram uma revisão da literatura que resultou na proposta de cinco escolas de pensamento da Ciência Aberta que possuem perspectivas próprias e, em alguns casos, podem se complementar. São elas:

Figura 1 - Escolas de pensamento sobre Ciência Aberta e suas características.

⁹ Disponível em <https://budapesteopenaccessinitiative.org>. Acesso em: 15 out. 2025.



Fonte: Adaptada pela autora com base em Fecher e Friesike (2013, p. 19).

Como é possível observar na Figura 1, há uma diversidade de discursos sobre o tema, dadas as várias abordagens, compreensões e aplicações do que se pretende por Ciência Aberta. Segundo Fecher e Friesike (2013) isso é compreensível, visto a multiplicidade das partes interessadas que são diretamente afetadas por um ambiente científico em mudanças.

O arcabouço conceitual multifacetado da Ciência Aberta é complexo e, por isso, é relevante apontar que a observação isolada de cada uma de suas faces (prática, teórica, política e epistemológica) é incapaz de revelar em completude sua essência e alcance (Sayão; Sales, 2024). Diante dessa complexidade, entende-se oportuno adotar como parâmetro norteador deste estudo o documento “*Making Open Science a reality*”, da OCDE (2015).

Nesse documento, clarifica-se que, apesar de não haver uma definição formal para a Ciência Aberta, o termo se refere aos esforços de diversos agentes em tornar os resultados primários de pesquisas financiadas com recursos públicos (publicações e dados) acessíveis ao público, em formato digital, sem restrições, ou com a mínima necessária, como meio de acelerar a pesquisa, aumentar a transparência e a colaboração, além de promover a inovação (OCDE, 2015).

Para o desenvolvimento desta tese, adotam-se os discursos compatíveis ao que se compreende por escola democrática, mais especificamente em sua corrente de dados abertos. Tal vertente, tem por foco assegurar a disponibilização de dados primários, coletados ao longo da pesquisa, em um formato que permita “não apenas a sua consulta, mas seu escrutínio e reutilização, de maneira conveniente, em pesquisas posteriores” (Albagli, Clinio e Raychtock, 2014, p. 437). Esta abordagem proporciona transparência e celeridade aos procedimentos científicos, além de motivar a inovação constante no campo, dada a possibilidade de (re)uso dos dados de pesquisa e seus metadados.

Face ao exposto, torna-se relevante trazer para essa análise não somente os benefícios associados à abertura da ciência, mas também os aspectos contraditórios que precisam ser considerados para embasar uma visão mais realista e pragmática, em razão dos discursos sobre o tema estarem fortemente fundamentados nessa dialética.

Sendo assim, após apresentar os marcos históricos e a definição adotada para esta pesquisa, reconhece-se que a implementação da Ciência Aberta enfrenta contradições e complexidades que, ao mesmo tempo que envolvem a busca de benefícios coletivos, encontram-se inseridas em cenários desafiadores para sua efetivação, como desigualdades, assimetrias de poder e tensões entre os ideais desse movimento e as práticas existentes no campo científico.

2.1.3 A Ciência Aberta diante de um campo com contradições e complexidades

Ao considerar o espectro abrangente da Ciência Aberta, é importante expor que a percepção positiva de suas consequências não é unânime entre os autores que a abordam. Apesar dos benefícios inerentes ao aprimoramento científico e social, mediante a possibilidade de ampliação do acesso ao conhecimento produzido e abertamente disponibilizado, alguns sujeitos, direta ou indiretamente, envolvidos na dinâmica de abertura da ciência retratam suas questões e direcionam para uma visão dialética.

Dentre essas questões, destaca-se a visão ainda existente, e muitas vezes eivada de resistência, na comunidade científica quanto ao ato de reusar dados abertos e compartilhados. Ainda que seja considerado que “dados abertos permitem sinergias de investigação e evitam duplicação na coleta de dados” (Fecher; Friesike, 2013, p. 26), muitos pesquisadores compreendem o ato de reusar e compartilhar como uma ação parasitária, conforme observado na edição 16 do *New England Journal of Medicine*, no

editorial intitulado “*Data Sharing*” (Longo; Drazen, 2016), em que se elencam duas questões relacionadas a esse escopo. A primeira é sobre o uso de dados por quem não esteve envolvido na geração e na coleta e o quanto isso pode vir a dificultar o entendimento das escolhas feitas na definição do parâmetro estabelecido para a pesquisa. A segunda refere-se à possibilidade de uma nova classe de pesquisadores que, mesmo não tendo nada a ver com a execução do estudo original, utiliza esses dados para seus próprios fins, seja para “roubar” a produtividade da pesquisa planejada pelos seus reais coletores, seja para refutar o que os pesquisadores originais postularam. Em outras palavras, “há preocupação entre alguns pesquisadores da linha de frente de que o sistema seja tomado pelo que alguns pesquisadores caracterizaram como parasitas de pesquisa” (Longo; Drazen, 2016, p. 27, tradução nossa).

Outra análise do aspecto contraditório envolvido na promoção da abertura da ciência reside na falta de uniformização de fronteiras conceituais da Ciência Aberta, o que gera, inevitavelmente, entraves na práxis científica. Além disso, as incongruências da implantação prática desse movimento global decorrem tanto da heterogeneidade de continentes, países e seus povos, quanto das suas diversas perspectivas e capacidades científicas, influenciadas por inúmeros fatores, sociais, culturais, econômicos e políticos.

A discussão sobre uma ciência equitativa, colaborativa e inclusiva demanda que se considere que o acesso a recursos e infraestruturas se dá de forma heterogênea. Tal análise pode ser feita em um espectro mais amplo, no qual se incluem diversas discrepâncias científicas entre o Norte e Sul globais, por exemplo. Contudo, sem a necessidade de recorrer a essas macros comparações, é possível identificar que tais discrepâncias também se evidenciam em uma análise sob nível nacional.

Nesse sentido, parte-se da compreensão que nem todos os pesquisadores e instituições possuem recursos necessários para atender às premissas da Ciência Aberta. Isso se manifesta tanto por não disporem de ferramentas de análise de dados, quanto pela ausência de acesso às revistas de publicação. Adicionalmente, ainda há a incapacidade de arcar com as altas taxas exigidas para divulgação dos resultados de suas pesquisas em revistas de acesso aberto. Esse cenário consolida um contraste no reconhecimento de impacto e relevância dos estudos e dos pesquisadores (entre aqueles mais e menos favorecidos financeiramente), o que influencia progressões de carreira. Ademais, torna o acesso à ciência menos sustentável e excludente.

Ao compreender que um dos motores do acesso aberto (às publicações e aos dados de pesquisa) é a democratização do acesso ao conhecimento, que possibilita abarcar

leitores de um público mais amplo. É importante destacar que, conforme Borgman (2018), o fato de ter acesso não significa ter conhecimento de domínio e capacidade linguística para explorar tais materiais, embora possa ser considerado um caminho para as questões de equidade.

Nesse aspecto, torna-se basilar a necessidade de que sejam revisitadas as políticas de reconhecimento, incentivo e recompensas nos âmbitos acadêmico e científicos. Atualmente, tais políticas se vinculam diretamente à publicação em periódicos de alto impacto mediante a cobrança de taxas significativas. Esta correlação não oferece estímulos suficientemente relevantes para que os pesquisadores se sintam motivados, ou aptos, a aderirem à abertura de seus estudos, o que os leva a buscarem caminhos alternativos.

Ademais, parte da comunidade científica ainda demonstra desconhecimento sobre o que, no escopo de suas pesquisas, pode ser considerado como objeto da Ciência Aberta. Mesmo quando há esse conhecimento, a ausência de capacitação que aprimore as práticas nessa direção, somada à falta de normativas institucionais ou políticas, impede o estabelecimento de metodologias e a oportunidade de uso de infraestrutura e ferramentas tecnológicas para essa finalidade.

É possível apontar, ainda, que desigualdades estruturais (gênero, raça e classe) representam desafios importantes para a construção de uma ciência equitativa. Tais desigualdades permeiam a baixa representatividade em equipes de pesquisa e a ausência de múltiplos idiomas na comunicação e na disseminação do conhecimento científico.

Em relação a uma eventual abertura dos dados e resultados de pesquisa é preciso considerar os aspectos éticos e legais envolvidos nas questões de direito autoral e propriedade intelectual. Nesse sentido, torna-se primordial o desenvolvimento de planos que resguardecam tais aspectos com a “aplicação dos princípios da razoabilidade e proporcionalidade” (Nascimento, Loureiro, Arruda, 2017, p.3) e o tratamento adequado aos dados sensíveis, como os relativos às pesquisas em saúde e/ou com informações pessoais, por exemplo, que se negligenciados podem causar inúmeros transtornos e consequências legais aos pesquisadores. Em suma, os pesquisadores precisam se sentir seguros e respaldados para aplicar a abertura em suas práticas científicas.

Em se tratando das contradições e complexidades do campo científico, também se ressalta o negacionismo científico e sua influência na visão que a sociedade tem da ciência.

A pesquisadora Carla Almeida (2020), em seu artigo intitulado “*Make the science great again?* O impacto da Covid-19 na percepção pública da ciência”, apresenta reflexões acerca das relações entre ciência e sociedade. Em sua pesquisa, a autora aponta que, embora os dados de uma pesquisa realizada pela *Wellcome Trust*, em 2018, com a participação de cerca de 140 mil pessoas, de 140 países, incluindo o Brasil e outros 17 países da América Latina, não possibilitasse afirmar que a ciência estivesse, àquela época, passando por uma crise de confiança, ao menos deixava nítido que se encontrava em um momento atormentador.

Esse momento, segundo a autora, se relacionava tanto ao fato dos substanciais cortes de recursos destinados às pesquisas científicas, quanto pela ebulição “de toda a sorte de ideias, teorias e movimentos que se contrapõem diretamente aos melhores dados produzidos pela ciência ao longo de séculos” (Almeida, 2020, p. 4).

No que se refere a percepção pública da ciência, desponta o movimento de negacionismo que, segundo Villela e Selles (2020), enquanto uma retórica do conservadorismo, utiliza a mentira de forma conveniente para atacar consensos científicos em prol de interesses político-ideológicos baseado na defesa de privilégios, levantando dúvidas a fim de produzir falsas controvérsias científicas e narrativas antifrustrações.

Apona-se, nessa mesma direção, para o espessamento de movimentos negacionistas, principalmente pelo advento das redes sociais e as diversas implicações éticas, morais, políticas, sociais e educacionais envolvidas em seu uso descompensado. Perini (2019), ao avaliar os problemas que o cenário de *fake news* e negacionismo impõe aos pesquisadores e cientistas, destaca o desafio para as democracias, tanto à nível individual, no qual os terraplanistas, por exemplo, mantêm um comportamento de reforçar suas crenças como um meio de defender seu papel identitário na comunidade contra a base evidencial, quanto à nível coletivo, visto que algumas decisões públicas dependem que o conhecimento científico seja aceito e considerado.

Dessa forma, movimentos negacionistas, tais como: os “negacionistas do clima”, os antivacina e os terraplanistas, mesmo que possam parecer inofensivos, instigam a descredibilidade dos métodos, processos e resultados científicos e, por consequência, impactam diretamente na vida das pessoas (Almeida, 2020). A autora reforça ainda que “mais do que desconfiança, parece haver na sociedade brasileira distanciamento, apatia e indiferença no que diz respeito à prática cotidiana da ciência” (Almeida, 2020, p. 4).

Portanto, na tentativa de combater o que se tem hoje por negacionismo científico é preciso uma aproximação da sociedade com a ciência. Para isso, Villela e Selles (2020) apontam o reforço e valorização do ensino da ciência desde as escolas, a solidariedade acadêmica e científica, a ampliação da divulgação científica, da educação não formal e da popularização da ciência são ações que devem ser firmadas.

[...] neste empreendimento **de encarar o negacionismo científico como algo a ser combatido** precisamos de mais ensino de Ciências nas escolas, mais divulgação científica, mais educação não formal, mais popularização da ciência de um modo que nos aproxime fortemente da raiz da palavra povo. Precisamos também de menos: menos prescrições, menos compromissos com o mercado, menos antagonismo entre as áreas, menos competição e mais solidariedade acadêmica e científica para dialogarmos com as escolas em processos de escuta (Villela; Selles, 2020, p. 1742, grifo nosso).

Diante da conjuntura apresentada, em que o conhecimento é capitalizado e a descrença na ciência é estimulada por grupos, muitas vezes induzidos por ideologias políticas e/ou religiosas, urge a necessidade de fomentar as reflexões e práticas da Ciência Aberta. Com o objetivo de ampliar o acesso ao conhecimento e tornar a ciência mais transparente, inclusiva e sustentável, a Ciência Aberta pode ser um caminho viável para integrar sociedade e comunidade científica.

A abertura do conhecimento, de uma forma geral, significa mais do que ser apenas uma oposição a sociedades fechadas. Ela se

[...] opõe a fatos enganosos (sobre questões sociais), dados ilegíveis (de pesquisas científicas), informações privatizadas (mantidas por plataformas de tecnologia) e, claro, documentos retidos (sobre atos governamentais ou corporativos) (Open Knowledge Foundation, 2004, tradução nossa¹⁰).

Além disso, em busca de avanços coletivos, cada vez mais elementos vêm sendo reconhecidos para comporem a conceituação do que se entende por Ciência Aberta, como o que se compreende como Ciência Cidadã. Nessa direção, Hecher *et al.* (2018) consideram a Ciência Cidadã um campo de pesquisa “incomum” em desenvolvimento que tem criado conscientização entre seus praticantes sobre a relevância de seus impactos tanto para o campo social quanto, e cada vez mais, para o político.

De acordo com Albagli e Rocha (2021), o termo Ciência Cidadã

[...] tem sido usualmente utilizado para designar um conjunto de ações que promovem a contribuição de não cientistas para a ciência, na expectativa de melhorar a qualidade dos resultados e reduzir os custos da pesquisa, além de

¹⁰ Disponível em: <https://okfn.org/en/who-we-are/>. Acesso em: 20 out. 2025.

ampliar o engajamento público na ciência. Tem consistido principalmente na colaboração cidadã voluntária e ativa em atividades de coleta e interpretação de dados relevantes para o avanço da ciência, incluindo o envio de imagens, sons e outros tipos de registros, bem como facilitando seus usos em questões de importância social e territorial (Albagli; Rocha, 2021, p. 490).

Dois exemplos ilustram a aplicação e a relevância da Ciência Cidadã no contexto da pesquisa e da gestão ambiental. O primeiro é um estudo qualitativo realizado por Altman et al. (2023) com profissionais da área ambiental. Essa pesquisa demonstra que uma comunicação científica inclusiva é fundamental para o avanço da alfabetização em saúde ambiental e para o apoio à gestão do meio ambiente. Os autores destacam que a compreensão limitada de processos científicos, a necessidade de tempo para construir confiança e a importância de incorporar o acesso no desenho dos programas são fatores determinantes para o engajamento público.

O segundo exemplo é o projeto brasileiro “Save Brasil”¹¹, que instituiu, em 2014, o “Projeto Cidadão Cientista” em quatro estados (Bahia, Rio de Janeiro, São Paulo e Paraná). Alinhada aos conceitos de Ciência Cidadã, a iniciativa convoca e capacita observadores de aves experientes e novos voluntários para colaborar no monitoramento da vasta biodiversidade ornitológica brasileira. Essa ação direta da população permite estimar tendências populacionais de espécies e avaliar a eficácia de medidas de conservação empregadas no país.

Essas ações apontam não apenas para a importância do compartilhamento de dados, mas também para a colaboração entre pesquisador e cidadão como subsídio para o enfrentamento ao negacionismo científico.

Dessa forma, é possível observar um contínuo crescimento em direção à promoção de abertura da ciência a partir de, dentre outros elementos, estímulos ao compartilhamento de dados de pesquisa, o que possibilita seu uso em outros contextos e incrementa inovações e avanços científicos e sociais.

2.2 A EMERGÊNCIA DA CULTURA DO COMPARTILHAMENTO E DO (RE)USO DE DADOS

¹¹ Disponível em: [Cidadão Cientista | SAVE Brasil](#) Acesso em: 04 ago. 2025.

Na Ciência Aberta, os termos compartilhamento e (re)uso caminham lado a lado, sendo “um par de conceitos que se complementam” (Caregnato *et al.*, 2021, p. 196). Contudo, “o valor do compartilhamento de dados de pesquisa científica é amplamente apreciado, mas os fatores que dificultam ou estimulam a reutilização de dados permanecem pouco compreendidos” (Curty, 2019).

Ainda que o compartilhamento de dados seja amplamente defendido, o (re)uso desses dados apresenta desafios e complexidades que precisam ser cuidadosamente considerados. Dessa forma, explorar as relações entre compartilhamento e (re)uso, analisando fatores relacionados, inclusive às dificuldades em definir e medir essa prática, e reconhecer que nem todos os dados compartilhados são passíveis de (re)uso, configura-se como uma proposta de interesse para esta pesquisa.

2.2.1 Compartilhar e (re)usar: relações e desafios imbricados

A partir do momento em que a Organização Mundial da Saúde (OMS) declarou como pandemia a infecção causada pelo coronavírus SARS-Cov2 (conhecida por COVID-19), em março de 2020, evidencia-se, em escala global, que a abertura e o compartilhamento de dados de pesquisa por pesquisadores, instituições e laboratórios de diversos segmentos e países poderiam ser os propulsores dos urgentes tratamentos médicos e produção de vacinas¹².

No que se refere à relação entre compartilhar e (re)usar dados de pesquisa, LaFlamme, Poetz e Spichtinger (2022) esclarecem que, apesar dos altos investimentos para garantir a abertura e o compartilhamento de uma enorme quantidade de dados, ainda se sabe pouco sobre o que determina se esse volume de dados compartilhados é realmente usado. Nesse sentido, no que tange à utilização do termo “reuso” para os dados de pesquisa, Van de Sandt *et al.* (2019) apontam que se trata de um conceito cujo significado ainda não parece estar definido e do qual a compreensão varia entre disciplinas e indivíduos.

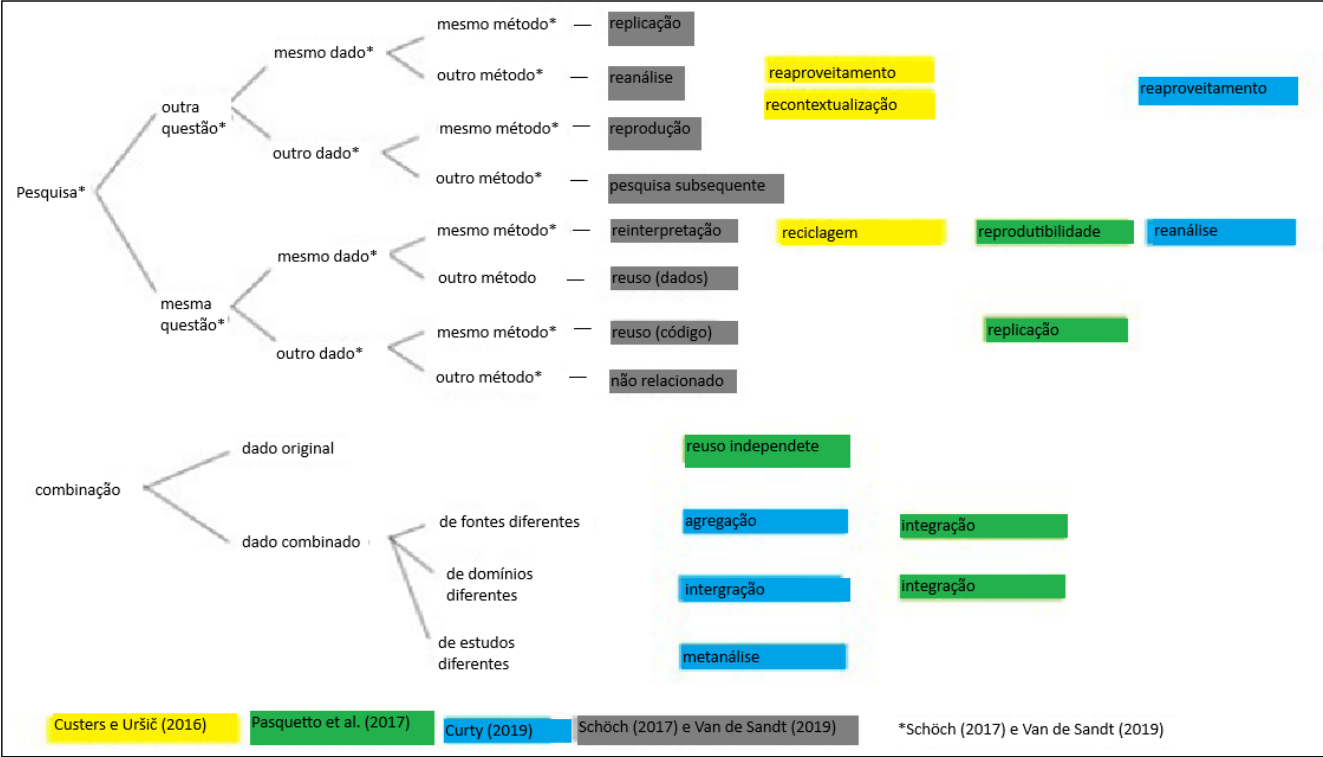
¹² É relevante ressaltar que a percepção dos benefícios de uma ciência conduzida abertamente já vinha sendo analisada no âmbito de pesquisas brasileiras. Pode-se destacar, por exemplo, a tese de doutorado de Vanessa Arruda Jorge (2018) sob o tema “Abertura e compartilhamento de dados para pesquisa nas situações de emergência em saúde pública: o caso do vírus Zika” que traz à luz da reflexão a importância do compartilhamento mas também a necessidade de uma maior abertura dos dados de pesquisa principalmente em casos de emergência de saúde pública, como o que foi por ela avaliado, onde a partir do cruzamento de informações foi possível estabelecer relações entre o vírus Zika e os casos de microcefalia e outras alterações neurológicas em bebês nascidos de mulheres que haviam contraído a doença.

Em seu artigo intitulado “*The definition of reuse*”, Van de Sandt *et al.* (2019) identificam (a partir de um levantamento realizado nas bases LISA, Google Scholar e Scopus que culminou na seleção de 20 (vinte) trabalhos de variadas autorias, anos e formações disciplinares) que quase todas as definições do termo (re)uso compartilhavam um elemento comum: o pensamento do reaproveitamento.

Cabe destacar que não há consensos na definição e categorização do termo (re)uso. Tal imprecisão advém da ampla abrangência do conceito, que engloba variáveis e especificidades que refletem as diversas possibilidades e interpretações no contexto dos dados e métodos da pesquisa.

Dessa forma, identifica-se que outros termos se correlacionam ao (re)uso como, por exemplo: replicabilidade, reprodutibilidade, reaproveitamento, replicação, reanálise, agregação, dentre outros. A partir da utilização de dois critérios gerais: tipos de (re)uso determinados pelo contexto da pesquisa e tipos de (re)uso determinados pela necessidade de combinação dos dados, Caregnato *et al.* (2021) desenvolvem a Figura 2.

Figura 2 – Categorização dos termos relacionados ao conceito de reuso de dados de pesquisa



Fonte: Caregnato *et al.* (2021, p. 204).

Com base na Figura 2, é possível verificar as inúmeras ramificações terminológicas que podem ser vinculadas ao ato de reuso dos recursos de pesquisa. Observa-se que autores como Pasquetto *et al.* (2017), fazem algumas distinções entre os diversos termos para embasar a necessidade de investimentos de recursos específicos a cada uma das categorias pontuadas, a partir do desdobramento do termo reuso. Os autores (2017) definem reuso como a utilização de dados por terceiros em pesquisas distintas das originais.

No entanto, embora existam distinções entre “uso” e “reuso”, Van de Sandt *et al.* (2019) argumentam que essa distinção é inadequada ao considerarem que os modelos de utilização do termo "reuso" refletem um padrão linear baseado na pesquisa em papel. Contudo, os processos científicos têm sido orientados por dados e, portanto, mais complexos e dinâmicos. Dessa forma, concluem que "uso" e "reuso" descrevem a mesma ação, independentemente do tempo, propósito, características ou usuário.

Tendo em vista essa gama de categorizações e classificações do termo (re)uso, adota-se para esta tese a definição de que “o reuso dos recursos de pesquisa é igual ao seu uso” (Van de Sandt *et al.*, 2019, p.15, tradução nossa). Esta definição, segundo esclarecem as autoras, pode apoiar um entendimento comum que independa de disciplinas ou abordagens de pesquisa, propiciando assim que futuras estratégias de Ciência Aberta sejam moldadas.

Ademais, ainda que as distinções sobre a adoção do termo e seu significado sejam menos relevantes no contexto atual, aspectos relativos à sua compreensão em um contexto mais pragmático, carecem de reflexões, debates e consensos. Uma das nuances que se julga relevante demonstrar é a percepção sobre o que motiva pesquisadores a serem usuários desses dados.

Em uma revisão de literatura sistemática, que contou com 32 estudos publicados entre 2004 e 2019, Zuiderwijk, Shinde e Jeng (2020) observaram a indicação de fatores que impulsionam ou inibem o compartilhamento e uso de dados de pesquisa compartilhados abertamente. Quanto ao (re)uso, especificamente, são listadas dez categorias de fatores¹³: a) formação do pesquisador; b) experiência e habilidades do pesquisador; c) motivações pessoais e intrínsecas, d) esforço; e) expectativa; f) confiança;

¹³ Quanto a isso, os autores disponibilizaram uma planilha detalhada em <https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0239283.t007>. Acesso em: 15 out. 2025.

g) influência e afiliação social; h) condições facilitadoras; i) requisitos e obrigações formais, legislação e regulamentos; e j) características dos dados.

Dentre outras discussões abordadas no artigo, as autoras (2020) concluem que os principais impulsionadores estão relacionados, principalmente, a motivações pessoais e intrínsecas e ao desempenho esperado dos pesquisadores. Já os inibidores estão relacionados ao esforço e às características dos dados.

A importância de estudos como esse reside na premissa da Ciência Aberta que preconiza que dados de pesquisa sejam acessíveis a todos com o mínimo de restrições possíveis. A defesa central em relação ao compartilhamento desses dados é que este ato torna o processo científico mais impactante e eficiente. Contudo, mesmo que a abertura e o compartilhamento de dados sejam realizados a contento, subsiste uma questão fundamental: como saber se eles realmente serão (re)usados e, consequentemente, produzirão valor adicional para a ciência e a sociedade?

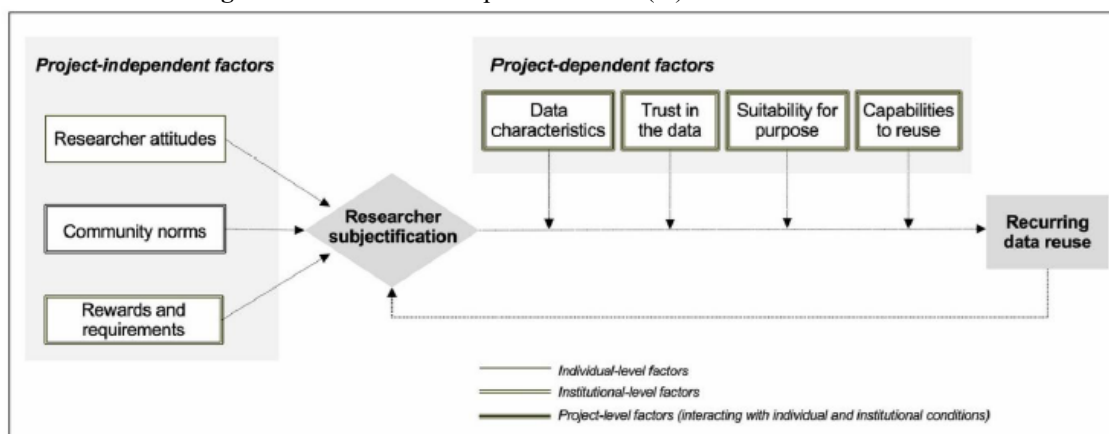
Na tentativa de responder a essa pergunta, LaFlamme, Poetz e Spichtinger (2022) conduziram uma revisão da literatura para identificar os fatores que influenciam o (re)uso de dados de pesquisa. Os autores constataram que esses fatores se agrupam em níveis individuais e institucionais.

No âmbito individual, o (re)uso é influenciado pela percepção do pesquisador de que essa prática seja menos dispendiosa em termos de tempo, esforço e custos do que a produção de novos dados e por suas experiências prévias com o (re)uso. Ou seja, se o pesquisador tiver experienciado previamente uma ocorrência fracassada, dificilmente vai optar pelo (re)uso novamente. No âmbito institucional, que envolve as condições materiais, dependências instrumentais de determinada área científica e suas práticas normativas (frequentemente sujeitas a mudanças), o incentivo à reutilização de dados é notadamente menor.

Os autores, ao referenciar a revisão de Zuiderwijk, Shinde e Jeng (2020), apontam para a falta de estruturação e interação entre as categorias de fatores e a pouca análise das relações de mediação e dinâmicas do processo. LaFlamme, Poetz e Spichtinger (2022) propõem a subjetivação, definida como o “processo multinível por meio do qual os pesquisadores passam a se ver como reusuários e intermediários de dados” (LaFlamme; Poetz; Spichtinger, 2022, p. 7, tradução nossa) como um conceito promissor para entender como os fatores influenciam o (re)uso.

Os autores realizaram entrevistas semi-estruturadas e, a partir dos resultados, propuseram um modelo que integra os fatores individuais e institucionais, considerando a subjetivação do pesquisador como um mecanismo influenciador no (re)uso recorrente de dados (ou seja, por mais de vez).

Figura 3 - Modelo de comportamento do (re)uso recorrente de dados



Fonte: LaFlamme, Poetz e Spichtinger (2022, p. 7).

LaFlamme, Poetz e Spichtinger (2022) concluem que uma mudança duradoura no comportamento dos pesquisadores em relação ao (re)uso de dados requer que instituições de pesquisa e decisores políticos criem “ambientes de subjetivação” que incentivem novos autoentendimentos compatíveis com essa prática, capacitando os pesquisadores a questionar as utilizações do (re)uso no contexto científico atual.

Além disso, o estudo destaca a preferência dos pesquisadores por intermediários, como centros de dados ou repositórios, devido à percepção de que estes oferecem um processo de revisão interna que garante um nível básico de qualidade, o que revela que os pesquisadores “encaram o (re)uso como uma forma de apostar sua reputação científica no trabalho de outra pessoa” (LaFlamme; Poetz; Spichtinger, 2022, p. 15, tradução nossa).

A análise dos autores aponta para uma questão relevante: a viabilidade e a efetividade do (re)uso de dados de pesquisa dependem, fundamentalmente, da qualidade dos dados compartilhados. Somente ao garantir a confiabilidade desses dados, os pesquisadores serão incentivados a realizar o (re)uso.

Assim, a próxima subseção aborda a questão da qualidade dos dados de pesquisa, explorando diferentes perspectivas sobre o tema e a importância da confiabilidade para o seu (re)uso.

2.3 DIMENSÕES, COMPLEXIDADES E PERSPECTIVAS DA QUALIDADE DOS DADOS

Embora o aspecto qualitativo seja um pilar fundamental para a prática efetiva da Ciência Aberta e haja ampla concordância sobre a necessidade de disponibilizar dados de pesquisa com alta qualidade, identifica-se na literatura uma notável falta de consenso em relação aos critérios para definição e avaliação dessa qualidade.

Estudiosos da temática já realizaram pesquisas com o intuito de propor definições sobre os atributos necessários, bem como metodologias e estratégias que pudessem embasar a avaliação da qualidade dos dados de pesquisa (como Strong, Lee; Wang, 1997; Wang; Strong, 1996; Wand; Wang, 1996). Todavia, Sadiq e Indulska (2017) apontam que incertezas sobre a qualidade dos dados disponibilizados persistem e que isso oportuniza uma ameaça ao valor que pode ser gerado a partir deles.

Aliado a isso, identifica-se que a imprecisão apontada pelos autores é agravada pela falta de consenso na terminologia, desde a utilização de termos como “dados” e “informação” de forma sinonímia (Oleto, 2006) até outras abordagens relacionadas à qualidade dos dados a partir de diferentes perspectivas.

Assim, nas próximas subseções identifica-se a diversidade de pontos de vista, apresentando as diferentes dimensões, atributos e fatores considerados relevantes para a qualidade dos dados, desde características intrínsecas aos dados até questões contextuais e de percepções do usuário, com objetivo de evidenciar a complexidade e a multidimensionalidade desse tema.

2.3.1 A centralidade da confiança como atributo essencial da qualidade de dados na Ciência Aberta: metadados de proveniência e a percepção do usuário

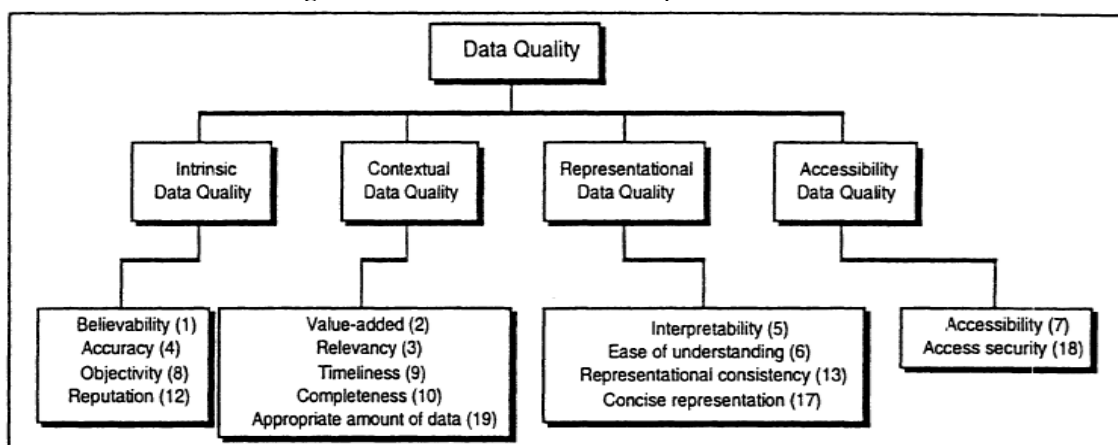
Como demonstrado anteriormente, dados abertos, por si só, não garantem que eles sejam (re)usados. Nesse escopo, a confiabilidade desses dados subsidia um (re)uso seguro, fortalecendo, assim, os objetivos da Ciência Aberta.

Strong, Lee e Wang (1997) analisam a qualidade dos dados sob a ótica dos sistemas de informação e identificam três agentes envolvidos: produtores, curadores e consumidores. Os produtores geram e fornecem o material bruto (a informação); os curadores fornecem e controlam os recursos para armazenamento, manutenção e fixação da informação; e os consumidores usam essa informação para realizar suas funções. A constatação dos autores indica que a participação de todos é aspecto elementar para que os dados tenham qualidade.

Wang e Strong (1996, s.p., tradução nossa), por sua vez, já haviam proposto que a qualidade é definida pela “adequação para uso pelos seus consumidores”. Essa abordagem enfatiza as dimensões qualitativas sob o ponto de vista do usuário. A qualidade de dados, portanto, não deve se fixar teoricamente ou com base na experiência, demandas e conhecimentos de seus fornecedores ou custodiadores. Segundo os autores, embora os dados não sejam “comprados” pelos seus consumidores, são eles que escolhem se os usarão ou não, e esse uso é o que pode atestar, na prática, a qualidade desses dados.

Diante disso, concluíram que “dados de alta qualidade devem ser intrinsecamente bons, contextualmente apropriados para a tarefa, claramente representados e acessíveis ao consumidor de dados” (Wang; Strong, 1996, s.p., tradução nossa) e representaram sua resolução em uma estrutura conceitual da qualidade de dados, conforme Figura 4.

Figura 4 – Estrutura conceitual da qualidade de dados



Fonte: Wang; Strong (1996, p. 20).

A estrutura de Wang e Strong (1996), que categoriza a qualidade em dimensões intrínseca, contextual, representacional e de acessibilidade, se alinha com os princípios da Ciência Aberta ao destacar a importância da acessibilidade e da representação clara para facilitar a compreensão e (re)uso de dados de pesquisa, sem perder de vista sua

credibilidade, acurácia, reputação, completude, relevância, entre outros fatores intrínsecos e contextuais dos dados.

Além dessa perspectiva, Wand e Wang (1996) também contribuem para essa discussão com as dimensões da qualidade de dados e suas relações internas e externas, conforme quadro abaixo:

Quadro 3– Dimensões da qualidade de dados e suas relações internas e externas

Dimensões	
Visão Interna (projeto, operação)	Relacionada aos dados: acurácia, confiabilidade, atualidade, completude, consistência e precisão
	Relacionada ao sistema: confiabilidade
Visão Externa (uso, valor)	Relacionada aos dados: atualidade, relevância, com conteúdo, importância, suficiência, usabilidade, clareza, concisão, imparcialidade, informativo, com detalhamento, com quantitativo, escopo, interpretabilidade, compreensibilidade
	Relacionada ao sistema: oportunidade, flexibilidade, formato e eficiência.

Fonte: Adaptado pela autora com base em Wand; Wang (1996, p. 92).

Ao analisarem o impacto da má qualidade dos dados e a ausência de um acordo geral sobre suas dimensões, que variam conforme o uso pretendido, dividindo os conceitos em visões internas e externas, Wand e Wang (1996) apontam:

a) a percepção do usuário (usuário), que seria a fonte mais importante de metadados de Qualidade Informacional (QI) em virtude de que ele é quem decide se alguma informação é qualitativamente boa ou não. Por serem considerados fornecedores de valiosas informações, em especial para critérios subjetivos, como *compreensão*, por exemplo;

b) a informação em si (fonte), sendo considerada para muitos critérios a origem das pontuações de QI, fornecendo critérios voluntariamente, como o preço, ou involuntariamente, como a integridade; e

c) o processo de acesso à informação (processo), como por exemplo o tempo de resposta, pode ser avaliado automaticamente sem a contribuição dos outros dois fatores. Diante disso, os autores propuseram métodos para avaliação de cada um dos critérios que identificaram, conforme figura abaixo.

Figura 5 - Classificação dos critérios de metadados de qualidade da informação

Classe de avaliação	Critérios de QI	Método de avaliação
critérios do usuário	credibilidade	experiência do usuário
	representação concisa	amostragem do usuário
	interpretabilidade	amostragem do usuário
	relevância	avaliação contínua do usuário
	reputação	experiência do usuário
	compreensibilidade	amostragem do usuário
	valor agregado	avaliação contínua do usuário
critérios do fonte	completude	análise, amostragem
	suporte ao cliente	análise, contrato
	documentação	análise
	objetividade	contribuição de especialistas
	preço	contrato
	confiabilidade	avaliação contínua
	segurança	análise
	atualidade	análise
	verificabilidade	contribuição de especialistas
critérios do processo	precisão	amostragem, técnica de limpeza
	quantidade de dados	avaliação contínua
	disponibilidade	avaliação contínua
	representação consistente	análise
	latência	avaliação contínua
	tempo de resposta	avaliação contínua

Fonte: Adaptada pela autora com base em Naumann e Rolker (2005, p. 7).

Assim, compreende-se que, da mesma forma que Wang e Strong (1996) apontam, Wand e Wang (1996) indicam a importância da análise de um contexto que não se limita à observação do dado em si para que seja possível verificar sua qualidade.

Já Lee e Strong (2003) expõem que, se o objetivo dos processos de geração de dados é produzir dados para consumidores, ter uma alta qualidade significa que dados são adequados para que sejam por eles usados. Nessa perspectiva os autores concentraram-se em cinco dimensões da qualidade dos dados, focando no dado em si:

- 1) Acessibilidade (os dados estão disponíveis ou podem ser facilmente e rapidamente recuperáveis?);
- 2) Relevância (os dados são aplicáveis e úteis para a tarefa em questão?);
- 3) Atualidade (os dados estão suficientemente atualizados para a tarefa em questão?);
- 4) Integridade (os dados não estão faltando e são de amplitude e profundidade suficientes para a tarefa em questão?); e
- 5) Precisão (os dados estão corretos e livres de erros e precisão?).

Segundo os autores, cada uma delas é uma meta de desempenho no processo de produção de dados no qual o objetivo geral é “produzir dados precisos, completos e atuais que sejam acessíveis aos consumidores de dados e relevantes para as suas tarefas” (Lee; Strong, 2003, p. 18, tradução nossa).

Cumprido esclarecer então que, pelo fato de a Ciência Aberta defender a abertura e o compartilhamento de dados, devem ser relevados os fatores envolvidos na possibilidade de que os dados tenham alta qualidade para serem não apenas acessíveis, mas também úteis e confiáveis para o (re)uso, promovendo a aceleração do progresso científico e aumento na transparência da pesquisa.

Assim, sob a ótica de Oletto (2006), a Ciência da Informação também contribui para o debate sobre a qualidade dos dados, apresentando uma dicotomia entre abordagens que avaliam os sistemas de informação a partir do próprio sistema (ênfase na informação como produto, com dimensões como abrangência, acessibilidade, atualidade, confiabilidade, objetividade, precisão e validade) e abordagens centradas no usuário (ênfase em aspectos subjetivos, com atributos como adequação da indexação e classificação, eficácia e eficiência da recuperação, impacto, relevância, utilidade, valor esperado, valor percebido e valor de uso).

Essa dicotomia reflete a tensão entre a qualidade intrínseca dos dados e sua relevância para o usuário e demonstra que tanto a qualidade técnica dos dados quanto sua adequação para diferentes usuários e propósitos são fatores relevantes.

De Lusignan *et al.* (2011) clarificam, em um estudo sobre dados clínicos, que “não há um consenso internacional sobre como descrever a qualidade dos dados e sua usabilidade para pesquisa” (p. 112, tradução nossa). Os autores apontam que, inicialmente, a qualidade de dados era definida em termos quantitativos a partir de sua completude e precisão e, posteriormente, foram adicionados conceitos de *currency* (moeda).

Em seguida, conforme os autores (2011), essa definição passou a ser dada pelo valor preditivo positivo, sensibilidade e especificidade. Concluem, então, que tem havido tentativas por avaliar a qualidade de dados de uma maneira mais analítica, com comparações por etapa e a análise de todo o processo com planejamento e procedimentos sistemáticos antes, durante e depois da coleta de dados.

Para esses autores, a descrição consistente de todo o processo, desde o registro de dados até o resultado da pesquisa, embora seja uma ação complexa, viabiliza uma melhor compreensão da sua validade e, por isso, deve ser formalizada no esquema de metadados. Assim, tais metadados devem ser publicados junto aos dados. Caso contrário, corre-se o risco de que estes sejam mal interpretados ao serem analisados por pessoas que não compreendem o contexto ao qual foram registrados.

Adicionalmente, De Lusignan *et al.* (2011) fortalecem a importância da proveniência enquanto um tipo de metadado que registra a história dos dados, o que inclui sua origem, alterações e possíveis versionamentos, tendo cinco funções em potencial:

- 1) compreender, e em última análise, melhorar a qualidade dos dados; 2) fornecer uma trilha de auditoria dos dados (o que abrange a linhagem); 3) gerar receitas de replicação para permitir a reprodução do processo; 4) fornecer atribuição e propriedade dos dados (uma característica importante do *pedigree*); e 5) possibilitar a descoberta de novas informações sobre o processo¹⁴ (De Lusignan *et al.*, 2011, p. 115, tradução nossa).

Dessa maneira, os autores ressaltam a limitada atenção dada à qualidade da informação em dados semiestruturados, não estruturados e multimídia. Também identificam desafios como a falta de experimentação metodológica, a necessidade de analisar a composição da qualidade em dados agregados em sistemas *web* e repositórios, e a carência de diretrizes metodológicas para abranger dimensões mais amplas da qualidade, como desempenho, disponibilidade, segurança e acessibilidade, bem como suas interdependências.

Para a análise das metodologias para avaliação e melhoria da qualidade dos dados, a definição de dimensões e métricas é crucial e elas podem se referir à extensão, aos valores ou ao esquema dos dados (De Lusignan *et al.*, 2011).

Ainda que a literatura forneça uma extensa classificação das dimensões da qualidade, Batini *et al.* (2009) apontam para a existência de discrepâncias devido à natureza contextual da qualidade e propõem um conjunto básico de dimensões: a) precisão, que segundo Wang; Strong (1996) se relaciona com a medida em que os dados são corretos, confiáveis e certificados; b) completude (grau em que os dados descrevem o conjunto correspondente de objetos do mundo real, relacionada a valores nulos); c)

¹⁴ Os autores propõem o uso de “proveniência” como termo abrangente, tendo os termos “pedigree” (termo com mais ênfase na qualidade e confiabilidade dos dados) e “linhagem” (estudo de onde os dados vem) como subordinados. Já a rastreabilidade é definida como a capacidade de manter a identidade de um produto e sua origem.

consistência (violações de regras semânticas); e d) atualidade (atualização dos dados ao longo do tempo).

No campo da indústria farmacêutica e biotecnologia, a qualidade dos dados tem sido uma preocupação central. Em 1990, a *Food and Drug Administration* (FDA) criou o princípio ALCOA, um acrônimo para *Atributable* (atribuível), *Legible* (legível), *Contemporaneous* (contemporâneo), *Original* (original) e *Accurate* (acurado), com o objetivo de garantir a integridade dos dados. Posteriormente, a Agência Europeia de Medicamentos (EMA) expandiu esse princípio, adicionando quatro atributos centrados na gestão dos dados: *Complete* (completo), *Consistent* (consistente), *Enduring* (duradouro) e *Available* (disponível), resultando no chamado ALCOA+. Mais recentemente, o termo *Traceable* (rastreável) foi incorporado, enfatizando a importância de rastrear os dados ao longo de seu ciclo de vida, o que levou à criação do acrônimo ALCOA++.

A respeito da percepção do usuário, Calazans (2008) argumenta que a avaliação da qualidade dos dados é baseada na própria percepção do consumidor e no atendimento de suas necessidades. Oleto (2006), por outro lado, observa a dificuldade dos usuários em distinguir os conceitos teóricos da qualidade da informação e sugere que a percepção da qualidade é muitas vezes intuitiva e baseada no senso comum.

Essa diversidade de perspectivas destaca a natureza contextual da qualidade dos dados, que é influenciada pelas necessidades e expectativas dos usuários, bem como pelo domínio de aplicação dos dados.

Em um panorama geral, todos os autores mencionados até o momento na seção 2 desta tese, de alguma forma, abordam questões, direta ou indiretamente, relacionadas à confiança e credibilidade dos dados.

Os termos *believability* (credibilidade) e *accuracy* (precisão) (Wang e Strong; 1996), *accuracy* (acurácia) e *reliability* (confiabilidade) (Wand e Wang; 1996), *integrity* (integridade) e *accuracy* (precisão) (Lee e Strong; 2003) são apresentados como dimensões de qualidade. Já Naumann e Rolker (2000), Oleto (2006) e Calazans (2008) abordam a percepção do usuário como um fator importante na qualidade à medida que confiança depositada nos dados e a facilidade de compreendê-los influenciam a sua perspectiva de qualidade. Adicionalmente, De Lusignan *et al.* (2011), ao reconhecerem a importância da descrição de todo o processo, apontam a proveniência como um elemento relevante para a avaliação de confiabilidade.

Nesse sentido, Bettivia, Cheng e Gryk (2022) apontam que a proveniência descreve o que aconteceu (proveniência retrospectiva), o que poderia acontecer (proveniência subjuntiva) ou o que acontecerá (proveniência prospectiva). Assume-se, portanto, para esta tese que uma das dimensões da qualidade dos dados de pesquisa é a sua confiabilidade, que perpassa a perspectiva do usuário e tem como um de seus elementos essenciais o registro de proveniência retrospectiva dos dados, o que deve ser realizado em metadados.

Dessa maneira, visto que a confiança é a “credibilidade ou conceito positivo que se tem a respeito de alguém ou de algo”, (Michaelis, [s.d]), entende-se que a análise sobre a confiabilidade também precisa levar em perspectiva a percepção do sujeito (crença ou sentimento) para o qual o compartilhamento e disponibilização de dados se direciona.

No entanto, mesmo que o foco seja no usuário, Lee e Strong (2003) destacam que os coletores dos dados com conhecimento sobre seu processo de produção contribuem para que eles sejam gerados com melhor qualidade visto que, no geral, o conhecimento deles é mais crítico do que o dos “guardiões de dados”, o que implica que os metadados sejam registrados de forma mais completa possível por esses agentes. Os autores enfatizam também a necessidade dos metadados serem publicados junto aos dados.

Isso posto, considera-se que os aspectos envolvidos na avaliação da qualidade dos dados são um tema que abarca diversas abordagens, contudo, reconhece-se que a finalidade inerente à disponibilização de determinado conjunto de dados é essencial para identificar os atributos qualitativos das informações que precisam estar registradas e, consequentemente, qual nível de qualidade é necessário ou possível de ser atingido.

Para avaliar a completude dos dados é fundamental, conforme sugerido por Batini *et al.* (2009), que se defina de maneira explícita o rol mínimo de informações julgadas basilares sobre os dados e seus respectivos processos de pesquisa. Somente após essa definição, torna-se possível, na perspectiva dos autores, verificar se o critério de completude foi ou não alcançado. Consequentemente, tal verificação é essencial para determinar o impacto na identificação da qualidade do dado disponibilizado. Ao mesmo tempo, Oletto (2006) questiona a possibilidade de definir um conjunto universal de atributos para garantir a qualidade da informação.

Piccolo *et al.* (2021) clarificam que a qualidade dos dados não é um parâmetro absoluto, mas sim dependente do contexto da pesquisa e de seus usos pretendidos e enfatizam a importância de considerar os requisitos do domínio de atuação da pesquisa.

Essa contextualização é fundamental pois, ao compartilhar dados, é imprescindível que sejam fornecidas informações detalhadas sobre o contexto de sua produção e as limitações de seu uso para que, assim, os usuários possam avaliar sua qualidade e adequação às próprias necessidades.

No contexto legislativo brasileiro, a partir da Lei de Acesso à Informação (Lei 12.527/2011) e do Decreto nº 7.724/2012, Nascimento (2022)¹⁵ afirma que os dados devem ser tratados de forma a preservar cinco propriedades.

Para os dados ostensivos, tem-se:

- Disponibilidade: se refere à preservação da informação para que esta esteja acessível sempre que necessário;
- Autenticidade: assegura a veracidade da origem e autoria da informação; e
- Integridade: a estabilidade do conteúdo registrado, ou seja, é a garantia de que esse conteúdo não foi alterado depois de concluído o seu registro.

Para os dados sigilosos, a propriedade de:

- Confidencialidade: relacionada à permissão de acesso, ou seja, é a garantia de que a informação será acessada apenas por quem tenha a necessidade de conhecê-la.

Já para todos os dados, as propriedades:

- Primariedade: refere-se a fonte primária, ou seja, o demandante tem o direito de receber os dados diretamente da fonte original sem intermediação e em seu estado bruto e não uma descrição do seu conteúdo; e
- Atualização: relacionada a possibilidade de receber os dados e informações mais recentemente produzidos.

Mais uma vez, é possível observar que autenticidade (veracidade) e integridade (fixidez) estão sendo identificadas como propriedades que devem ser preservadas.

¹⁵ Material disponibilizado na aula 5 sobre “Segurança da informação” do “Curso de Ciência Aberta” da Fiocruz, realizada por esta pesquisadora.

2.3.2 Confiança para (re)uso de dados de pesquisa: o valor de evidência e a autenticidade na perspectiva arquivística.

Considera-se que dados só importam se forem úteis (Webster, 2014), portanto a definição de seus atributos é fundamental, mas não garante que dados de pesquisa sejam efetivamente (re)usados. Nesse sentido, a documentação, incluindo seus metadados, é um fator importante e frequentemente enfatizado como uma condição essencial para o sucesso do re(uso) de dados (Caregnato *et al.*, 2019).

A confiança nos dados, que está intimamente ligada à sua autenticidade, também é influenciada por fatores subjetivos relacionados ao (re)uso. Considerando as indicações de Laflamme, Poetz e Spichtinger (2022) sobre o aspecto subjetivo do (re)uso e o que foi exposto por Van de Sandt *et al.* (2019) acerca das questões que levam um pesquisador a se reconhecer como (re)usuário de dados de pesquisa, destaca-se que a confiança nesses dados é um dos fatores condicionantes para essa escolha. É preciso que o pesquisador tenha segurança suficiente para usar dados que não foram por ele produzidos em sua pesquisa, o que implica, inclusive, em sua reputação e na reputação de seus estudos (Van de Sandt *et al.*, 2019).

Parte-se da premissa de que dados de pesquisa apresentam valor de evidência, um termo comumente conhecido pelos profissionais da área arquivística. Conforme expõem Rocha e Rondinelli (2016), os termos “prova” ou “evidência” são constantemente utilizados para qualificar os documentos arquivísticos e “diz-se que esses documentos se constituem em prova/evidência de ação” (Rocha; Rondinelli, 2016, p. 64).

No entanto, ao citar Meehan (2006, *apud* Rocha; Rondinelli, 2016), as autoras apontam que evidência não é algo inerente ao documento arquivístico, ou seja, não pode estar contida nele, o que se tem no documento, então, são inferências sobre determinado evento.

A autenticidade é uma “característica importante para que o documento arquivístico seja considerado prova de uma atividade” (Rocha; Rondinelli, 2016, p. 68). Dividida em três aspectos, a autenticidade pode ser: legal (onde sua genuinidade é garantida por uma autoridade pública), diplomática (conformidade com as práticas da época e local, e assinatura da pessoa competente) e histórica (atestação de eventos reais

ou informações verdadeiras, o que se aproxima da “confiabilidade” quanto à veracidade do conteúdo) (CONARQ, 2012, p.3).

Ao considerar esses diferentes aspectos da autenticidade, Rocha e Rondinelli (2016) enfatizam que, mesmo sendo independentes, eles são fundamentais para apoiar o potencial de prova de um documento e que, no contexto digital, a autenticidade diplomática (manutenção do documento sem alterações) é particularmente ameaçada pela facilidade de alteração sem rastros. Fato que se coloca como um grande desafio para os arquivistas na garantia da integridade dos documentos digitais ao longo do tempo, segundo as autoras.

Adicionalmente, Gava e Flores (2021) destacam que a transformação digital coloca em risco a autenticidade dos documentos a cada transmissão no tempo ou no espaço, quando, por exemplo, o objeto digital sai de um ambiente de produção para um ambiente de preservação. Pode-se apontar, ademais, a necessidade constante de migrações de formatos de arquivo e as reformatações de suporte enquanto fatores impactantes a esse risco.

Portanto, sendo o ambiente digital altamente vulnerável, faz-se necessário que a autenticidade desses documentos seja continuamente mantida e acrescida de registros cumulativos que a apoiem durante todo seu ciclo de vida, não apenas em momentos estanques e desconexos.

Conforme documento intitulado “Diretrizes do Preservador”, desenvolvido no âmbito do Projeto Interpares¹⁶, algumas garantias de autenticidade podem ser ofertadas por uma cadeia de custódia ininterrupta, pelo conhecimento de práticas de manutenção de documentos e por verificação.

De acordo com o *ICA Multilingual Archival Terminology* (ICA, [s.d.]), o termo cadeia de custódia diz respeito a “documento ou trilha que demonstra a sucessão de entidades coletivas ou pessoas que tiveram posse, custódia e controle sobre os documentos”. Contudo, no cenário digital, a demonstração de uma cadeia de custódia

¹⁶ O Projeto InterPares é “uma iniciativa acadêmica de pesquisa em preservação digital, especificamente voltada para os documentos arquivísticos, de enorme alcance. Tem influenciado diversas iniciativas de arquivos digitais, como dos Arquivos Nacionais dos Estados Unidos e de Portugal, e a elaboração de normas e padrões. No Brasil, os trabalhos desenvolvidos pela Câmara Técnica de Documentos Eletrônicos, do Conselho Nacional de Arquivos (Conarq), e as iniciativas de documentos digitais do Arquivo Nacional têm sido também orientados por este projeto”. (Fonte: <https://pt.scribd.com/document/235814347/Projeto-InterPARES>).

mantida ininterruptamente não é suficiente. Dessa forma, o Projeto InterPares¹⁷ apresenta o que denomina de Cadeia de Preservação (*Chain of Preservation – CoP*).

Essa Cadeia, conforme Glossário de Documentos Arquivísticos Digitais, engloba “sistema de controles que se estende por todo o ciclo de vida dos documentos, a fim de assegurar sua autenticidade ao longo do tempo”, ou seja, “que garantem uma cadeia ininterrupta dos sistemas informatizados” (CONARQ, 2020)¹⁸, fundamentando um ambiente seguro e rastreável que garanta sua confiabilidade, autenticidade e acurácia (ou precisão) desses registros (InterPares 2 Project).

De acordo com o disposto no InterPares Glossary Online:

- A confiabilidade diz respeito à sustentação dos fatos ao qual o registro se refere, sendo estabelecido pelo exame da completeza de sua forma e o grau de controle exercido no processo de produção.
- A autenticidade é estabelecida pela avaliação da identidade e integridade¹⁹ do registro. Essa integridade deve ser passível de ser verificada em todos os momentos do seu ciclo de vida. Ela deve conter: o que é um registro, quando foi criado, por quem, em que ação ou matéria participou e quais são suas implicações jurídicas/administrativas, culturais e contextos documentais. Além disso, deve ser verificada pela totalidade e solidez do registro: se ele está intacto ou, se não, o que está faltando; e
- A acurácia enquanto o grau ao qual dados, informações, documentos e documentos arquivísticos são precisos, corretos, verdadeiros, livres de erros ou distorções, ou pertinentes ao assunto.

¹⁷ O Projeto InterPares é “uma iniciativa acadêmica de pesquisa em preservação digital, especificamente voltada para os documentos arquivísticos, de enorme alcance. Tem influenciado diversas iniciativas de arquivos digitais, como dos Arquivos Nacionais dos Estados Unidos e de Portugal, e a elaboração de normas e padrões. No Brasil, os trabalhos desenvolvidos pela Câmara Técnica de Documentos Eletrônicos, do Conselho Nacional de Arquivos (Conarq), e as iniciativas de documentos digitais do Arquivo Nacional têm sido também orientados por este projeto”. (Fonte: <https://pt.scribd.com/document/235814347/Projeto-InterPARES>).

¹⁸ Em 2 de julho de 2019, no III Simpósio do Patrimônio Histórico e Cultural Militar, a autora Claudia Lacombe realizou uma apresentação elucidativa quanto a essa questão, intitulada “Autenticidade do documento arquivístico digital e a Cadeia de Preservação ininterrupta”. Disponível em https://www.gov.br/defesa/pt-br/arquivos/ajuste-01/2019/mes06/simposio_protocolo/7a_claudiaa_lacombea_autenticidadea_doca_arquiva_digitala_cadeia_a_dea_preserva_ininterrupta.pdf. Acesso em: 20 jun. 2025.

¹⁹ No glossário do e-Arq Brasil – versão 2 (CONARQ, 2022, p. 213), aponta-se que o termo integridade não deve ser confundido com “integridade arquivística” cujo objetivo, decorrente do princípio da proveniência, “consiste em resguardar um fundo de misturas (sic) com outros, de parcelamentos e de eliminações indiscriminadas. Também chamado integridade do fundo”. (Arquivo Nacional, 2005, p. 108”).

Ao compreender, portanto, que os dados inerentes ao desenvolvimento de pesquisas devem também possuir valor de evidência, clarifica-se a relevância do planejamento de estratégias de ação que consubstanciem a presunção de sua autenticidade.

Ao considerar as demandas imbricadas no aspirado (re)uso, entende-se que a confiança nos dados de pesquisa compartilhados e disponibilizados é um fator determinante na decisão do usuário por essa escolha. Dessa forma, torna-se essencial incorporar informações que viabilizem a análise da capacidade do dado transmitir, com precisão, a mensagem que motivou sua produção, de modo a atender aos objetivos previstos (integridade) e possibilitar que ele seja caracterizado como um registro único (identidade) (CONARQ, 2012). Além disso, todo o arcabouço informacional que ampara sua devida contextualização e proveniência, também deve ser considerado.

Sayão e Sales (2024) apontam a necessidade da implementação de ações transversais para que a colaboração, transparência e confiança, qualidade e engajamento, enquanto objetos de uma Ciência Aberta, possam ser cumpridos e para que a ampla disponibilização dos resultados e processos de pesquisa viabilize um escrutínio rigoroso pela comunidade científica e da sociedade. Os autores elencam parâmetros que precisam ser considerados e dentre eles, a qualidade dos dados emerge como um elemento fundamental.

Observa-se, portanto, que não basta “abrir”, é preciso que haja qualidade dos dados, o que coloca a confiabilidade do dado de pesquisa como um pré-requisito fundamental para seu (re)uso eficaz.

Como foi demonstrado, embora não haja consenso sobre métodos e métricas para avaliação da qualidade dos dados de pesquisa, entende-se que a transparência e o detalhamento do registro da proveniência desses dados contribuem para avaliar a partir da compreensão do seu contexto de produção.

Opta-se por considerar, para fins desta tese, a conceituação arquivística no que tange à autenticidade englobar a integridade. No entanto, ao invés de correlacionar a confiabilidade a quesitos inerentes exclusivamente à produção, relacionando-a mais à autenticidade histórica, compreende-se que, no contexto dos dados de pesquisa, a confiabilidade perpassa diversas camadas. Contudo, nesta tese, será dada ênfase à confiabilidade que se ampara na manutenção da autenticidade diplomática, observando-

se o seguinte encadeamento: a confiabilidade decorre da aferição da presunção de autenticidade, sustentada por evidências acrescentadas pelo registro de proveniência.

Identifica-se como relevante o papel da curadoria digital para garantir uma adequada documentação, organização, gestão e preservação dos dados de pesquisa, na intenção de manter sua qualidade para (re)uso futuro e maximizar seu impacto na pesquisa.

A discussão sobre a qualidade dos dados é central para esta tese, pois uma metodologia de gestão de dados de pesquisa que possa amparar sua preservação digital qualitativa deve necessariamente abordar como garantir o registro de informações que demonstrem a autenticidade, enquanto uma qualidade dos dados ao longo do tempo, de modo a assegurar seu (re)uso eficaz e confiável no contexto da Ciência Aberta.

Faz-se mister, portanto, aprofundar na próxima seção a compreensão sobre a curadoria digital, a qual engloba “todas as atividades envolvidas na gestão de dados, desde o planejamento de sua criação, melhores práticas em digitalização e documentação e garantia de sua disponibilidade e adequação para descoberta e reutilização no futuro” (DCC, 2008).

Finalmente, esta seção delineou os fundamentos epistemológicos e multidisciplinares que embasam a Ciência Aberta, desde suas raízes nas discussões sobre o acesso ao conhecimento até as complexas questões de qualidade dos dados. Adicionalmente, localizou a opção teórico-metodológica de apropriar-se da conceituação de autenticidade desenvolvida na área arquivística.

Assim sendo, para que esses fundamentos se traduzam em práticas efetivas, é essencial uma abordagem robusta para a gestão dos dados de pesquisa ao longo de seu ciclo de vida.

3 O ESCOPO DA CURADORIA DIGITAL PARA O (RE)USO CONFIÁVEL DE DADOS DE PESQUISA

Esta seção explora o papel da curadoria digital, que engloba a preservação digital e a gestão de dados, como o elo que conecta os princípios da Ciência Aberta à realidade da pesquisa. Seu objetivo é garantir que os dados não apenas tenham a capacidade de serem abertos e compartilhados como, também, preservados a longo prazo com qualidades suficientes que maximizem seu potencial de (re)uso.

Dessa forma, cumpre esclarecer que a curadoria digital emerge como um conceito central para abordar os desafios informacionais e tecnológicos dos dados de pesquisa. Seu escopo envolve a gestão e a preservação de dados e informação ao longo do tempo, desde sua criação até seu (re)uso. Essa abordagem a distingue da “curadoria de dados”, mesmo que na literatura seja possível observar o uso de ambas como sinônimos uma da outra.

A curadoria digital é exposta nesta seção enquanto um processo holístico que abrange todo o ciclo de vida dos dados de pesquisa e que exige a integração de diferentes modelos e *frameworks* para sua efetivação. Para tanto, são apresentados os modelos do Ciclo de Vida dos dados de pesquisa (da *Joint Information Systems Committee* - JISC), e do Ciclo de Vida da Curadoria Digital (do *Data Curation Centre* - DCC), o Plano de Gestão de Dados (PGD) e os princípios FAIR, destacando suas contribuições e limitações.

Apresenta-se, ainda, a preservação digital como um componente essencial nesse cenário por ser um conjunto de estratégias e ações que garantem a manutenção da disponibilidade e o acesso contínuo aos dados de pesquisa. Enfatiza-se, portanto, que a efetivação da gestão de dados de pesquisa é condição *sine qua non* para a preservação digital qualitativa desses objetos digitais por possibilitar registros de metadados contínuos e cumulativos que demonstrem sua proveniência e sustentem sua presunção de autenticidade, o que maximiza sua confiabilidade e ampara o (re)uso.

3.1 A CURADORIA DIGITAL DADOS DE PESQUISA: MODELOS E FRAMEWORKS

Nos últimos anos, o aumento do volume e a aceleração dos processos de criação e transmissão de dados de pesquisa tem trazido reflexões sobre os critérios a serem

estabelecidos para gestão, armazenamento, compartilhamento e preservação desses objetos digitais. Diante dos desafios informacionais e tecnológicos dos dados de pesquisa digitais, o termo “curadoria digital” ganha relevância.

Segundo Ball (2010), essa expressão surgiu em 2001 como título de um seminário nas áreas de arquivo, biblioteca e *e-Science*. De acordo com o *Digital Curation Centre* (DCC, [s.d])²⁰, o termo *digital curation* (curadoria digital) abrange a gestão e preservação de dados/informação ao longo do tempo, incluindo manutenção, preservação e agregação de valor durante seu ciclo de vida.

Ainda hoje, a escolha pelo uso do termo *digital curation* (curadoria digital) ou *data curation* (curadoria de dados) não é bem definida na literatura. Sobre esse fato, Ball (2010) entende que isso ocorra em virtude de que, em alguns contextos, os termos “dados” e “objetos digitais” têm sido tratados quase como sinônimos, o que leva à adoção de “curadoria de dados” como “curadoria digital”.

No entanto, o autor reforça que o termo “curadoria de dados”, que já era anteriormente utilizado, refere-se à “seleção, normalização, anotação e integração de dados de periódicos, relatórios ou bases de dados de terceiros em um banco de dados sobre um determinado tópico, a fim de mantê-lo atualizado e relevante” (Ball, 2010, p. 5). Ou seja, a curadoria digital possui um escopo ampliado ao ser comparada com o que se entende por curadoria de dados.

Assim, ao considerar os dados de pesquisa digitais como objeto nuclear desta tese, adota-se que as questões inerentes à sua gestão e preservação constituem a “curadoria digital de dados de pesquisa”. Dados estes que estejam em formato digital, nato digitais ou representantes digitais (oriundos do processo de digitalização), mesmo que ainda não se encontrem abertos diante da perspectiva da Ciência Aberta.

Por visar garantir o acesso e valor dos dados de pesquisa enquanto evidência e recurso valioso a ser usufruído em futuras pesquisas, a curadoria digital possibilita a efetivação de um fazer científico mais transparente, colaborativo, reprodutível e confiável. Sua aplicação e efetividade determinam a qualidade do que será disponibilizado. De toda forma, “É importante destacar que o grau de confiança, a utilidade e o valor desses dados variam enormemente: enquanto alguns são estruturados

²⁰ Disponível em: <https://www.dcc.ac.uk/about/digital-curation>. Acesso em: 20 out 2025.

e recebem curadoria, outros são simplesmente disponibilizados” (Coligman *et al.*, 2021, p. 196, tradução nossa).

Todavia, Ball (2010) reforça que a *research data curation* (curadoria de dados de pesquisa) ainda se constitui uma “disciplina imatura”. Isto porque, segundo o autor, essa curadoria mesmo não sendo uma atividade recente, passa por novidades consideráveis. Por exemplo, sua realização de forma inteiramente digital, impulsionada pelo formato dos dados, e sua adoção em um cenário mais generalista, fora de centros de dados especializados (Ball, 2010).

Essa visão de Ball (2010) pode ser ratificada quando, uma década depois, Sayão e Sales (2021) expõem que a curadoria digital ainda é um conceito em evolução. No entanto, os autores apontam que, mesmo em desenvolvimento, há o estabelecimento de que essa curadoria abarca “a gestão atuante e a preservação de recursos digitais durante todo o ciclo de vida de interesse do mundo acadêmico e científico, tendo como perspectiva o desafio temporal de atender a gerações atuais e futuras de usuários” (Sayão; Sales, 2021, p. 185).

Em acréscimo, Strong (2012) explica que a curadoria digital é, em essência, um processo holístico que analisa todo o ciclo de vida dos dados, englobando sua criação, manutenção e preparação quanto a ajustes e migrações necessárias futuramente, tratando-se de tomadas de decisões informadas desde o início do ciclo de vida dos dados para garantir que sejam úteis mais tarde.

Atualmente, diversos modelos e *frameworks* têm sido desenvolvidos para apoiar a realização da curadoria digital de dados de pesquisa. Eles sustentam sua efetivação enquanto estruturas conceituais e práticas e tem por objetivo, de maneira geral, fornecer um conjunto de princípios, diretrizes e fluxos de trabalho que possam garantir a confiabilidade, utilidade e acesso aos dados de pesquisa ao longo do tempo.

3.1.1 Modelo do ciclo de vida dos dados de pesquisa – JISC

Instituições têm estabelecido modelos para o ciclo de vida dos dados de pesquisa com o intuito de fornecer uma estrutura que considere as diversas operações a serem realizadas em cada uma de suas fases.

Para demonstrar esse panorama, Ball (2012) realizou um levantamento para revisar oito modelos que foram criados sob contextos variados. São eles:

- Modelos da *Australian National Data Service* (ANDS) e do *UK Data Archive*: que demonstram o ciclo de vida dos dados em cenários especificamente organizacionais.
- Modelos do *Infrastructure for Integration in Structural Sciences* (I2S2), do *Research360 Institutional Research Data lifecycle* e do *Capability Maturity Model for Scientific Data*: apresentam o ciclo de vida dos dados de pesquisa sob o olhar de um projeto de pesquisa.
- Modelos da *Digital Curation Centre* (DCC), *DataONE* e *DDI Combined Life Cycle Model*: apresentam vieses voltados a uma comunidade específica.

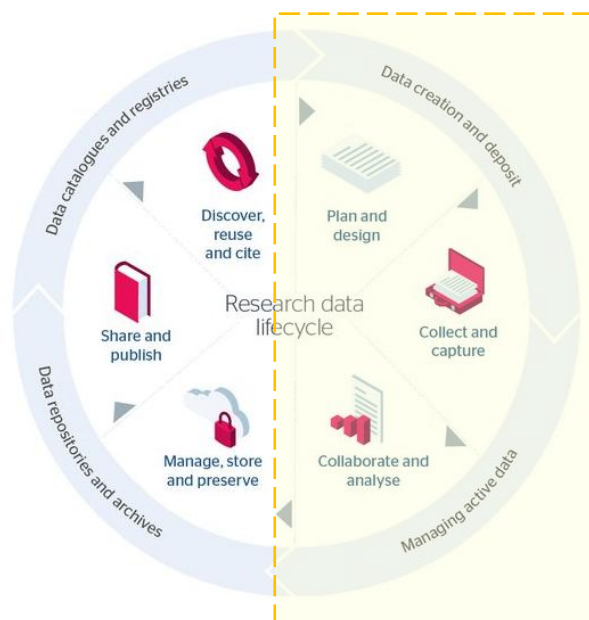
Ao realizar essa revisão, Ball (2012) apontou que, mesmo com características distintas, todos possuem um ponto em comum: o fato de que o início do ciclo parte de um planejamento. Esse aspecto é importante porque, ao planejar, o pesquisador consegue traçar os passos necessários para tratar adequadamente os dados antes mesmo de serem criados.

Ao serem identificados e preparados com antecedência ou durante as atividades de pesquisa, as ações de gestão tornam-se mais fáceis de serem realizadas. Ball (2015) aponta que o exercício de destrinchar os estágios do ciclo de vida dos dados viabiliza a proposição de uma lista relativamente detalhada dos processos para uma possível verificação das práticas, necessidades, agentes e infraestruturas envolvidas que merecem atenção e devem ser considerados desde os momentos preliminares da pesquisa.

Dessa forma, traz-se o Modelo do ciclo de vida dos dados desenvolvido pela JISC (*Joint Information Systems Committee*). Sua adoção para esta tese se deu em virtude de que, mesmo elaborado para o contexto de pesquisas acadêmicas das Universidades do Reino Unido, adota uma ótica aplicável a outros cenários de pesquisa. Além disso, é visualmente mais simples e didático. Cabe expor, adicionalmente, que, segundo Sá e Pereira (2024)²¹, as Universidades têm se tornado um referencial mais amadurecido quanto às questões e soluções tangidas pela gestão de dados de pesquisa.

²¹ O estudo foi apresentado pelas autoras Ivone Sá e Renata Pereira, em 2024, em reunião ao Comitê Gestor do Arca Dados, da Fiocruz, e ainda não foi publicado. Tendo sido enviado por e-mail a esta pesquisadora.

Figura 6 – Modelo do ciclo de vida dos dados de pesquisa da JISC



Fonte: Adaptado pela autora com base em site da JISC.

Com base na Figura 6 é possível identificar quatro grandes etapas: *Data creation and deposit* (Criação dos dados e depósito), *Managing active data* (Gerenciando dados ativos), *Data repositories and archives* (Repositórios e arquivos de dados) e *Data catalogues and registries* (Catálogos e registros de dados).

Ao analisá-lo junto ao *Checklist* de verificação da Gestão de Dados de Pesquisa (JISC), é possível compreender que a primeira fase de “criação dos dados e depósito” abrange dois tópicos:

- a) “Planejar e projetar” que engloba:
 - a.1) “Planejamento de gestão de dados” onde deve ser observada a preparação de um Plano de Gestão de Dados (PGD)
 - a.2) “Conformidade com a política” onde o pesquisador deve tomar ciência dos requisitos do financiador e da instituição/organização ao qual está vinculado.
 - a.3) “Princípios FAIR na gestão de dados de pesquisa” onde deve ser observado como esses princípios se aplicam ao projeto de pesquisa em questão.

- a.4) “Custeio” onde devem ser calculados os custos das atividades de RDM e se os serviços oferecidos pela instituição/organização foram considerados para isso.
- b) Uma parte de “Coletar e capturar” que engloba:
 - b.1) “Armazenamento e backup de dados ativos” onde são avaliadas se as estratégias para salvar os dados ativos estão claramente descritas e se foram consideradas as opções de backup desses dados salvos.
 - b.2) “Catalogação de dados de pesquisa (questões de metadados)” onde devem ser escolhidos os modelo e padrão de metadados, levando em consideração se são apropriados para a área da pesquisa em questão, e se o *software* escolhido pelo pesquisador permite adição de metadados.

Já a segunda fase, identificada como “Gerenciando dados ativos”, abrange os dois tópicos a seguir:

- c) A outra parte de “Coletar e capturar” que engloba:
 - c.1) “Gerenciamento e formatos de arquivos” onde deve ser documentada a hierarquia de pastas e a convenção de nomenclatura de arquivos, além da escolha de formatos de arquivo apropriados para permitir o reuso a longo prazo.
- d) “Colaborar e analisar” que engloba:
 - d.1) “Pesquisa colaborativa” onde são analisadas: como a comunicação é gerenciada na equipe, como os dados são compartilhados entre os membros da equipe, se as responsabilidades foram formalmente atribuídas aos membros da equipe, se foi criada ou se o pesquisador participa de uma plataforma que permite pesquisa colaborativa, se a colaboração levou em conta a segurança dos dados e se existe uma abordagem documentada para controle de versão de arquivos.

Adiante, na terceira fase “Repositórios e arquivos de dados” estão abrangidos os dois tópicos:

- e) “Gerenciar, armanezar e preservar” que engloba:
 - e.1) “Preservação” onde devem ser selecionados os resultados apropriados para compartilhamento de longo prazo, as saídas devidamente

documentadas e a escolha de uma solução de sistema/*software* para compartilhamento de longo prazo.

- e.2) “Segurança” onde devem ser tomadas as providências para proteger os dados de pesquisa.
- e.3) “Regulamentação de proteção de dados” onde o pesquisador confirma ciência quanto às suas obrigações legais ao lidar com informações de identificação pessoal.
- e.4) “Protegendo informações de identificação pessoal” onde são tomadas providências práticas para proteger informações de identificação pessoal com a escolha de uma abordagem apropriada para anonimização.
- e.5) “Programas” onde é selecionado o *software* apropriado para compartilhamento de longo prazo, a documentação do *software* que se deseja compartilhar é preparada e o meio para publicação do *software* é escolhido.
- f) Uma parte de “Compartilhar e publicar” que engloba:
 - f.1) “Propriedade intelectual e direitos autorais” onde o pesquisador ratifica sua ciência sobre seus direitos ao compartilhar/publicar dados, os direitos de outras pessoas quando ele reusa o trabalho delas e sobre o que os outros podem/não podem fazer com os dados que ele compartilha/publica.
 - f.2) “Licenciamento” onde é formulado um plano para licenciar os dados de pesquisa para permitir reuso futuro.
 - f.3) “Identificadores” onde são considerados identificadores persistentes ou links permanentes (permalinks) para compartilhar os dados.

E por último, na quarta fase “Catálogos e registros de dados” os dois tópicos abaixo são abrangidos:

- g) A outra parte de “Compartilhar e publicar” que engloba:
 - g.1) “Vinculando dados a uma publicação” onde é escolhido o local em que os dados são compartilhados a partir de considerações sobre as suas vantagens e desvantagens.
- h) “Descobrir, reusar e citar” que engloba:
 - h.1) “Citação de dados e métricas” onde devem ser citadas todas as fontes apropriadas ao divulgar os dados, o pesquisador deve estar ciente das

métricas de citação de dados e do seu significado e saber onde encontrar dados relevantes para reutilizar em sua pesquisa.

Ao analisar cada etapa do Modelo de ciclo de vida dos dados de pesquisa da JISC, expõe-se que a opção em demarcar na Figura 6 as duas primeiras fases teve por fundamento a indicação de que nelas, geralmente, o pesquisador é o responsável por decidir sobre, e executar, as atividades necessárias para a gestão desses dados. Isso se dá, principalmente, pelo fato de os dados de pesquisa ainda estarem em um ambiente sob sua custódia e responsabilidade.

Outra condição que justifica essa demarcação relaciona-se ao fato de que o tratamento a ser facultado nesses momentos preliminares do ciclo de vida dos dados de pesquisa possui impacto direto no nível de qualidade desses dados ao longo de todo o restante do diagrama.

Esses dois fatores pontuados demonstram a importância do estabelecimento de orientações e na escolha de recursos e ferramentas capazes de instrumentalizar e apoiar a execução dos procedimentos adequados para criação e depósito dos dados de pesquisa. Essa fase de criação e depósito alicerça o gerenciamento dos dados ativos, para a qual é relevante considerar o apoio de profissionais capacitados a orientar o pesquisador na realização dessas atividades de maneira adequada e efetiva. Considerando, inclusive, a possibilidade de que esse profissional assumira diretamente a execução dessas atividades.

A partir disso, são dinamizadas as atividades de gerenciar, armanezar e preservar (letra “e”) e assegura-se que as decisões quanto a abertura e FAIRificação²² ao compartilhar e publicar (letra “f”) sejam tomadas de forma consciente e responsável.

O Modelo da JISC demonstra o ciclo de vida dos dados de pesquisa enquanto uma série de ações complexas e correlacionadas entre si. A modernização contínua e acelerada das tecnologias de informação e comunicação utilizadas para criar, gerir, armazenar, compartilhar, preservar e acessar esses dados incrementam ainda mais essa complexidade. Dados de pesquisa criados, simulados e disponibilizados mediante uso de recursos tecnológicos modernos e atuais tem sua transmissão, no tempo e espaço, agilizada. Todavia, esses recursos se tornam obsoletos em prazos cada vez mais curtos.

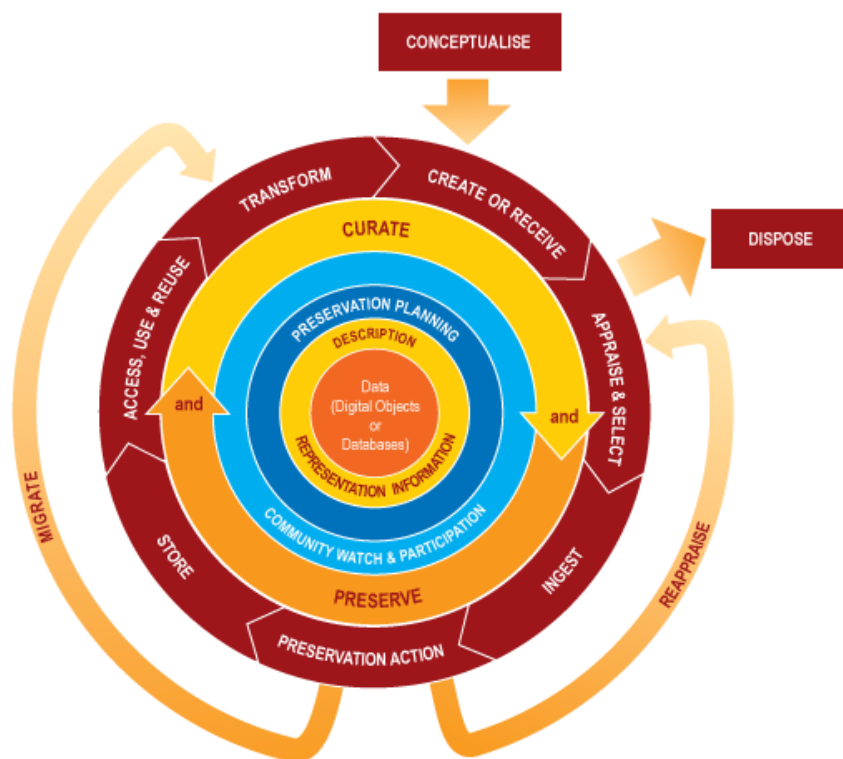
²² Termo usado para se referir aos processos voltados à implementação prática dos princípios FAIR, detalhados em <https://www.go-fair.org/fair-principles/fairification-process/>. Acesso em: 13 out. 2025.

Dessa forma, clarifica-se que o estabelecimento do ciclo de vida dos dados de pesquisa precisa estar integrado a uma dinâmica de curadoria digital que esquematize as ações e os recursos necessários para sua manutenção.

3.1.2 Modelo do Ciclo de Vida da Curadoria Digital (DCC)

É possível observar a existência de modelos que esquematizam as ações da curadoria digital. Dessa forma, julga-se relevante expor o que tem sido mais amplamente divulgado e referenciado no contexto em questão. Opta-se, portanto, por apresentar o Modelo de ciclo de vida da curadoria, desenvolvido por Higgins (2008) e, posteriormente, adotado pelo *Digital Curation Centre* (DCC).

Figura 7 – Modelo do ciclo de vida da curadoria - DCC



Fonte: The Digital Curation Lifecycle Model by Sarah Higgins, IJDC Issue 1, Volume 3 (2008).

A Figura 7, de maneira geral e simplificada, apresenta uma forma visualmente clara de compreender o Modelo. Conforme é exposto por Lindlar *et al.* (2020), os dados

digitais, que podem ser simples, complexos ou bancos de dados, apresentam-se no centro do círculo, circundados por níveis do ciclo de vida da curadoria. Os três primeiros níveis são entendidos como “atividades contínuas” e seus metadados (descritivos e informações de representação), que tem por objetivo manter os dados (centro da esfera) com significado. Os autores (2020) acrescentam que o modelo tem em seu núcleo os dados digitais, mas se concentra no processo de sua curadoria.

Ao apresentar todos os níveis desse ciclo, Ball (2012), aponta que

- No primeiro nível está a “descrição” e a (gestão da) “Informação de Representação”, ou seja, a criação, coleção, preservação e manutenção de metadados suficientes para permitir que os dados sejam usados e reutilizados enquanto tiverem valor para justificar a curadoria contínua.
- No segundo nível tem-se o “Planejamento de preservação” que abarca estratégias, políticas e procedimentos para todas as ações de curadoria.
- No terceiro nível encontra-se a “Vigilância e participação da comunidade Designada”, denominado pelo Modelo de Referência *Open Archival Information System* (OAIS)²³ (apresentado na subseção 4.1 desta tese) como um grupo predeterminado de partes interessadas nos dados. O objetivo desse nível é acompanhar as mudanças de requisitos, o que engloba a participação no desenvolvimento de padrões, ferramentas e *softwares* relevantes para manutenção e acesso aos dados.
- No quarto nível tem-se as atividades de “Curar” e “Preservar” que são relativas à maioria das ações no modelo, mas, especificamente, representam a execução do gerenciamento planejado e das ações administrativas de apoio à curadoria.
- No quinto, e último, nível, tem-se as ações sequenciais, em bordô, as quais não tem preocupação exclusiva com a curadoria, mas sim com os estágios do ciclo de vida dos dados que devem ter a curadoria como componente.

Nessas ações sequenciais, segundo a descrição disponível no site da DCC, são propostas as seguintes etapas:

²³ Um OAIS é um sistema de arquivo que consiste em hardware, software, informações e políticas baseadas em processos e procedimentos colocados em prática e operados por uma organização que aceitou a responsabilidade de preservar informações e disponibilizá-la para uma comunidade designada, e sua equipe (CCSDS 650.0-M-3, 2024).

- a) “Conceituar” que abrange conceber e planejar a criação ou coleta de objetos digitais, incluindo métodos de captura de dados e opções de armazenamento.
- b) “Criar ou receber” que abarca a produção ou recebimento de objetos digitais e a atribuição de metadados administrativos, descritivos, estruturais e técnicos.
- c) “Acesso e uso” onde é certificado que os usuários designados podem acessar facilmente objetos digitais no dia-a-dia. Alguns objetos digitais podem estar disponíveis publicamente, enquanto outros podem ser protegidos por senha.
- d) “Avaliar e selecionar” que compreende a avaliação dos objetos digitais e a seleção daqueles que exigem curadoria e preservação a longo prazo. Sendo preciso que sejam aderentes a orientações, políticas e requisitos legais documentados.
- e) “Descartar” objetos digitais não selecionados para curadoria e preservação a longo prazo. Orientações, políticas e requisitos legais documentados podem exigir a destruição segura desses objetos.
- f) “Ingestão” compreende a transferência de objetos digitais para um arquivo, repositório digital confiável, *data center* ou similar, novamente aderindo a orientações, políticas e requisitos legais documentados.
- g) “Ação de preservação” que empreende ações para garantir a preservação e retenção a longo prazo da natureza autoritária dos objetos digitais.
- h) “Reavaliar” indica o retorno de objetos digitais que falham nos procedimentos de validação para avaliação e resseleção adicionais.
- i) “Armazenar” abarca a manutenção dos dados de maneira segura, conforme descrito pelos padrões relevantes.
- j) “Acesso e reutilização” onde são certificados que os dados estejam acessíveis aos usuários designados para uso e reutilização pela primeira vez. Alguns materiais podem estar disponíveis publicamente, enquanto outros dados podem ser protegidos por senha.
- k) “Transformar” significa a criação de novos objetos digitais a partir do original, por exemplo, migrando para um formulário diferente.

Ao apresentar essa estrutura, o modelo da DCC propõe a integração do ciclo de vida dos dados junto ao ciclo de vida da curadoria. Essa proposta demonstra aspectos que devem ser observados tanto sobre os dados quanto às atividades envolvidas em sua gestão. De toda forma, é preciso expor que desde sua elaboração, em 2008, esse modelo vem sendo analisado e avaliado por pesquisadores e gestores da informação quanto à sua aplicação em cenários e contextos atuais.

Um exemplo recente, elaborado em razão da rápida mudança do cenário global quanto à curadoria digital, chama-se d-KISTI. Essa contraproposta a modelos de ciclo de vida existentes à época foi publicada por Rhee (2022). O d-KISTI é apresentado como uma nova proposta de Modelo de ciclo de vida, com intuito de atualizar o modelo da DCC com novas aplicações e significados (Rhee, 2022). Segundo o autor, esse novo modelo coreano visa facilitar a curadoria digital de forma amplamente aplicável, eficiente e sistemática, como forma de oferecer atividades de curadoria mais representativas para as organizações de informação.

Apesar de ser uma proposta de aprimoramento do modelo DCC, o modelo d-KISTI não foi considerado para esta tese devido ao seu pouco tempo de desenvolvimento e outros estudos que abordem evidências quanto a seu nível de maturidade e aplicação prática.

Assim, embora o modelo DCC possa vir a ser incrementado diante das demandas atuais inerentes à curadoria digital, sendo passível de ajustes, sua contribuição para a execução das atividades envolvidas na manutenção dos dados de pesquisa é reconhecida na área da Ciência da Informação. Quanto a isso, Gracy e Kahn (2012) sinalizam, inclusive, que o modelo da Digital Curation Centre pode ser considerado como um complemento ao Modelo de Referência OAIS.

Dessa forma, justifica-se utilizar o modelo DCC como parâmetro para as questões a serem discutidas nesta tese. Essa escolha é reforçada quando Gracy e Kahn (2012) apontam que Harvey (2010) o considera uma proposta “particularmente valiosa porque transmite o conhecimento crítico que a preservação e o reuso bem-sucedido de dados arquivados dependem de um bom design de dados e captura de informações essenciais sobre os dados desde o ponto de criação” (Harvey, 2010, *apud* Gracy; Khan, 2012, p. 55, tradução nossa).

Assim, ao analisar o modelo adotado, nota-se que a preservação digital se configura como um subsídio fundamental no escopo da curadoria digital à medida que

visa proporcionar a manutenção dos dados de pesquisa para possíveis acessos e (re)usos futuros. Sales e Sayão (2012) expõem que os dados digitais são passíveis de riscos inerentes à fragilidade das mídias aos quais se encontram e que o surgimento do termo “curadoria digital” se deu exatamente como resposta à necessidade de tratá-los e gerenciá-los de modo que sua preservação por longo prazo, acesso, autenticidade e (re)uso em novas pesquisas pudessem ser assegurados.

Para os autores (2012) o termo envolve a gestão de dados de pesquisa desde o seu planejamento, assegurando a sua preservação, descoberta, interpretação e reutilização, tendo por principal desafio “[...] preservar não somente o conjunto de dados, mas de preservar, sobretudo, a capacidade que ele possui de transmitir conhecimento para uso futuro das comunidades interessadas” (Sayão; Sales, 2012, p. 180).

Portanto, como foi apontado anteriormente, o modelo DCC especifica a preservação em dois momentos. O primeiro que, junto à curadoria no nível 4, compõe as ações administrativas de apoio e gerenciamento planejado a todas as demais etapas do ciclo. O segundo, compondo as ações da própria curadoria, no nível 5, enquanto ação voltada a garantir a preservação e retenção a longo prazo dos objetos digitais.

Essa perspectiva coaduna com o conceito de curadoria digital, e mesmo ao utilizar o termo *curate* e não *management* (gestão), junto à preservação, compreende-se, para fins desta tese, que o modelo expõe, de forma mais clara, a necessidade de ações de gestão e preservação ao longo de todo o ciclo. Esse ponto amplia o entendimento da preservação no que é exposto no ciclo de vida dos dados da JISC. Todavia, ambos os modelos reforçam a importância do planejamento, sendo que o modelo do DCC realça seu registro seja formalizado e contínuo.

3.1.3 Os Princípios FAIR

Em um cenário onde a importância dos dados para a sociedade e para o processo científico se apresenta proporcional à dificuldade que pessoas e aplicações computacionais encontram para localizar e utilizar esses dados disponíveis (Vidotti; Torino; Coneglian, 2021), surgem os “princípios FAIR”. O termo FAIR é um acrônimo em que cada uma das letras possui um significado, *Findable* (localizável), *Accessible* (acessível), *Interoperable* (interoperável) e *Reusable* (reusável).

Esses princípios se constituem como um conjunto de práticas orientadoras voltadas à boa gestão de dados e sua implementação faz parte do ecossistema da Ciência Aberta. Criados em janeiro de 2014, os princípios FAIR são resultado de um encontro ocorrido em Leiden (Holanda), chamado de “*Joint Designing a Data Fairport*”. Encontro que contou com a participação de especialistas, representantes da academia, agências de fomento à pesquisa, editores científicos e representantes da área industrial.

Em 2016, os princípios FAIR foram inseridos na pauta de discussões do *High Level Expert Group on the European Open Science Cloud* (EOSC), da Comissão Europeia. Sales *et al.* (2021) esclarecem que o objetivo desse grupo de especialistas era apresentar recomendações para que ciência, negócios, governo e, eventualmente, a indústria, pudessem aproveitar da melhor maneira os frutos da revolução do *big data*. Fato que, segundo os autores (2021), veio a causar “inquietação” na comunidade acadêmico-científica.

No entanto, De Oliveira Veiga *et al.* (2019) esclarecem que, em 2017, sua aplicação foi consolidada quando a Comissão Europeia passou a exigir que pesquisas por ela financiadas deveriam adotar Planos de Gestão de Dados com base nesses princípios.

Segundo disposto no módulo “O que é Ciência Aberta?” do Curso online modular sobre Ciência Aberta, oferecido pela Vice-presidência de Educação, Informação e Comunicação (VPEIC), da Fiocruz, FAIR tem as seguintes proposições:

- “F” propõe que os dados possam ser descobertos por outros. Para isso, tanto os dados quanto os metadados devem adotar um Identificador Persistente Único e serem indexados em repositórios confiáveis. Além disso, os metadados devem ser ricos;
- “A” propõe que os dados possam ser acessados por outros e isso pressupõe o uso de protocolos de comunicação padronizados, abertos e gratuitos, que ofereçam autenticação e acesso aos metadados mesmo quando o dado não estiver mais disponível;
- “I” propõe a integração de dados com outros dados ou facilmente “entendidos” por computadores, sendo necessário, portanto, o uso de linguagem de representação do conhecimento, vocabulários e/ou ontologias que adotem os princípios FAIR, além de dados e metadados interligados; e

- “R” propõe que os dados possam ser reusados por outros ou em novas pesquisas, o que prescinde que dados e metadados possuam múltiplos atributos, usem licenças apropriadas, tenham suas procedências descritas e usem padrões específicos da sua comunidade.

Para atender a essas proposições, FAIR se divide em 15 (quinze) princípios, dos quais, especificamente 4 (quatro): 1 (um) princípio que se desdobra em outros 3 (três), amparam as perspectivas do (re)uso (Wilkinson *et al.*, 2016). Em vistas a operacionalizá-los, Henning *et al.* (2019) propõem as seguintes aplicações:

- Para o princípio “R.1: (Meta)dados são descritos com uma pluralidade disponibilizada de atributos precisos e relevantes”, sendo essencial fornecer metadados detalhados o suficiente para que um pesquisador possa avaliar a adequação e o potencial de reuso dos dados para suas necessidades;
- Em relação a “R.1.1: (Meta)dados devem ter licenças de uso claras e acessíveis”, o responsável pelos dados e metadados deve definir explicitamente as condições de acesso, finalidade e permissões de uso por meio de licenças claras.
- No que concerne a “R.1.2: (Meta)dados devem estar associados à sua proveniência”, especificar a linhagem dos dados é crucial não apenas para avaliar sua utilidade, mas também para garantir o devido crédito aos seus criadores e mantenedores. Informações relevantes sobre a proveniência incluem:
 - a) o processo de obtenção dos dados (geração ou coleta);
 - b) limitações ou particularidades relevantes para outros pesquisadores;
 - c) data de geração, condições experimentais, responsáveis pela preparação, configurações de parâmetros e detalhes do *software* utilizado (nome e versão);
 - d) a distinção entre dados brutos e processados; e
 - e) a identificação clara e documentada das versões dos dados arquivados e reutilizados.

- Para “R.1.3: (Meta)dados devem estar alinhados com padrões relevantes do seu domínio”, além de aderir aos padrões específicos da área, é importante seguir as boas práticas de arquivamento e compartilhamento estabelecidas pela comunidade de pesquisa.

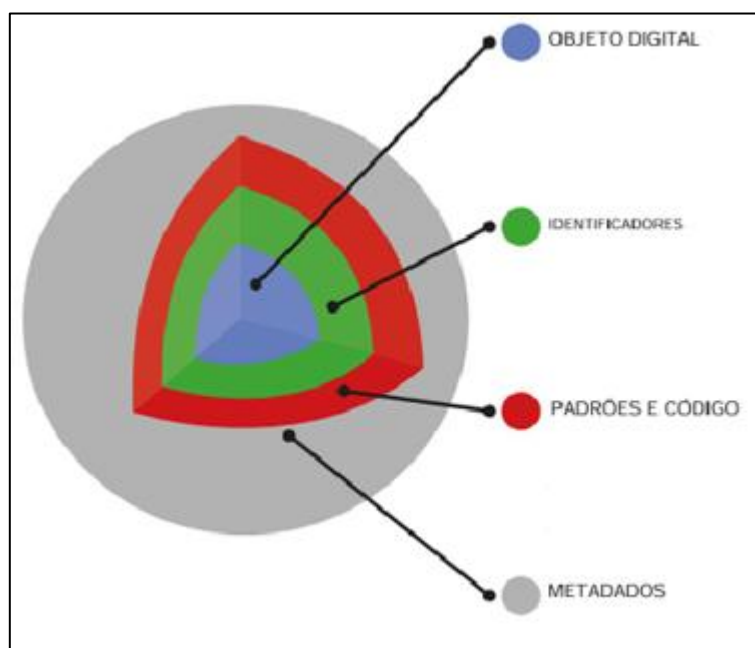
A relação de princípios FAIR é um “refinamento importante dos conceitos necessários para dar maior valor aos dados e aumentar a sua propensão para reutilização, por seres humanos e em escala por máquinas” (Comissão Europeia, 2018, p. 10, tradução nossa) e referem-se a três tipos de entidades: dados (ou qualquer objeto digital), metadados (informações sobre esse objeto digital) e infraestrutura (GOFAIR, [s.d.]).

Apesar de ter enfoque na infraestrutura tecnológica necessária para que dados possam ser encontrados, acessados, interoperados e reusados (Vidotti; Torino; Coneglian, 2021), FAIR não constitui um padrão e não especifica formatos ou tecnologias, sendo, portanto, fornecedor de contexto e direção para tornar os (meta)dados mais úteis e reusáveis (Mons, 2018).

Assim, de acordo com o Relatório da Comissão Europeia “*Turning FAIR into Reality: Report and Action Plan*” (2018), a efetivação desses princípios requer mudanças significativas na cultura e prática de pesquisa. Além disso, sua implementação depende que seja constituído um Ecossistema FAIR. Esse Ecossistema é um “modelo que apresenta os componentes mínimos necessários para permitir a criação, curadoria e reuso de Objetos Digitais FAIR de forma eficaz e sustentável” (Comissão Europeia, 2018, p. 17). Isso porque FAIR deve ser aplicado de forma ampla a todos os objetos (incluindo metadados, identificadores, software e DMPs) que são essenciais para a prática da pesquisa (HODSON *et al.*, 2018., slide 10, tradução nossa).

Sayão, Sales e Carvalho Segundo (2024) apontam que a ineficiência das práticas atuais de gestão de dados, que frequentemente leva à invisibilidade e inacessibilidade de muitos ativos digitais, motivou a proposta desses Objetos Digitais FAIR (*FAIR Digital Object* - FDO). O FDO é uma unidade de conhecimento identificada por Identificadores Persistentes (*Persistent Identifiers* - PIDS) e rica em metadados que facilitam sua interação com sistemas automatizados. Ainda em desenvolvimento, o modelo de FDO tem sido amplamente testado e aplicado na academia e em setores industriais.

Figura 8 – Modelo de Objetos Digitais FAIR - FDO



Fonte: Comissão Europeia (2018, p. 35).

A Figura 8 ilustra que um objeto digital em sua forma mais básica é um fluxo de *bits* que, para adquirir significado e aderir aos princípios FAIR, necessita de camadas semânticas que o enriqueçam e viabilizem o reuso (Comissão Europeia; 2018). Isso implica não apenas no uso de PIDs, mas também de formatos comuns, dados acompanhados de código, metadados ricos e licenciamento claro (Hodson *et al.*, 2018, slide 10)

O relatório da Comissão Europeia (2018), portanto, especifica que

- Os identificadores precisam sejam únicos e persistentes (exemplos: DOI, Handle e URN). Essa característica garante a criação de *links* estáveis para o objeto, rastreando citações e reusos. Essa mesma lógica deve ser aplicada a outros elementos relacionados, como autores de dados (ORCIDs), projetos (RAIDs), financiadores e recursos de pesquisa associados (RRIDs).
- Os padrões e códigos devem estar em formatos de arquivo comuns e, idealmente, abertos. Essa escolha facilita o reuso por um público mais amplo, tanto pela familiaridade do formato quanto pela disponibilidade de *software* compatível. Uma documentação clara desses formatos é essencial para a preservação. Adicionalmente, os dados devem ser acompanhados pelo código utilizado em seu processamento e análise.

- Os objetos digitais, para que sejam avaliáveis e reutilizáveis, devem ser complementados por metadados e documentação suficientes para sua descoberta. Contudo, para compreender o contexto de criação (como, porquê, quando e por quem), informações de proveniência, enquanto a história dos objetos e sobre como eles vieram a ser o que são²⁴, devem estar mais detalhadas possíveis. Além disso, para promover um (re)uso mais amplo, os objetos digitais devem ser enriquecidos com uma variedade de atributos relevantes e uma licença de uso clara e acessível.

Diante dessa perspectiva, Sayão, Sales e Segundo (2024) apresentam que estudos e testes têm revelado que o conceito de FDO altera, de maneira significativa, a noção sobre o que deve ser um *dataset* (conjunto dados). Os autores complementam que esse cenário traz impactos relevantes para a gestão de dados, a qual espera-se que, a partir de sua concreta implementação, os dados cumpram, ao menos, uma parte do que é preconizado pelos Princípios FAIR.

É importante, no entanto, ressaltar que “a realização do FAIR depende de um ecossistema de componentes essenciais: Políticas, Planos de Gestão de Dados, Identificadores, Padrões e Repositórios (Hodson *et al.*, 2018., slide 11, tradução nossa). Assim, por se trata de um modelo com abordagem generalista, é preciso que a implantação de infraestruturas e práticas se deem a partir da observação do contexto específico das comunidades científicas e suas possibilidades de adoção (Sayão; Sales, 2021).

A pretensão de implantar os princípios FAIR, todavia, não é um consenso para todas as comunidades de pesquisa. A Astronomia e a Linguística, por exemplo, já utilizam princípios com fins semelhantes antes da definição de FAIR. Além disso, a Comissão Europeia (2018) revela que algumas áreas encontram-se relutantes, seja por não subscreverem o compartilhamento de dados ou fluxos de trabalho automatizados, seja pela falta de recursos disponíveis para tal implementação ou pelo fato de já contarem com uma forma “suficientemente boa” de compartilhamento de seus dados.

Nesse sentido, julga-se relevante apresentar como exemplo os princípios CARE, criado para governança de dados indígenas. O CARE, assim como FAIR, é um acrônimo

²⁴ Bettivia; Cheng; Gryk (2023) chamam essa proveniência de “proveniência retrospectiva”.

para *Collective benefit* (benefício coletivo), *Authority to control* (autoridade para controlar), *Responsibility* (responsabilidade) e *Ethics* (ética). Esses princípios foram elaborados no ano de 2018, em Botswana, na Semana Internacional de Dados, na Plenária da *Research Data Alliance* “Princípios de Soberania de Dados Indígenas para a Governança de Dados Indígenas”. Desde então, CARE vem ganhando espaço nas reflexões acerca da implementação de tratamentos adequados a esses dados²⁵.

Segundo disponível no site *Global Indigenous Data Alliance*²⁶, os princípios existentes dentro do movimento de dados abertos se concentram principalmente nas características dos dados com fins de aumentar seu compartilhamento entre entidades, desconsiderando diferenciais de poder e contextos históricos.

Portanto, o exemplo dos Princípios CARE é oportuno por transparecer a complexidade do universo imbricado no tratamento de dados de pesquisa e reforçar que os Princípios FAIR possuem um escopo com limitações que devem ser complementadas por novas propostas que venham a contribuir na superação de lacunas geradas por necessidades específicas.

De toda forma, Caregnato *et al.* (2019) esclarecem que os Princípios FAIR são um marco importante na busca por orientações quanto à gestão de dados de pesquisa e a ampliação de seu valor para reuso. Ao constatarem isso, os autores (2021) apontam que a definição de FAIR foi motivada a partir da defesa do compartilhamento e da luta por superar obstáculos quanto à sua reutilização.

Para as áreas que tem interesse em sua implementação, a *GOFAIR Initiative*²⁷ aponta que o processo de FAIRificação possui sete etapas: 1) recuperar dados não FAIR, 2) analisar esses dados, 3) definir os modelos semânticos, 4) tornar esses dados linkáveis, 5) atribuir licenças, 6) definir conjunto de metadados e 7) implantar os recursos de dados FAIR. A GOFAIR reforça que metadados ricos e adequados são importantes para apoiar todos os aspectos dos princípios FAIR. Contudo, reconhece que maioria dos requisitos

²⁵ Dados indígenas incluem dados coletados por governos e instituições sobre os povos indígenas e seus territórios, sendo intrínsecos à capacidade desses povos em exercer seus direitos humanos e responsabilidades.

²⁶ Disponível em: <https://www.gida-global.org/care>. Para um entendimento mais aprofundado, é possível encontrar textos recentemente publicados nesse site.

²⁷ A GO FAIR foi criada em 2017 na Europa como “uma iniciativa de baixo para cima orientada pelas partes interessadas e autogovernada que visa implementar os princípios FAIR a partir da oferta de um ecossistema aberto e inclusivo para indivíduos, instituições e organizações que trabalham juntos por meio de Redes de Implementação que atuam em três pilares: GO CHANGE, voltado à mudança cultural, GO TRAIN, voltado à capacitação, e GO BUILD, voltado à tecnologia.

para que os dados sejam localizáveis e acessíveis pode ser alcançada no nível de metadados, ao passo que a interoperabilidade e o reuso desses dados exigem mais esforços no nível dos dados (GO FAIR, [s.d.]).

Considera-se que ser localizável precede aos outros princípios, enquanto ser (re)usável implica numa dependência a eles. Em suma, o primeiro passo para (re)usar dados, objetivo final do FAIR, é encontrá-los. Para isso, metadados legíveis por máquina são considerados um componente essencial do processo de FAIRificação.

Uma vez que os dados necessários são encontrados, é preciso que o usuário saiba como eles podem ser acessados (se é necessário autorização ou autenticação, por exemplo). Já para que sejam analisados, armazenados e processados, esses dados, que geralmente são integrados com outros, precisam interoperar com aplicativos ou fluxos de trabalho.

Como a principal característica dos princípios FAIR é a oferta de um conjunto de orientações concisas, de alto nível, que valem para qualquer domínio e que devem ser aplicadas não somente aos dados, mas também aos metadados, aos identificadores, ao *software* e aos planos de gestão de dados, então eles se apresentam como facilitadores e orientadores do reuso (Caregnato *et al.*, 2019, p. 198, grifo nosso).

No entanto, deve ser compreendido que o (re)uso, conforme explicitado anteriormente nesta tese, também possui uma intrínseca relação entre a subjetividade, apresentada por LaFlammer *et al.* (2022) e o papel fundamental da confiança nos dados para sua adoção por outros pesquisadores, de acordo com Van de Sandt *et al.* (2019).

Adiciona-se a importância da proveniência dos dados para avaliar sua utilidade e atribuir créditos aos responsáveis por sua produção e manutenção, conforme apontam Moreira *et al.* (2019). Dessa forma, ao trazer os entendimentos concernentes à confiabilidade, discutida na subseção 2.3.2 desta tese, aponta-se para a necessidade de estruturação e registro cumulativo dos metadados de proveniência para fins de documentar dados mas, também, agentes e eventos envolvidos ao longo de todo seu ciclo de vida.

É interessante reforçar que, conforme expõe Wilkinson *et al.* (2016), os princípios FAIR surgem após o reconhecimento da necessidade de que as vastas quantidades de informação geradas pela ciência contemporânea intensiva em dados fossem encontráveis, acessíveis, interoperáveis e reutilizáveis. Todavia, esses princípios se limitam a destacar

o que espera-se para o tratamento desses (meta)dados, sobretudo sob a perspectiva dos aspectos técnicos.

Como apontam Boeckhout *et al.* (2018), os Princípios FAIR indicam o caminho para o compartilhamento de dados mais sistemático, desde que desafios éticos, metodológicos e organizacionais também sejam considerados.

FAIR também não diz nada sobre a qualidade do objeto. Os princípios dizem respeito apenas aos aspectos técnicos do fornecimento de acesso, não oferecendo nenhum julgamento sobre a qualidade do conteúdo do objeto ou como isso será mantida. Esta é uma área em **que os conceitos de RDM²⁸ e DP²⁹ podem fortalecer significativamente o FAIR** (Lindlar *et al.*; 2020, p. 13, tradução nossa, grifo nosso).

Conforme Sales *et al.* (2021, p. 26) afirmam, os princípios FAIR representam um avanço irreversível e cada vez mais presente nas práticas globais de gestão de dados. Para os autores, a perspectiva futura é que os dados se tornem insumos essenciais para a geração de conhecimento, inovação e até mesmo para a sobrevivência, o que depende diretamente de como são e serão gerenciados ao longo de seu ciclo de vida.

Apesar de fornecerem diretrizes valiosas para o (re)uso, esses princípios priorizam aspectos como formatação e acessibilidade. Em contrapartida, a Gestão de dados de pesquisa (RDM) e a Preservação digital (DP), abarcadas pela curadoria digital dos dados de pesquisa, abordam questões mais amplas e necessárias para o (re)uso efetivo e qualitativo a longo prazo, ao incorporar quesitos de confiabilidade a esses dados.

Diante disso, é preciso reconhecer que os princípios FAIR fornecem diretrizes para maximizar o valor imediato e potencial de descoberta dos dados de pesquisa. Todavia, enquanto FAIR foca na otimização dos dados de pesquisa em termos de acesso e uso, Higman, Bangert e Jones (2019) esclarecem que a gestão de dados de pesquisa facilita sua recuperação e compreensão, diminui sua suscetibilidade de perdas e aumenta sua capacidade de serem utilizados durante um projeto ou no futuro.

3.2 A GESTÃO DE DADOS DE PESQUISA COMO FUNDAMENTO PARA A PRESERVAÇÃO DIGITAL QUALITATIVA

²⁸ Research Digital Management (RDM) que traduzido para o português significa: gestão de dados de pesquisa.

²⁹ Digital Preservation (DP) que traduzido para o português significa: preservação digital.

Diante do cenário científico atual, baseado em produção de dados complexos e volumosos, Marín-Arraiza e Vidotti (2019) informam que, cada vez mais, as tarefas de gestão de dados vêm se constituindo como uma ação muito importante. Os dados da pesquisa eram produzidos, esquecidos, desorganizados e não acessíveis nos arquivos pessoais dos pesquisadores ou das instituições (Estevão; Arns; Strauhs, 2019). Todavia, com a aplicação das ações da gestão de dados de pesquisa³⁰ (*Research Data Management – RDM*), eles passam a constituir um novo insumo para a prática científica.

De acordo com a Terminologia da CODATA RDM (2023), a gestão de dados de pesquisa engloba

Armazenamento, acesso e preservação de dados criados ou coletados no curso da pesquisa. **As práticas de gerenciamento de dados de pesquisa abrangem todo o ciclo de vida dos dados**, desde o planejamento da investigação até sua condução, e desde o backup dos dados conforme são criados e usados até a preservação a longo prazo dos dados entregues após a conclusão da investigação da pesquisa. Atividades e questões específicas que se enquadram na categoria de gerenciamento de dados incluem: **Nomeação de arquivos (a maneira correta de nomear arquivos de computador); controle de qualidade de dados e garantia de qualidade; acesso a dados; documentação de dados (incluindo níveis de incerteza); criação de metadados e vocabulários controlados; armazenamento de dados; arquivamento e preservação de dados; compartilhamento e reutilização de dados; integridade de dados; segurança de dados; privacidade de dados; direitos de dados; protocolos de notebook** (laboratório ou campo) conforme necessário (CODATA RDM, 2023, p. 30, tradução nossa, grifo nosso).

Essa definição amplia e abrange atividades que, atualmente, na literatura, se encontram sobrepostas ao escopo do conceito de curadoria digital. Assim, complementa-se o entendimento sobre a gestão de dados com o que é exposto na “Política de gestão, compartilhamento e abertura de dados para pesquisa – princípios e diretrizes”, elaborada pela Fiocruz, em 2020. Nessa Política considera-se que

A gestão adequada dos dados compreende uma série de etapas - identificação, catalogação, armazenamento, preservação- que são relevantes para permitir o seu reuso pelo próprio investigador e seu grupo de pesquisa futuramente, bem como para **subsidiar as decisões do grupo sobre o tipo, a forma, o grau e o momento de compartilhamento e de abertura dos dados** para outros grupos de pesquisa e o público em geral (FIOCRUZ, 2020, p.5, grifo nosso).

³⁰ Nesse sentido, considerando que já há um forte consenso no uso deste termo e de sua conceituação, conforme já foi melhor detalhado nesse estudo, opta-se por utilizar gestão de dados de pesquisa mas sob a compreensão de que nela estão incluídos também os metadados.

A abordagem da Fiocruz (2020) refina certa diferenciação conceitual entre a curadoria digital e a gestão de dados de pesquisa. À medida em que aponta a gestão de dados como subsídio às decisões sobre seu compartilhamento e abertura, essa definição possibilita entender a importância dessa atividade antes que sejam disponibilizados em repositórios com fins de gestão e compartilhamento.

Compreende-se que a gestão de dados de pesquisa quando adequadamente implementada possibilita o aprimoramento da curadoria digital desses dados. Conforme Stasser (2015), dados quando bem gerenciados são mais facilmente usados e reusados, o que reflete em mais colaboração para pesquisadores e o máximo de retorno sobre o recurso investido.

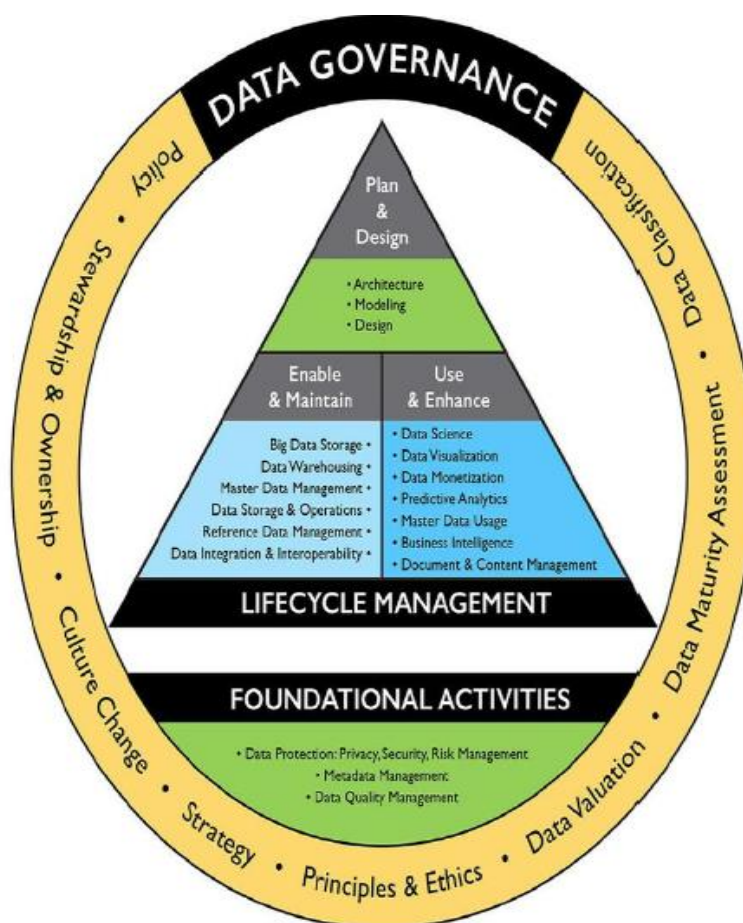
É relevante apontar, que a gestão de dados de pesquisa não depende da execução da curadoria digital para ser efetivada. Enquanto a realização da curadoria digital pressupõe-se a participação de um curador, a gestão de dados de pesquisa, geralmente é realizada pelo próprio pesquisador ou equipe de pesquisa.

A gestão de dados de pesquisa, para ser efetivada, precisa estar inserida sob um panorama de governança que fomente sua implementação sustentável. De acordo com 2ª edição do *Data Management Association International - Data Management Body of Knowledge*³¹ (DAMA - DMBOK, 2017), a estrutura da gestão de dados engloba atividades de planejamento e projeção, ativação e manutenção, uso e aprimoramento. Além disso, essa publicação (2017) indica que essas atividades devem estar fundamentadas por ações essenciais que devem ser realizadas durante todo o ciclo de vida dos dados: proteção dos dados, a gestão de metadados e a gestão da qualidade dos dados.

O DAMA – DMBOK (2017) aponta que a proteção de dados envolve quesitos de segurança visto que dados seguros demandam que os riscos associados a eles sejam mitigados. A gestão de metadados implica que metadados sejam geridos enquanto uma forma de dados, até porque, a gestão dos outros dados depende que isso seja feito. Já a gestão da qualidade ocupa grau de importância pelo fato de que dados de baixa qualidade representam custo e risco, em vez de valor. A publicação, então, reforça a necessidade de que sejam estabelecidos padrões explícitos para aferição dessa qualidade.

³¹ “DAMA® Data Management Body of Knowledge (DAMA-DMBOK®) é uma estrutura reconhecida globalmente que define os princípios básicos, as melhores práticas e as funções essenciais do gerenciamento de dados. Desenvolvido pela DAMA International®, este guia fornece uma base estruturada para profissionais, organizações e educadores entenderem, implementarem e melhorarem continuamente as práticas de gerenciamento de dados”, disponível em <https://dama.org/learning-resources/dama-data-management-body-of-knowledge-dmbok/>. Acesso em: 10 out. 2025.

Figura 9 – Estrutura de gestão de dados – DAMA-DMDBOK



Fonte: DAMA-DMBOK (2017, p. 87).

Na Figura 9, é possível visualizar que as ações essenciais para viabilizar agregação de valor aos dados a partir de melhorias no *design* dos dados, facilitação do seu uso e redução de custos para sua manutenção, encontram-se como centrais e envoltas pelas ações inerentes à governança de dados, a qual possa garantir que esses dados sejam usados de forma estratégica com práticas de gestão de dados alinhadas a políticas, princípios e estratégias adequadas, além de promover a mudança organizacional necessária para o sucesso de sua implementação.

Alguns princípios norteiam a gestão de dados, dentre os quais, ao considerar os dados de pesquisa como objeto nuclear desta tese, destacam-se:

[...]

- **A gestão de dados deve levar em conta uma variedade de perspectivas:**
Os dados são fluidos. A gestão de dados deve evoluir constantemente para

acompanhar as formas como os dados são criados e usados e os consumidores de dados que os utilizam;

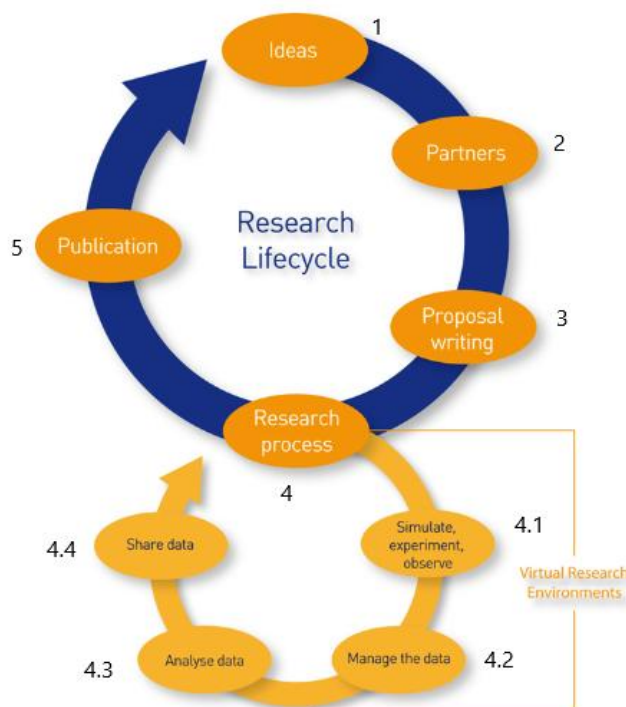
- **A gestão de dados é a gestão do ciclo de vida:** os dados têm um ciclo de vida e a gestão de dados requer a gestão de seu ciclo de vida. Como os dados geram mais dados, o próprio ciclo de vida dos dados pode ser muito complexo. As práticas de gestão de dados precisam levar em conta o ciclo de vida dos dados;

- **Diferentes tipos de dados têm diferentes características de ciclo de vida:** E por esse motivo, eles têm diferentes requisitos de gestão. As práticas de gestão de dados precisam reconhecer essas diferenças e ser flexíveis o suficiente para atender a diferentes tipos de requisitos de ciclo de vida de dados;

- **A gestão de dados inclui o gerenciamento dos riscos associados aos dados:** além de serem um ativo, os dados também representam risco para uma organização. Os dados podem ser perdidos, roubados ou mal utilizados. As organizações devem considerar as implicações éticas de uso de seus dados. Os riscos relacionados aos dados devem ser gerenciados como parte do ciclo de vida dos dados (DAMA-DBOK, 2017, p. 53, tradução nossa, grifo nosso).

Por tratar de um processo contínuo que deve considerar uma variedade de perspectivas e com diferentes características de ciclo de vida, compreende-se a necessidade de que essa gestão seja estruturada e alinhada ao contexto no qual se insere. Portanto, para esta tese, traz-se a gestão de dados sob a ótica da pesquisa científica. Dessa forma, entende-se que essa gestão deve estar integrada não apenas ao ciclo de vida dos dados de pesquisa como também do ciclo de vida da própria pesquisa.

Figura 10 - Ciclo de vida da pesquisa integrada ao ciclo de vida de seus dados



Fonte: Adaptado pela autora com base em *JISC Research Lifecycle* (arquivado em webarchive.nationalarchives.gov.uk).³²

Em observância à figura 10, compreende-se que o ciclo de vida da pesquisa engloba 5 (cinco) grandes etapas. Segundo Liu (2017), essas etapas se distribuem em “ideias” (1) e “parcerias” (2), que configuram o estágio preparatório da pesquisa. Já a “proposta de redação” (3) junto ao “processo de pesquisa” (4) e as ações a ele vinculadas (simular, experimentar, observar (4.1), gerir (4.2), analisar (4.3) e compartilhar os dados (4.4)) incorporam o estágio de pesquisa e desenvolvimento. Por fim, tem-se a “publicação” (5) que, segundo o autor, coroa a fase de resultados ou saídas.

Ao analisar o diagrama da JISC (figura 10) junto à proposta de gestão de dados do DAMA-DMBOK (2017), considera-se que as visões até então apresentadas nesta tese: curadoria digital, ciclo de vida dos dados, ciclo de vida da pesquisa, demonstram a inserção da gestão de dados no cenário do desenvolvimento científico.

No entanto, sob a perspectiva de abertura da ciência e do incremental estímulo ao (re)uso de dados, dinâmicas que embasem a qualidade e a efetividade da gestão de dados de pesquisa são requeridas para apoiar sua efetividade prática.

Nota-se que o ciclo da pesquisa não se finda quando seu produto final é publicizado. Dessa forma, a partir da observância de novas demandas que devem ser atendidas, o ciclo de vida dos dados de pesquisa passou a conter outras atividades imbricadas. Por conseguinte, reflexões mais aprofundadas e com cunho de proposta passam a ser incorporados ao escopo dessa gestão.

3.2.1 A gestão, a abertura e a FAIRificação de dados de pesquisa

Ao reconhecer a forte conjunção entre o movimento de abertura dos artefatos de pesquisa, em especial dos dados e seus metadados, e o incentivo à adesão dos princípios FAIR, entende-se que abrir e FAIRificar não são termos concorrentes. Mesmo que cada um deles tenda a fortalecer o outro, tão pouco são co-dependentes. Um dado de pesquisa pode ser aberto e FAIR, aberto ou FAIR ou nenhum dos dois.

³² Disponível em:

<https://webarchive.nationalarchives.gov.uk/ukgwa/20140702225057/http://www.jisc.ac.uk/whatwedo/campaigns/res3/jisichelp.aspx#ideas>. Acesso em: 05 out. 2025.

De acordo com o Relatório da Comissão Europeia (Hodson *et al.*, 2018), a extração máxima dos benefícios aspirados pela Ciência Aberta - o (re)uso em larga escala, pode vir a ocorrer quando há o enquadramento de ambos os conceitos às práticas científicas.

Sejam os dados abertos, abertos com restrição ou fechados, para permitir o melhor aproveitamento da revolução digital, acelerar a pesquisa, fomentar o poder de análise das máquinas em escala, para fins de transparência, reprodutibilidade e utilidade social, os objetos digitais que tenham sido criados e utilizados para o desenvolvimento de uma pesquisa científica precisam ser FAIR (Hodson *et al.*, 2018).

Considera-se, portanto, que os dados podem ser compartilhados sob algum controle de acesso. Lindlar *et al.* (2020) reforça ser possível fornecer apenas metadados sobre um objeto até então fechado. Quando é esse o caso, as expectativas relativas à FAIRificação recaem sobre os metadados, o que também pode ser aplicado a outras saídas, como o código, por exemplo. Para algumas disciplinas, inclusive, como Engenharia e Ciência da computação, frequentemente, o acesso aos metadados pode ser mais relevante do que a visualização do dado em si.

Facultar metadados detalhados e bem descritos viabilizam a descoberta e a correta referência da pesquisa realizada. Ademais, o acesso a metadados pode amparar a produção de novos conhecimentos. Tem-se, por exemplo, um conjunto de dados disponível no Portal de Dados Abertos da NASA intitulado “*meteorite landings*” (aterrisagem de meteoritos). Essa coleção franqueada abertamente ao acesso se refere a todos os meteoritos conhecidamente pousados na Terra, contabilizando 45.716, até o momento consultado (junho/2024).

Sua apresentação é composta por uma planilha intitulada *Visualization*, cujas linhas correspondem a um pouso e para cada uma tem-se 10 colunas que as correlacionam a informações catalogadas, em padrão de metadados (*name, id, nametype recclass, mass(g), fall, year, reclat, reclong, geoLocation*).

Nesse caso, é possível identificar que o dado em si é um objeto físico³³: o meteorito. Portanto, mesmo que coletado, armazenado adequadamente e disponibilizado para eventuais consultas, ele possui um potencial científico limitado para fins de (re)uso.

³³ Segundo CCSDS 653.0-M-1 (2024), objeto físico é um objeto com propriedades fisicamente observáveis que representam informações consideradas oportunas para serem adequadamente documentadas para preservação, distribuição e uso independente.

O que importa na oportunidade de geração de novos usos são as informações inerentes ao meteorito. Ou seja, são os metadados que apoiam o desenvolvimento de novas análises.

Sob essa perspectiva, é possível observar a permanência de um cenário científico que conta com dados físicos e digitais e que, independem da forma (podendo ser trabalhados em conjuntos híbridos, inclusive). Contudo, dados de pesquisa sem informações que os contextualize podem não ter significado ou confiabilidade suficientes para serem (re)usados.

Ao avaliar especificamente o exemplo da NASA, é possível entender que tais meteoritos com as referidas informações correlacionadas, que nesse caso estão dispostas enquanto dados registrados digitalmente, corroboram com a expressão que explicita o conceito de metadados: “dados sobre dados” (Sayão; Sales, 2015, p. 29). Esses dados sem dados, poderiam ser identificados como uma rocha apenas.

Diante disso, reforça-se a importância do registro de metadados com informações julgadas relevantes, mesmo que o dado em si não esteja franqueado ao acesso, pelo menos não em formato digital. A disponibilização dos metadados propicia visibilidade e disseminação amplificadas sobre a pesquisa em questão.

Todavia, para embasar a decisão de quais recursos da pesquisa podem e/ou devem ser abertos é preciso que algumas particularidades sejam identificadas e consideradas. Essa decisão deve ser norteada no momento do planejamento da pesquisa, o que possibilita que os propícios tratamentos sejam adequadamente executados.

Quanto a isso, questões sobre: de que forma e sob quais critérios esses recursos poderão ser acessados e compartilhados? Caso haja necessidade, quais fundamentos motivam sua restrição ou fechamento? Há interesses e condições para aplicação dos princípios FAIR - ou não- e em qual momento da pesquisa ou do ciclo de vida do dado? Quais os agentes precisam estar envolvidos nessa dinâmica? Dentre outras indagações.

Assim, identifica-se que “FAIR” e “aberto” são termos conceitualmente distintos, mas passíveis de serem sobrepostos. De toda forma, é relevante entender que entre eles há um ponto em interseção: suas implementações estão embricadas na realização de ações inerentes à gestão dos dados de pesquisa.

Isso porque o planejamento da criação/coleta de dados e das atividades inerentes ao seu gerenciamento adequado, quando realizado durante os estágios preliminares e iniciais da pesquisa, com a definição de fontes, fluxos, formatos, padrões, processos,

métodos de registro e a documentação necessária, interferem diretamente na capacidade de abrir e/ou FAIRificar os dados.

3.2.2 Planos de Gestão de Dados – PGD (*Data Management Plan – DMP*)

Instituições de fomento à pesquisa têm incentivado a abertura, o compartilhamento e o (re)uso dos dados de pesquisa. Para que isso ocorra, conforme já exposto na seção primária 3, é preciso que sejam efetivadas ações inerentes à curadoria desses dados. Dessa forma, conforme indicam Sayão; Sales (2012), cada vez mais tem sido exigida a elaboração de um Plano de Gestão de Dados (PGD) como requisito para o financiamento de projetos. Quanto a isso, podem ser citados os seguintes exemplos: Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, *National Institutes of Health* -NIH e *Horizon Europe*.

Revistas científicas de alto impacto também vêm exigindo documentação relativa à disponibilização dos dados utilizados nas pesquisas por elas publicados. Essa documentação exige previsões e planejamentos que, mesmo indiretamente, incentivam que Planos de Gestão de Dados sejam elaborados. A exemplo disso, tem-se a *Nature*, *The Lancet*, *PLOS One*.

De acordo com o “Guia de Gestão de Dados de Pesquisa para Bibliotecários e Pesquisadores”, produzido no âmbito da Comissão Nacional de Energia Nuclear (CNEN) por Sayão e Sales (2015, p.15), “o PGD não é um documento burocrático [...]. Ele deve ser tratado como uma carta de intenções que considere o que realmente é necessário para a preservação, compartilhamento e reuso dos dados”.

O Plano de Gestão de Dados (PGD) abarca, de uma forma geral, informações relativas à Coleta de dados, Metadados e Documentação, Ética e conformidade legal, Armazenamento e *Backup*, Seleção e Preservação, Compartilhamento de dados e Responsabilidades e Recursos. Segundo o DCC (2013), esse documento fornece contexto sobre as intenções e escolhas dos criadores e, portanto, oferecem informações úteis tanto para os curadores quanto para possíveis novos usuários dos dados, o que incrementa ainda mais sua relevância.

Cabe destacar que desde setembro de 2019 vem sendo difundido por um grupo de trabalho da *European Open Science Cloud* (EOSC), o *Data Management Plans Machine*

Actionable - maDMPs (Planos de gestão de dados acionáveis por máquina). No entanto, em um artigo intitulado “*Making Data Management Plans Machine Actionable: Templates and Tools*”, Philipson, Hasan e Moa (2023) expõem que “o desenvolvimento de ferramentas para maDMPs está progredindo rapidamente e parece convergir para um padrão comum. No entanto, ainda há uma imensa quantidade de trabalho para chegar lá”.

Para padronizar a elaboração de PGD, a Fiocruz desenvolveu a proposta de *template* do Plano de Gestão de Dados da Fiocruz (disposto no Anexo A). Apresentado no artigo “Plano de gestão de dados fair: uma proposta para a Fiocruz” (De Oliveira Veiga *et al.*, 2019), esse padrão de formulário foi elaborado por profissionais do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT/Fiocruz).

Para os autores (2019),

A criação de um PGD alinhado aos princípios FAIR que respeite os imperativos inerentes aos diferentes domínios disciplinares do campo da saúde **é um desafio necessário** que está sendo enfrentado pela Fiocruz **para garantir o sucesso da gestão, compartilhamento e reuso dos seus dados de pesquisa.** Esse desafio foi assumido pelo Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), indo **ao encontro do fortalecimento do novo paradigma das práticas de gestão de dados contidas no âmbito da Ciência Aberta mundial** (De Oliveira Veiga *et al.*, 2019, p. 280, grifo nosso)

Assim como a Fiocruz, a Universidade de Utrech³⁴ disponibiliza em seu site um guia para auxiliar os pesquisadores na elaboração desse planejamento. Sua finalidade é oferecer orientações práticas para preparar, manusear, preservar e compartilhar dados de forma ética e legal. Este guia online, chamado de *Research Data Management Support Guide* (Guia de suporte ao gerenciamento de dados de pesquisa), aborda desde o planejamento inicial até as considerações legais e as ações pós-pesquisa, incluindo boas práticas para o armazenamento e a preservação dos dados.

Ao adentrar na aba “durante a pesquisa”, que orienta sobre como encontrar e armazenar dados e como conduzir a pesquisa de forma eficiente e segura, esse guia se desdobra em mais quatro subníveis, das quais destaca-se a aba “armazenando dados”, onde são descritas boas práticas no armazenamento de dados de pesquisa durante a coleta de dados e na preservação de seus dados após o término do projeto.

³⁴ Tal escolha se deu visto a reconhecida tradição holandesa nas reflexões e práticas relativas à preservação digital e por se tratar de um guia com orientações simples e de fácil aplicação.

Por conseguinte, essa aba se ramifica em três itens: “1 - Boas práticas no armazenamento de dados” (com orientações do I ao VI), “2 – Boas práticas na preservação de dados” (com orientações do VII ao XII) e “3 – Soluções de TI para armazenamento e preservação de dados” (que, inclusive traz um *link* para o documento “ferramentas para armazenar e gerenciar dados”) (Utrecht, [s.d.]).

No tópico VII do item “2 – Boas práticas na preservação de dados”, encontra-se o “especificar quais dados preservar”, conforme quadro abaixo:

Figura 11 – Extrato do guia de suporte ao gerenciamento de dados de pesquisa

VII – Especificar quais dados preservar

Quais dados você selecionará para preservação? As opções são, por exemplo:

- Você vai preservar apenas os dados que sustentam uma publicação científica, ou também outros dados?
- Você preservará os dados quando eles estiverem completamente estáticos (sem alterações esperadas) ou permitirá versões?
- O local onde você decide armazenar seus dados é apropriado para preservar dados pessoais (sensíveis à privacidade)?
- Em que momento e de acordo com quais especificações seus dados serão removidos?

O que exatamente preservar também depende do seu propósito:

- Se você preservar dados principalmente para fins de verificação, de preferência resultados intermediários ou materiais e métodos de análise também são armazenados junto com o fluxo de trabalho.
- Se os dados forem armazenados porque podem ser reutilizáveis no futuro por você mesmo ou por outras pessoas, seu foco será preservá-los de forma a possibilitar novas análises. Nesses casos, você deve armazenar os dados da forma mais bruta possível.

Em ambos os casos, é necessário adicionar documentação suficiente para tornar os dados compreensíveis.

DADOS EXTERNOS DE TERCEIROS

Se você fez uso de dados de outras partes, você terá que contabilizá-los também. Você tem duas opções:

- Combinar com os proprietários que armazenem os dados e os disponibilizem para fins de verificação, pelo menos pelo período obrigatório de armazenamento (dez anos). Você pode então simplesmente consultar o seu armazenamento.
- Tente organizar uma cópia local, que você mesmo pode armazenar pelo período necessário.

Fonte: Adaptado pela autora com base em site Universidade de Utrecht.

Portanto, conforme é possível observar na Figura 11, a criação de um PGD, longe de ser um mero cumprimento de exigências formais, configura-se como um instrumento essencial para orientar o planejamento, o armazenamento e a preservação dos dados de pesquisa. Em complemento, Strasser (2015) esclarece que o PGD é essencial para o planejamento técnico do tratamento dos dados de pesquisa o que, para Turkeyilmaz-Van Der Velden *et al.* (2020), garante sua usabilidade e preservação a longo prazo.

É preciso compreender, todavia, que o PGD deve ser revisitado e atualizado pelos pesquisadores periodicamente, ou sempre que necessário, ao longo do percurso da pesquisa, para garantir a sua consistência. Ao fornecer o contexto da pesquisa e apoiar o planejamento das ações de curadoria, o PGD estabelece um alicerce para a gestão e preservação digital dos dados de pesquisa.

Nos dias de hoje, é desejável que a elaboração de PGD incorpore boas práticas de gestão e preservação alinhadas a princípios como os FAIR, desenvolvidos justamente para refinar os aspectos técnicos que viabilizam o objetivo da curadoria digital.

3.2.3 Repositórios de gestão de dados de pesquisa

Para gerir os dados de pesquisa de forma eficaz, é fundamental dispor de um arcabouço tecnológico e gerencial que apoie as ações ao longo de todo o seu ciclo de vida, principalmente quando se trata de dados em formato digital. Nesse sentido, diante do atual cenário científico, surgem os repositórios digitais de dados de pesquisa.

Esses repositórios, que objetivam gestão e compartilhamento dos dados de pesquisa, tornam-se essenciais na infraestrutura de pesquisa global devido às suas necessidades de expansão e organização. Segundo Sayão e Sales (2016), o desenvolvimento desses repositórios teve por impulso duas demandas principais: as políticas obrigatórias de agências de financiamento e a adoção dos princípios da Ciência Aberta por instituições e pela sociedade.

Os autores indicam que,

O repositório de dados, na qualidade de um sistema digital que integra diversas funções, tem como perspectiva oferecer um ambiente dinâmico e flexível – principalmente pela natureza heterogênea dos dados - para dar apoio à execução dos processos de gestão de dados de pesquisa (Sayão; Sales, 2016, p. 99).

Ademais, conforme disponível no site da Re3Data.org, esses repositórios fazem parte de uma infraestrutura de informação sustentável que deve apoiar o acesso e armazenamento a longo prazo de dados de pesquisa.

A Universidade de São Paulo (USP), ao apresentar seu repositório de dados, o qual adota a nomenclatura de “repositório de dados científicos”, define que trata-se de “ambientes digitais para auxiliar pesquisadores no armazenamento, preservação, disponibilização, gerenciamento e acesso de dados científicos, elementos fundamentais para o ciclo de vida da própria pesquisa” (USP, [s.d.]).

Nesses repositórios são gerenciadas as diferentes versões dos dados que surgem ao longo do seu ciclo de vida, incluindo dados coletados, processados e aqueles que poderão ser (re)usados em pesquisas futuras. De acordo com Rocha *et al.* (2021), eles precisam lidar com a evolução dos dados desde a sua criação até o seu potencial (re)uso.

Adicionalmente, Sayão e Sales (2016) reconhecem a variedade de formatos de repositórios existentes, apontando que podem ser categorizados como institucionais, disciplinares ou temáticos, multidisciplinares e orientados por projetos.

A categoria de repositórios de gestão de dados de pesquisa institucionais é

[...] caracterizada por ser gerenciada e funcionar no âmbito de uma instituição acadêmica, como universidades ou institutos de pesquisa, e são voltados para arquivar dados que são, geralmente, provenientes unicamente das atividades acadêmicas dessas instituições (Sayão; Sales, 2016, p. 101).

Os repositórios de gestão de dados de pesquisa disciplinares ou temáticos, ao seguir normas de campos específicos “são extremamente variados e heterogêneos, refletindo a multiplicidade de disciplinas e a diversidades de dados gerados no contexto da pesquisa científica mundial” (Sayão; Sales, 2016, p. 103).

Enquanto os repositórios de gestão de dados de pesquisa multidisciplinares “reúnem coleções de dados coletados ou gerados por atividades de pesquisa em várias áreas de conhecimento [...] e que sejam provenientes de diferentes instituições de pesquisa” (Sayão; Sales, 2016, p. 103), o que lhes confere a denominação de repositórios de “cauda longa” (Sayão; Sales, 2015).

Já os repositórios de dados de pesquisa orientados por projetos, de acordo com Sayão e Sales (2016, p. 104), “são repositórios cujas coleções de dados são resultados de projetos de pesquisa ou resolução de problemas específicos” e que podem se estender a “coleções de dados armazenados em outros sistemas de informação e banco de dados fora

do escopo científico, gerenciados por órgãos governamentais, empresas privadas e ONGs”.

Sendo assim, independentemente da categoria ao qual se enquadrem, é possível afirmar que os repositórios de gestão de dados de pesquisa são ferramentas tecnológicas voltadas a armazenar, organizar, preservar, descobrir, acessar e compartilhar dados que foram gerados e/ou utilizados, ou ainda em uso, no processo de pesquisa científica. Dessa maneira, esses repositórios servem como um ponto de encontro entre pesquisadores (que depositam seus dados) e partes interessadas ou possíveis (re)usuários.

A utilização desses repositórios vem sendo incentivada e incorporada aos processos científicos que buscam ser alinhados às premissas da Ciência Aberta. Isso porque sua implementação é capaz de promover transparência, reprodutibilidade, colaboração, avanço e inovações do conhecimento científico através da disponibilização e possibilidade de compartilhamento dos dados de pesquisa. Por sua vez, pode promover o (re)uso desses dados.

Com relação a esses repositórios de gestão de dados de pesquisa, é interessante expor que em 2017, em um esforço cooperativo entre o *Data Seal of Approval* (DSA) e o *World Data System do International Science Council* (WDS), sob a égide da *Research Data Alliance* (RDA), foi lançado o *CoreTrustSeal*. Ele define requisitos para oferecer certificados de nível básico para repositórios de dados confiáveis que mantêm dados para preservação a longo prazo.

Essa certificação é concebida como o primeiro passo em uma estrutura global para certificação de repositórios que inclui a certificação de nível estendido (Nestor-Seal DIN 31644) e a certificação de nível formal (ISO 16.363). Na lista de requisitos para a obtenção desse selo, constam informações sobre Contexto, Infraestrutura organizacional, Gestão do objeto digital e Tecnologia de informação e segurança.

Conforme a *CoreTrustSeal Standards and Certification Board* (2022), o tópico Gestão do objeto digital é dividido em 7 assuntos, dentre os quais são apontados:

- R07 Proveniência e autenticidade: “o repositório garante a autenticidade do objeto digital e provê informação de proveniência”.
- R09 Plano de preservação: “o repositório assume a responsabilidade pela preservação a longo prazo e gerencia isso de forma planejada e documentada”.

Dessa forma, entende-se que a curadoria digital dos dados de pesquisa, ao envolver o planejamento, a organização, o armazenamento, o compartilhamento e a preservação ao longo de todo o ciclo de vida desses dados, constitui um fator estruturante para garantir sua qualidade e, conseqüentemente, seu potencial de (re)uso. Esses aspectos são indispensáveis à efetivação da Ciência Aberta.

Ao integrar instrumentos como os Planos de Gestão de Dados e os repositórios de gestão, tal curadoria estabelece as bases conceituais e operacionais que sustentam uma preservação que ultrapassa a dimensão meramente técnica, incorporando atributos relacionados à proveniência, à autenticidade e à confiabilidade dos dados de pesquisa.

Reconhece-se, assim, a relação de interdependência entre gestão e preservação, na qual a primeira orienta e qualifica a segunda, configurando-se como o alicerce sobre o qual se estrutura o fundamento desta tese: a preservação digital qualitativa.

4 PRESERVAÇÃO DIGITAL QUALITATIVA: O REGISTRO DE METADADOS PARA O FORTALECIMENTO DO (RE)USO DOS DADOS DE PESQUISA

A preservação digital, enquanto campo teórico e prático, configura-se como etapa imprescindível no ciclo de vida dos dados de pesquisa, sobretudo diante dos desafios impostos pela crescente produção, complexidade e diversidade dos recursos digitais. Nesta seção, busca-se discutir os fundamentos conceituais que sustentam a preservação digital, com ênfase no Modelo de Referência *Open Archival Information System* (OAIS), reconhecido internacionalmente como base para orientar as ações e responsabilidades envolvidas na manutenção de objetos digitais autênticos, acessíveis e compreensíveis ao longo do tempo.

Destaca-se, nesse contexto, o papel central dos metadados como elementos estruturantes da preservação, uma vez que registram as informações necessárias à verificação da proveniência, evidenciando sua autenticidade e assegurando sua confiabilidade e potencial de (re)uso. Ao abordar o OAIS e suas conexões com outros padrões e instrumentos correlatos, esta seção pretende evidenciar como a gestão adequada dos metadados e das informações de preservação contribui para o fortalecimento de práticas que ultrapassam o âmbito técnico, incorporando dimensões organizacionais, informacionais e arquivísticas que se articulam à proposta desta tese de qualificar a preservação digital dos dados de pesquisa.

4.1 O MODELO DE REFERÊNCIA OAIS COMO BASE CONCEITUAL PARA A PRESERVAÇÃO DIGITAL QUALITATIVA

A realização da preservação digital³⁵ requer uma combinação flexível de conceitos, precauções, políticas, estratégias e ações planejadas e executadas contextualmente, considerando benefícios e riscos. Seu objetivo é garantir a utilidade, a autenticidade contínua e o acesso significativo do conteúdo digital foco da preservação, independentemente de todos os desafios que possam perpassá-lo e afetá-lo ao longo do

³⁵ Esclarece-se que, quando se trata de preservação digital, considera-se a sua aplicação tanto a objetos natos digitais, criados já em formato digital, quanto a representantes digitais, oriundos de processos de digitalização.

tempo. A obsolescência de recursos digitais, advinda dos rápidos e constantes avanços tecnológicos, é um exemplo desse aspecto desafiador enfrentado atualmente (Brown, 2013; Webb, 2014; Baucom, 2019).

Com o fito de superar as complexidades envolvidas na preservação digital, alguns autores, como Arellano (2008), Santos e Flores (2019) e Formenton e Gracioso (2023), agrupam as principais ações recomendadas em estratégias estruturais e estratégias operacionais.

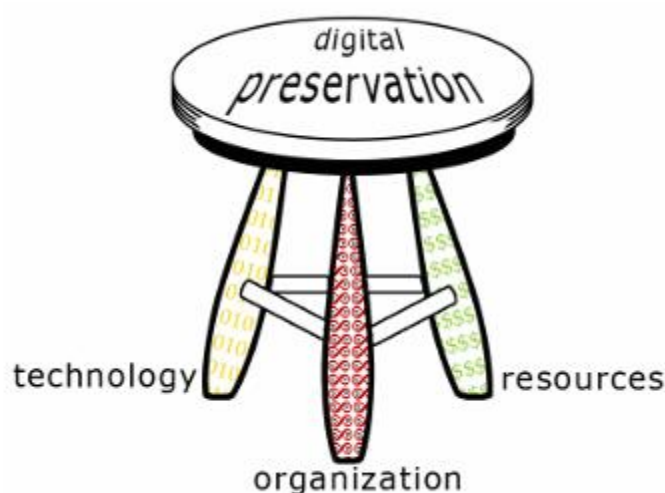
As estratégias estruturais consistem nos investimentos institucionais iniciais direcionados à construção e preparação de um ambiente apropriado para a preservação digital. Elas definem questões como infraestrutura física e tecnológica adequada, assim como o planejamento das atividades em longo prazo. A adoção de padrões abertos, definição de políticas e estratégias institucionais, cálculo de orçamentos e custos envolvidos, seleção para preservação e conformidade legal, treinamento e desenvolvimento de pessoal, adoção de padrões e esquemas de metadados, investimento e montagem de infraestrutura tecnológica e formação de redes de colaboração são exemplos dessas estratégias (Arellano, 2008; Santos; Flores, 2019; Formenton; Gracioso, 2023).

Já as estratégias operacionais, de acordo com Santos e Flores (2020), abarcam ações reais que devem ser executadas nos três níveis dos objetos digitais:

- a) o “físico”, que se restringe ao *hardware* e ao suporte, integrando o refrescamento ou recopia de conteúdos em suportes mais atuais,
- b) o “lógico”, que se refere aos *softwares* e as sequências de bits, englobando emulação, preservação de tecnologia e encapsulamento ou a reunião de tudo o que seja requerido para acessar e interpretar os objetos digitais; e
- c) o “conceitual”, cujo enfoque está na representação visual do objeto, conforme é visto pelo humano, incluindo, portanto, a migração.

É relevante reforçar o entendimento de que a preservação digital não se restringe a aspectos técnicos e tecnológicos. Ela deve contemplar, de maneira conjunta, o apoio organizacional, a infraestrutura tecnológica e os recursos disponíveis, o que Anne Kenney e Nancy McGovern intitularam, em 2003, por *Three-Legged Stool* (conhecido como “Banquinho de três pernas”).

Figura 12 – Banquinho de três pernas da preservação digital



Fonte: Kenney; McGovern (2003).

Na Figura 12, a primeira perna, *Organization* (organização), engloba o reconhecimento da necessidade de preservação digital, o desenvolvimento de políticas, a implementação de programas, o planejamento de preservação e os sinais de comprometimento institucional. Para isso, as autoras (2003) recomendam a utilização do documento “*Trusted Digital Repositories: Attributes and Responsibilities*” (RLG, 2002), como diretriz.

A segunda, *Tehcnology* (tecnologia), inclui *hardware*, *software*, formatos e armazenamento, as redes e segurança, as funções e fluxo de trabalho, os procedimentos, protocolos e documentação e as habilidades técnicas e de arquivamento disponíveis. Kenney e McGovern (2003) recomendam o Modelo de Referência *Open Archival Information System* - OAIS (CCSDS 650.0-M-3, 2024) como documento que pode fornecer boas diretrizes para uma infraestrutura tecnológica forte.

Por último, as autoras (2023) explicitam a terceira perna, *Resources* (recursos), que reflete os ativos como equipe, tecnologia e treinamento.

Kenney e McGovern (2003) consideram que cada perna tem sua importância, mas só as três juntas são capazes de manter o banquinho em pé. Se qualquer uma delas for inadequada ou ausente, o “banco” irá balançar ou tombar, reforçam as autoras.

Adicionalmente, em 2012, no II Seminário de Serviços em Informação em Museus, Millard Schisler propôs a adição do conceito de “educação e formação de profissionais” como a amarração das três pernas deste banquinho. O que, além de reforçar

a terceira perna de Kenney e McGovern (2003), reflete sobre a importância da capacitação de profissionais para a adequada e efetiva execução de todas as ações previstas.

Diante do exposto, compreende-se Conway (2001) quando expõe que o gerenciamento da preservação digital abrange todas as políticas, procedimentos e processos que lidam com a deterioração dos objetos, a prorrogação de sua informação e a intensificação de sua importância funcional.

Reforça-se, então, que preservar abarca a intervenção “direta” sobre os fatores que complexificam e impactam na manutenção do objeto digital e de sua capacidade de ser recuperado e (re)usado. Assim como também engloba o esforço lógico-racional dos agentes envolvidos no planejamento e operacionalização dos recursos (humanos, financeiros e materiais).

Contudo, pela amplitude de todo o escopo inerente à preservação digital, esta tese foca no direcionamento das análises relativas aos aspectos de infraestrutura tecnológica. Tais aspectos estão inseridos nas estratégias operacionais (Arellano, 2008; Santos; Flores, 2019; Formenton; Gracioso, 2003) e na segunda perna do banquinho, que inclui questões sobre tecnologia.

Ademais, frisa-se que “nenhuma estratégia de preservação única oferece todas as vantagens e desvantagens” (Brown, 2013, p. 213). Portanto, a preservação digital deve ser planejada com a perspectiva de que sejam implementadas diferentes estratégias ao longo do tempo. Para tanto, é preciso considerar as especificidades dos recursos digitais e os contextos ao qual estão inseridos. Além disso, há a necessidade de que sejam identificadas as prioridades para sua intervenção.

A essa afirmativa acrescenta-se que “grandes ou pequenas, as soluções implementadas são necessariamente parciais e exigem monitoramento e adaptação contínuos ou pelo menos recorrentes” (Webb, 2014, p. 112).

Assim, ao constatar a demanda de uma abordagem flexível e adaptável com a implementação de diversas estratégias e a consideração de múltiplas variáveis, torna-se fundamental adotar um modelo de referência para orientar as ações de preservação digital e garantir sua efetividade e eficácia.

Nesse sentido, conforme Kenney e McGovern (2003), tem-se o Modelo de Referência OAIS. Esse Modelo apresenta uma base conceitual robusta e adotada por profissionais e instituições responsáveis por ações de preservação digital. De acordo com

CCSDS 653.0-M-1 (2024, p. 1-3, tradução nossa), trata-se de “um dos padrões arquivísticos mais amplamente reconhecidos e aplicados disponíveis hoje”.

A ampla convergência entre profissionais da área reforça a recomendação de que as estratégias de preservação digital se baseiem no CCSDS 650.0-M-3, 2024 “Modelo de Referência *Open Archival Information System* (OAIS)”.

Tal Modelo é caracterizado como uma Prática Recomendada desenvolvida pelo *Consultative Committee for Space Data Systems* (CCSDS). Foi criada em 2002 e recentemente atualizada, em 2024 (atualizando, consequentemente, a Norma ISO 14721:2003 para ISO 14721:2025).

Esse Modelo de referência considera que a preservação digital não pode ser abordada, exclusivamente, sob o ponto de vista técnico, devendo levar em consideração questões organizacionais, legais, industriais, científicas e culturais.

Os autores Donaldson, Zegler-Poleska e Yarmey (2020), consideram o OAIS essencial na prática da preservação digital. Formenton e Gracioso (2023, p. 7), reforçam que o OAIS é “um dos principais modelos de referência para padronização das atividades de preservação digital”. Nesse sentido, esse Modelo vem sendo amplamente adotado ao longo desses últimos 20 anos³⁶.

O OAIS (2024) inclui em seu escopo as modelagens: funcional e de informação, sem que haja a pretensão de definir ou exigir métodos específicos para sua implementação. Assim sendo, em seu modelo informacional o processo de preservação em longo prazo adota uma estrutura conceitual de pacotes de informação que funcionam como um contêiner.

Um Pacote de Informação é um contêiner lógico composto por Objetos de Informação opcionais. Associadas a esse Pacote de Informação estão as Informações de Empacotamento, que descrevem como os componentes de um Pacote de Informação estão vinculados lógica ou fisicamente e como identificar e extrair os componentes (CCSDS 650.0-M-3, 2024, p. 2-5, tradução nossa).

A estrutura conceitual desses pacotes de informação abarca dois objetos de informação: a Informação de Conteúdo (*Content Information* - CI) e a Informações Descritivas de Preservação (*Preservation Description Information* - PDI), que podem ser associados com outros dois tipos de objetos de informação: a Informação de

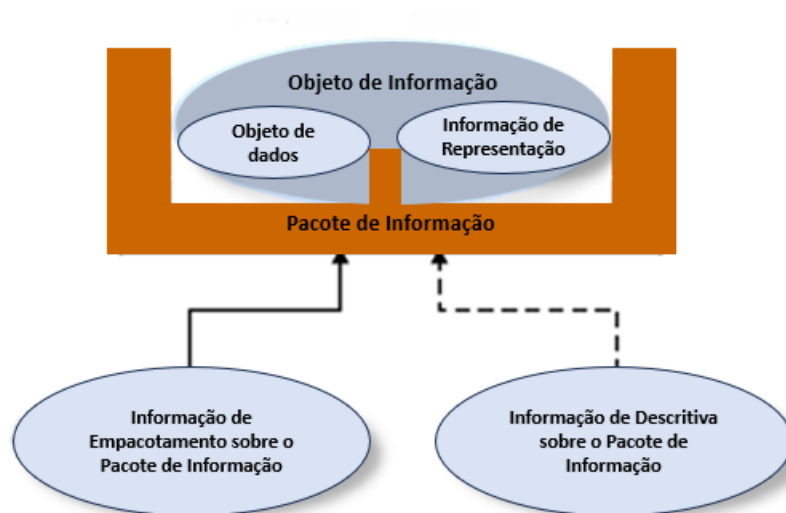
³⁶ Em dezembro de 2024, a CCSDS disponibilizou a versão atualizada do Modelo OAIS.

Empacotamento (*Package Information*) e a Descrição do Pacote (*Descriptive Information*).

Essa integração de diversas informações caracteriza o pacote de informação como “um conjunto de objetos digitais encapsulados” (Santos; Flores, 2020, p. 24).

A Informação de Conteúdo (IC) é constituída da seguinte maneira:

Figura 13 - Conceitos e relacionamentos de um Objeto de Informação em um Pacote de Informação



Fonte: Adaptada pela autora com base em CCSDS 650.0-M-3 (2024, p. 6).

Conforme ilustrado na Figura 13, a IC é composta pelo Objeto de Informação, o qual associa o objeto físico ou digital que se pretende preservar: o Objeto de Dados (OD) (composto por uma ou mais cadeia de bits) às informações necessárias para interpretá-lo, chamadas de Informação de Representação (IR).

Segundo disposto no OAIS (2024), essa Informação de Representação (IR) torna as cadeias de bits significativas por meio de “informações estruturais” (nome ou posição relativa), que geralmente necessitam de “informações semânticas”. Essas informações semânticas podem ter muitas variedades e níveis de complexidade ao incluir “significados especiais associados a todos os elementos de informação estrutura, operações a serem executadas em cada tipo de dado e seus inter-relacionamentos” (CCSDS 650.0-M-3, 2024, p. 66).

Quanto a isso, “a escolha entre preservar a quantidade mínima necessária ou o máximo possível da informação de representação deve considerar também os custos envolvidos para a sua preservação em longo prazo” (Santos; Flores, 2019, p. 12). No

entanto, embora os custos devam ser considerados para fins de planejamento e execução, a importância da manutenção dessas informações de representação, ou seja, os metadados, para que, de fato, a preservação digital possa ser efetivada, também precisa ser relevada.

Os metadados descrevem quais dados uma organização possui, o que eles representam, como são classificados, de onde vieram, como se movem dentro da organização, como evoluíram através do uso, quem pode e não pode usá-los, e se são de alta qualidade. **Os dados são abstratos. Definições e outras descrições de contexto permitem que ele seja compreendido. Eles tornam os dados, o ciclo de vida dos dados e os sistemas complexos que contêm dados compreensíveis. O desafio é que os metadados são uma forma de dados e precisam ser gerenciados como tal.** As organizações que não gerenciam bem seus dados geralmente também não gerenciam seus metadados (DAMA-DMBOK, 2017, p. 59, tradução nossa, grifo nosso).

Em consequência, a preservação digital implica preservar outras informações além daquela que essencialmente é o objeto da preservação em si. Essa questão envolve complexidades organizacionais, técnicas e tecnológicas. Sobre isso, observa-se que, nas últimas décadas, tem havido uma crescente divulgação de estudos que defendem a inexorabilidade da preservação do dado de pesquisa em si, a partir do estabelecimento de estratégias voltadas ao seu acesso e uso, junto às suas propriedades significativas.

Até porque, de acordo com Pires e Rocha (2020)³⁷, também devem ser consideradas as transformações que podem vir a ocorrer no objeto digital ao longo do tempo a partir de decisões que precisam levar em consideração, em maior ou menor grau, as percepções da sua comunidade de usuários que é e está passível de mudanças.

Atualmente, as propriedades significativas vêm sendo representadas no OAIS (2024) enquanto um tipo de “propriedade de informação” chamada de “transformacional” que depende de informações de representação específicas, incluindo informações semânticas, para denotar o que ela significa e como é codificada, além de serem importantes para contribuir com evidências sobre autenticidade.

Para operacionalizar a preservação desses dados e suas propriedades significativas ao longo do tempo, o Modelo OAIS indica uma estrutura conceitual baseada em diferentes tipos de pacotes de informação: Pacote de Informações para Submissão (*Submission Information Package* - SIP), Pacote de Informações para Arquivamento

³⁷ Apresentação sob título “Propriedades Significativas na curadoria digital e suas relações com a Representação da Informação do modelo OAIS” realizada no 3º Fórum de Estudo em Informação, Sociedade e Ciência nos dias 29 e 30 de outubro de 2020, no formato online.

(*Archival Information Package* - AIP) e Pacote de Informações para Disseminação (*Dissemination Information Package* - DIP)

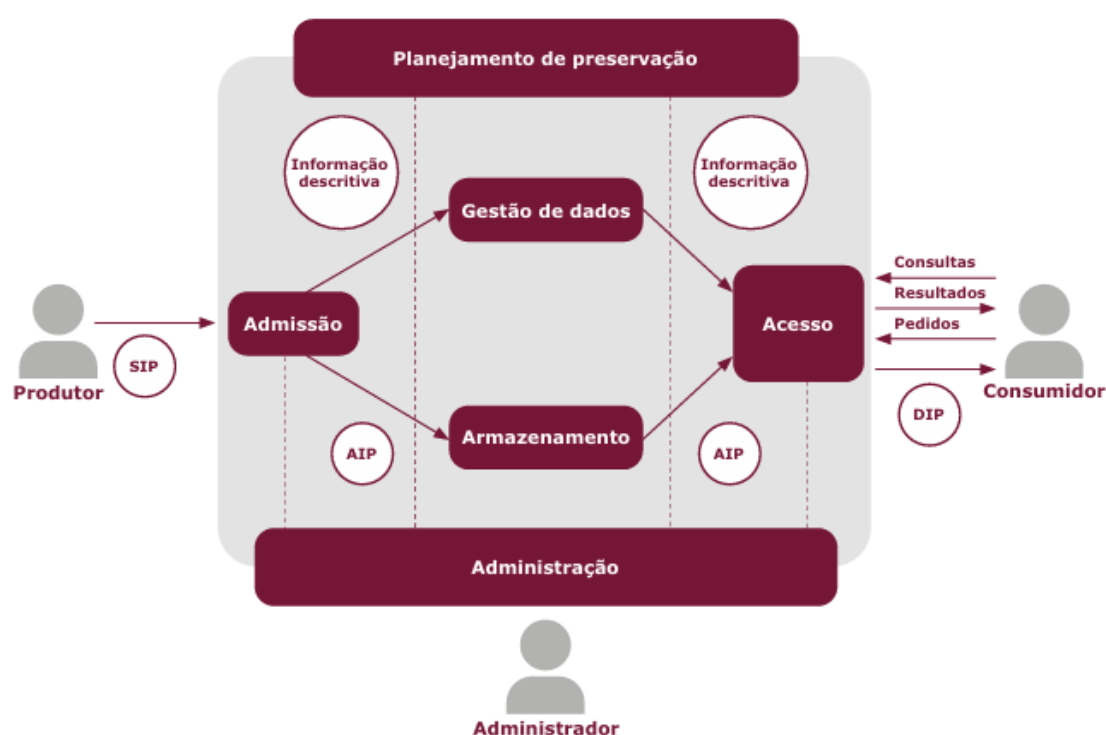
O SIP se constitui pelo empacotamento da Informação de Conteúdo (IC) e os metadados registrados até o momento em que é enviado pelo produtor para admissão em um repositório de preservação digital.

O AIP, objeto alvo da preservação, é constituído pelo SIP processado pelo repositório de preservação digital, quando a ele são agregadas novas informações adicionais correspondentes ao conjunto de ações realizadas durante seu processo de admissão no repositório. Os registros relativos a esse processo são necessários para manter o IC significativo, compreensível, autêntico, acessível e utilizável ao longo do tempo.

Já o DIP é produzido quando o acesso ao objeto digital é solicitado, constituído por IC e metadados que sejam necessários para sua recuperação e (re)uso pela comunidade designada.

Esses 3 (três) pacotes de informações descritos no Modelo Informacional tramitam pelas entidades funcionais do OAIS de acordo com o seguinte fluxo:

Figura 14 – Entidades funcionais do OAIS



Fonte: CONARQ (2023, p. 20).

Nessa perspectiva, conforme pode ser observado pela Figura 14, alguns agentes são envolvidos ao longo do processo de preservação digital. Em conformidade com o CONARQ (2023), os primeiros são os produtores, pessoas ou sistemas que fornecem a informação a ser preservada. Logo após, tem-se o administrador, que estabelece as políticas gerais que governam o repositório de preservação digital. Por fim, há os consumidores, pessoas ou sistemas interessados na informação preservada e que interagem com os serviços OAIS para poder acessá-la.

Assim sendo, o SIP é o pacote que inicia as ações relativas ao processo de preservação digital em um ambiente de repositório. Afirma-se, portanto, que a sua adequada constituição é determinante para a qualidade do que será preservado, inclusive em termos de informações significativas com qualidade.

Entende-se que essas informações são apresentadas pelo registro de metadados. Para Machado (2022), as informações sobre informação, que em relação a um documento, seriam seu título, a sua apresentação e o tema, por exemplo, são chamadas de “metainformações”. Já a CCSDS 650.0-M-3 (2024), sob um espectro mais amplo das informações registradas em metadados, denomina essas informações como “informações adicionais”. São elas que, no contexto do OAIS (2024), compõem o que se define como Informações Descritivas de Preservação (*Preservation Description Information -PDI*).

As PDI são informações essenciais que acompanham a Informação de Conteúdo (IC) e têm como objetivo apoiar a manutenção de sua capacidade em ser preservada e acessada ao longo do tempo. Em outras palavras, a PDI fornece o contexto e a história da IC, ajudando a garantir sua compreensibilidade, autenticidade e confiabilidade para a Comunidade Designada e outros possíveis usuários, tanto no presente quanto no futuro.

Na composição da PDI são identificadas 5 (cinco) informações: Informação de referência, Informação de contexto, Informação de proveniência, Informação de fixidez e Informação de direitos de acesso.

Em relação à IC de um pacote, a “Informação de referência” identifica e, se necessário, descreve os mecanismos usados para o fornecimento de seus identificadores.

A Informação de Referência identifica e, se necessário, descreve um ou mais mecanismos usados para fornecer identificadores atribuídos ao Objeto de Dados de Conteúdo. Também fornece aqueles identificadores que permitem que sistemas externos se refiram, de forma inequívoca, a este Objeto de Dados de Conteúdo em particular. Exemplos desses sistemas incluem sistemas taxonômicos, sistemas de referência e sistemas de registro. No Modelo de

Referência OAIS, a maior parte, senão todas, essas informações são replicadas nas Descrições de Pacotes, que permitem que os Consumidores acessem Informações de interesse (CCSDS 650.0-M-3, 2024, p.4-35, tradução nossa).

A “Informação de contexto” documenta as suas relações com seu ambiente, incluindo o motivo de sua criação e como se relaciona com outros objetos de informação de conteúdo existentes em outros locais.

Já a “Informação de proveniência”, que pode ser vista como um tipo especial de informação de contexto, documenta todo o histórico, relata sua fonte, ou origem, as alterações que possam ter ocorrido desde sua criação, seus custodiadores e uma trilha de auditoria.

A Informação de proveniência documenta a história do Objeto de Dados de Conteúdo. Isso indica a origem ou fonte do Objeto de Dados de Conteúdo, suas Propriedades de Informação a serem preservadas (Propriedades de Informação Transformacional), quaisquer alterações que possam ter ocorrido desde sua criação e quem teve a custódia dele desde sua origem, fornecendo um registro de auditoria para o Objeto de Dados de Conteúdo. Isso dá aos futuros usuários alguma garantia quanto à provável confiabilidade do Objeto de Dados de Conteúdo, à medida que contribui para evidências que suportam a Autenticidade (CCSDS 650.0-M-3, 2024, p.4-35, tradução nossa, grifo nosso).

Quanto à Informação de proveniência é importante frisar que “o arquivo é responsável por criar e preservar as informações de proveniência do ponto de ingestão; no entanto as informações de proveniência anteriores devem ser fornecidas pelo produtor” (CCSDS 650.0-M-3, 2024, p. 1-15). Isso indica que a Informação de proveniência deve, então, conter metadados, de forma cumulativa e contínua, que demonstrem informações sobre todo o ciclo de vida do objeto digital.

Ademais, a “Informação de fixidez” fornece verificações de integridade de dados ou chaves de validação/verificação usadas para garantir que não houve alteração de forma não documentada.

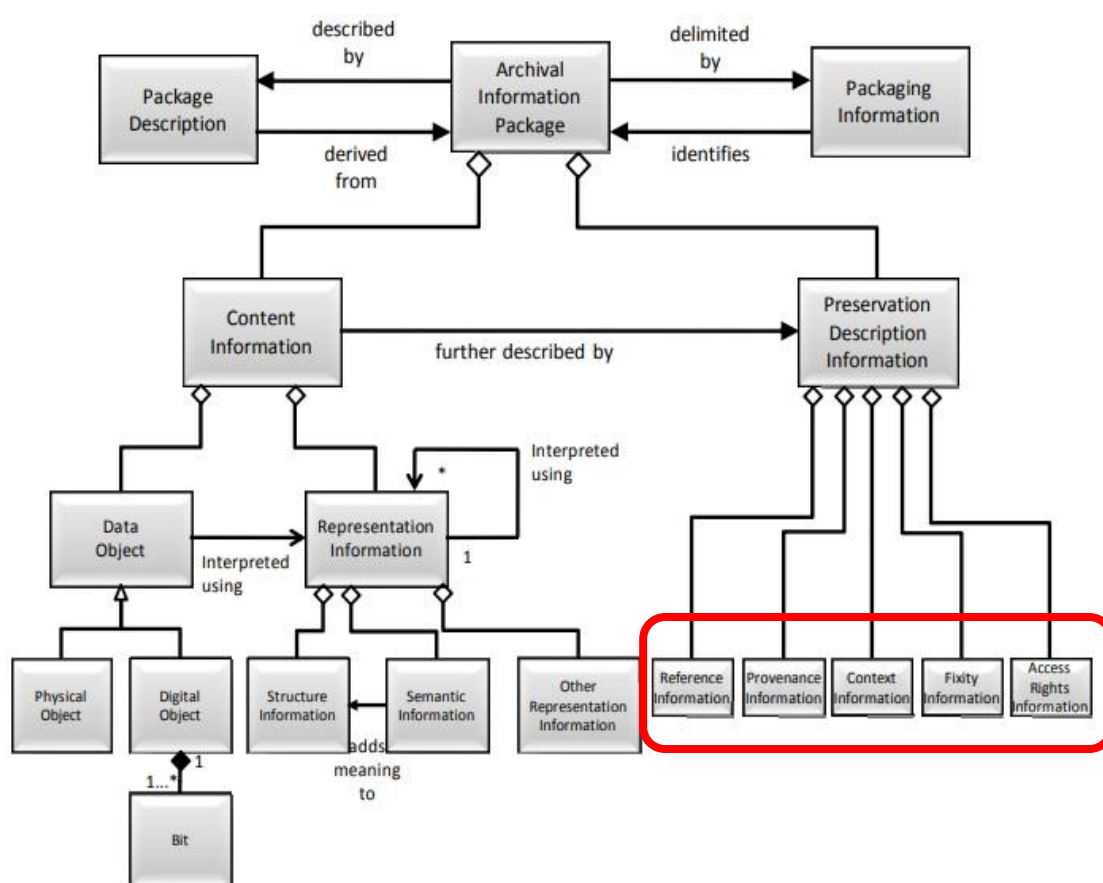
A Informação de fixidez pode incluir codificações especiais e esquemas de detecção de erros que são específicos para instâncias de Objetos de Dados de Conteúdo. A Informação de fixidez não inclui os mecanismos de preservação de integridade fornecidos pelos serviços subjacentes do OAIS, nem a proteção contra erros fornecida pela mídia e pelos drivers de dispositivo usados pelo Armazenamento Arquivístico. A Informação de fixidez pode especificar requisitos mínimos de qualidade de serviço para esses mecanismos (CCSDS 650.0-M-3, 2024, p.4-35-36, tradução nossa).

Por fim, a “Informação de direitos de acesso” identifica as restrições de acesso, incluindo a estrutura legal, termos de licenciamento e acesso ao controle.

A Informação de direitos de acesso [...] contém as condições de acesso e distribuição estabelecidas no Acordo de Submissão, relacionadas tanto à preservação (pelo OAIS) quanto ao uso final (pelo Consumidor). Também inclui as especificações para a aplicação de medidas de execução de direitos (CCSDS 650.0-M-3, 2024, p.4-36, tradução nossa).

As informações de PDI se correlacionam da seguinte maneira com os demais conceitos do OAIS.

Figura 15 - Visão detalhada das PDI na estrutura de um AIP



Fonte: CCSDS 650.0-M-3 (2024, p. 40).

A integralidade do registro das informações da PDI é opcional para os pacotes SIP e DIP, mas obrigatória para os Pacotes AIP. Retoma-se, no entanto, que a análise da

acurácia³⁸ e completeza³⁹ dos metadados de uma PDI parte da identificação de registros contínuos e cumulativos durante todo o ciclo de vida do objeto digital em preservação.

Por essa razão, embora não seja obrigatório, considera-se relevante o rol de informações da PDI na construção de um SIP que será submetido a um repositório. Isso porque, ao considerar a escolha por (re)usar um objeto digital que, conforme já abordado nesta tese, envolve a perspectiva subjetiva do usuário, a capacidade de demonstrar essas informações incrementa parâmetros pragmáticos à confiabilidade dos dados de pesquisa compartilhados e disponibilizados.

Dessa forma, reforça-se que a presunção de autenticidade dos dados de pesquisa é um fator que fortalece sua confiabilidade. Assim, a incorporação de informações que amparam a devida contextualização e proveniência desses dados deve ser efetivada. Ademais, informações que viabilizem a análise sobre sua capacidade em transmitir a mensagem que levou à sua produção de maneira a atingir seus objetivos (integridade) e em ser caracterizado como único (identidade), aspectos que integram a sua presunção de autenticidade, segundo CONARQ (2012), também devem ser registradas.

4.1.1 A relevância dos metadados na preservação digital qualitativa

Como tem sido demonstrado nesta seção 4, as estratégias e ações de preservação devem ser abrangentes o suficiente para abarcar todo o ciclo de vida do objeto digital. Para isso, deve-se levar em consideração aspectos imbricados tanto na sua criação quanto na fixação de sua integridade e manutenção ao longo do tempo.

Dentre as estratégias elencadas na literatura, é possível apontar ações de preservação que devem ser executadas na criação, na integridade e na manutenção, conforme quadro abaixo:

Quadro 4 – Estratégias de preservação digital

³⁸ Grau de precisão, correção, verdade e ausência de erros e distorções existentes (InterPares Project, disponível em: [ip2_creator_guidelines_booklet--portuguese.pdf](https://www.interpares.org/pt-br/guidelines-booklet/). Acesso em: 20 out. 2025).

³⁹ Segundo o Glossário da Câmara Técnica de Documentos Eletrônicos é “atributo de um documento arquivístico que se refere à presença de todos os elementos intrínsecos e extrínsecos exigidos pela organização produtora e pelo sistema jurídico administrativo a que pertence, de maneira a ser capaz de gerar consequências (CONARQ, 2020, p. 17).

Na criação	Na integridade	Na manutenção
Especificações técnicas claras e completas	Documentação de todas as políticas, estratégias e procedimentos	Robusta infraestrutura de computação e rede
Produção de arquivos mestres confiáveis	Uso de identificadores persistentes	Armazenamento e sincronização de arquivos em vários sites
Metadados descritivos, administrativos e estruturais suficientes para garantir acesso futuro	Origem registrada e histórico de mudança para todos os objetos	Monitoramento contínuo e gerenciamento de arquivos
Controle detalhado de qualidade dos processos	Mecanismos de verificação	Programas para refrescamento, migração e emulação
	Atenção aos requisitos de segurança	Criação e teste de planos de prevenção e recuperação de desastres
	Auditorias de rotina	Revisão periódica e atualização de políticas e procedimentos

Fonte: Adaptado pela autora com base no site da American Library Association (ALA).

Com base no Quadro 4, é possível observar que a maioria das estratégias de preservação digital é intrínseca e diretamente vinculada ao uso de metadados. Seja no registro de informações relevantes sobre o objeto digital em si, visando sua recuperação, acesso e renderização, seja sobre os processos aos quais ele é submetido ao longo do tempo, os agentes que executam esses processos e o ambiente em que é criado e tramitado.

De acordo com o site *Digital Preservation Coalition* (DPC Online), existem inúmeros fatores que tornam os metadados e a documentação (manuais de *software*, *designs* de pesquisa e guias do usuário, fornecidas por um criador e o repositório que complementa os metadados e fornece informações suficientes para permitir o uso do recurso por outras pessoas) críticos para a continuidade dos materiais digitais. São eles:

- **Tecnologia:** para serem inteligíveis, os recursos digitais dependem de hardware e *software* e seus requisitos técnicos precisam estar registrados para que sejam tomadas decisões apropriadas sobre possíveis estratégias de preservação e acesso;
- **Mudança:** os materiais digitais estão sujeitos a diversas ações e agentes (diferentes operadores e, possivelmente, instituições) que influenciam seu gerenciamento por um longo período de tempo, portanto, registrar essas

ações que são executadas no recurso e as alterações que ocorrem como resultado fornecem uma chave para seus futuros gerentes e usuários;

- **Autenticidade:** Os metadados e a documentação são os principais, e podem ser os únicos, meios confiáveis de estabelecer a autenticidade do material após as diversas alterações que passam ao longo do tempo;
- **Gerenciamento de direitos:** os recursos digitais precisam ser copiados para permanecerem acessíveis e, portanto, seus gestores necessitam saber que têm o direito de copiar para fins de preservação e, caso se apliquem, quais foram as tecnologias usadas para controlar o gerenciamento de direitos e quais implicações existentes para controlar esse acesso;
- **Reutilização futura:** sem a documentação adequada, pode não ser possível que outras pessoas usem o material; e
- **Custo:** criar metadados manualmente é caro e os metadados de preservação nem sempre podem ser facilmente gerados automaticamente, o que implica que metadados adicionais para necessidades de preservação digital exijam compensações cuidadosas de custo/benefício.

Nesse sentido, reforça-se que

Metadados são dados estruturados que descrevem objetos de informação e são um veículo chave para acessá-los, gerenciá-los e entendê-los, por isso **as propriedades que são por eles descritas devem ser cuidadosamente escolhidas para que sejam mais úteis para as tarefas que eles precisam suportar** (Dappert; Guenther; Peyrard, 2018, p.1, tradução nossa, grifo nosso).

Pode-se somar à importância dessa escolha, o fato de que os metadados são o caminho que ampara a possibilidade de (re)uso dos objetos digitais. Aliás, “Dados têm valor apenas quando são realmente usados ou podem ser úteis no futuro” (DMBOK, 2017, tradução nossa).

Conforme expressado pelo Diretor Executivo da NISO, Todd Carpenter (2017, tradução nossa), “em um mundo digital, as informações sobre conteúdo podem muitas vezes ser mais importantes do que o conteúdo em si. Sem bons metadados, a informação efetivamente desaparece”. Para isso, segundo Saramago (2004), os metadados devem conter informações de cunho técnico e administrativo relativas às decisões e ações de preservação. Além disso, segundo a autora, devem registrar os efeitos das estratégias adotadas para a conversão de dados, visando a manutenção da autenticidade dos recursos

digitais e as informações inerentes à gestão das coleções e aquelas acerca dos próprios metadados.

Em suma, os metadados desempenham um papel fundamental na documentação do histórico dos processos realizados ao longo do ciclo de vida dos objetos digitais. Isso para que, mesmo diante de migrações e atualizações, a confiança desses objetos não seja prejudicada.

É interessante observar que isso não é uma questão recente. No final da década de 90, Bullock (1999) já apontava alguns problemas relativos à preservação digital e seu status àquela época. O autor elenca nove aspectos sob os quais era preciso estar atento e cuja garantia de manutenção a longo prazo se fazia relevante, são eles: fixar o objeto como um todo, preservar a presença fixa (no caso, a sequência de 1 e 0), preservar o conteúdo, preservar a apresentação, preservar a funcionalidade, preservar a autenticidade, preservar a proveniência, preservar o contexto e a localização e possibilidade de consulta.

Esse entendimento de Bullock (1999) sobre a preservação já reforçava que não basta preservar o objeto em si: o dado de pesquisa, é preciso que informações relativas a ele, as quais sejam capazes de manter sua interpretação, compreensão e atestar sua confiabilidade, também sejam preservadas, tendo inclusive o mesmo nível de importância.

Reforça-se, assim, que a criação e o registro de metadados é uma ação relevante em todos os esforços reais que devem ser exercidos na práxis da preservação digital. Isso se dá pelo fato de que elas estão baseadas em métodos voltados à conservação de *softwares* e *hardwares*, mas também em reformatações que modificam os bits da cadeia binária que compõe o objeto digital original.

Podem ser dados como exemplo a conversão de formato de um arquivo a partir da alteração de codificação de texto, da compressão de imagens e/ou da reestruturação de tabelas. Nesses casos, há uma alteração do *bitstream* do objeto digital originalmente criado. Assim, é preciso haver o registro de metadados que demonstrem o controle da ação tomada e que o processo foi realizado sob requisitos planejados e responsabilidades conhecidas. Esse registro evidencia que o objeto não foi corrompido ou alterado inadvertidamente. Já a ausência dessas informações impacta a aferição da presunção de autenticidade desse objeto.

Reforça-se a noção de que metadados extensivos são os melhores recursos para minimizar os riscos de um objeto digital se tornar inacessível (Besser, 2010), uma vez

que “se revelam como um mecanismo capaz de registrar toda e qualquer ação realizada na produção e manutenção” (Praxedes; Rangel, 2017, p. 20)

Diante da complexidade da preservação no universo digital, compreende-se que “uma parte do problema só será resolvida com a identificação de um conjunto de dados e informações, expressos na forma de metadados, que ancorem os processos de gestão da preservação digital” (Sayão, 2010, p. 3). Dessa maneira, a concepção de pacotes de informação, com base no OAIS (2024), onde constem registros de metadados com o máximo de completeza e acurácia de informações desde o início do ciclo de vida dos dados de pesquisa, propicia a execução do Princípio R 1 de FAIR “(Meta)dados são descritos com uma pluralidade de atributos precisos e relevantes”.

É preciso compreender, porém, que no contexto da preservação digital os diferentes tipos de metadados: descritivos, administrativos (que englobam os técnicos, de preservação e de direitos), estruturais e de marcação de linguagem são importantes e nenhum deles deve ser desconsiderado enquanto um conjunto cumulativo de informações que se complementam.

Percebe-se que, no contexto da curadoria digital, é dada uma maior visibilidade aos metadados descritivos que possibilitam a interoperabilidade, a recuperação e o acesso. Seguido dos metadados técnicos, enquanto recursos capazes de viabilizar a manutenção da renderização dos dados, garantindo sua apresentação adequada. Todavia, é importante que também seja dada atenção aos metadados de preservação (que se consubstanciam a partir de informações registradas desde a criação/coleta do objeto digital).

Adicionalmente, é interessante que se reforce a noção de que, para possibilitar a confiança necessária a um (re)uso efetivo, os metadados de preservação devem registrar, de forma contínua, as ações e decisões que impactam o objeto digital. De acordo com o disponível no site do DPC, esses metadados tem por objetivo apoiar a manutenção da disponibilidade, identidade, persistência, capacidade de renderização, compreensibilidade e autenticidade dos objetos digitais por longos períodos de tempo.

Para tanto,

A preservação digital tem sido associada à proveniência retrospectiva: as definições que usamos em nossa área se baseiam fortemente em arquivos ao conceber a proveniência como a documentação das origens e mudanças na propriedade. O modelo de referência OAIS adiciona explicitamente a ideia de documentar as mudanças gerenciadas que os preservacionistas digitais fazem nos componentes físicos e lógicos dos alvos de preservação (Bettivia; Cheng; Gryk, 2022, p. 9, tradução nossa).

Julga-se relevante apontar que esses metadados de preservação não devem se limitar a registros de informações obtidas somente a partir da admissão do objeto digital em um repositório ou quando passa a ser tratado pela figura do curador. Há a necessidade premente de que todas as informações capazes de embasar a verificação de proveniência e contextualização sejam armazenadas associadamente. Até porque, de acordo com Esanu *et al.* (2004), nomes e conceitos podem mudar ao longo do tempo, e a preservação dos dados requer a preservação de contextos históricos.

Cabe refletir, então, sobre o fato de que os dados de pesquisa podem passar por diversas transmissões, seja de tempo ou espaço, ao longo de seu ciclo de vida, antes que sejam submetidos a um repositório de gestão de dados. Isso os leva a ficarem sujeitos a possíveis flancos metódico-gerenciais caso não haja registros adequados e cumulativos, ocasionando lacunas informacionais sobre esses trajetos precedentes. Esse fato impacta, diretamente, na presunção de sua autenticidade e, por consequência, podem vir a comprometer sua confiabilidade.

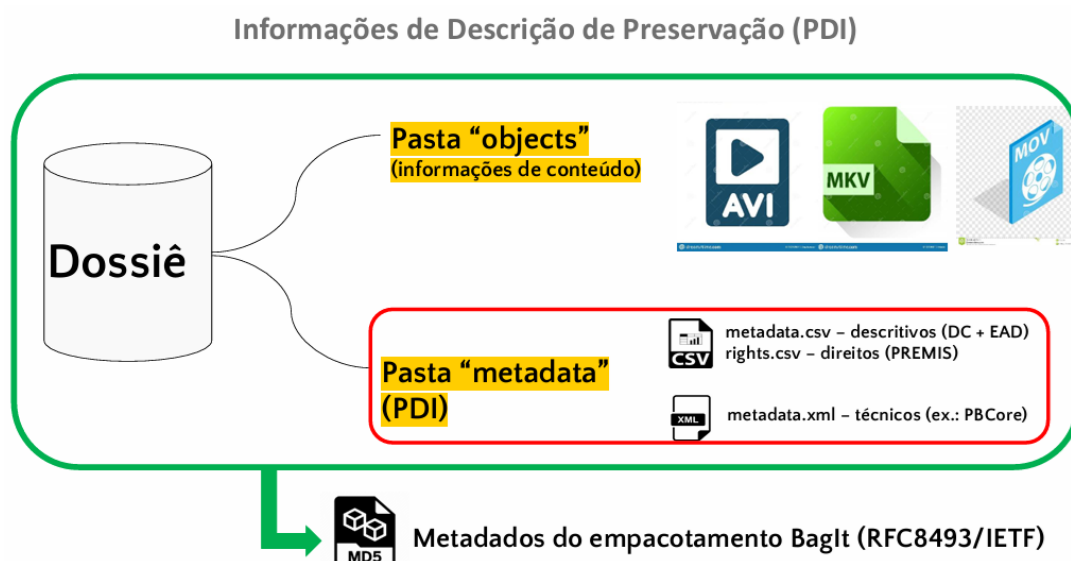
De acordo com Sayão (2010), a preservação digital possui cinco categorias fundamentais: proveniência, autenticidade, atividades de preservação, ambiente técnico e gestão de direitos. Para o autor,

- a proveniência afirma que os metadados registram informações do objeto desde sua origem, traçando sua cadeia de custódia e de propriedade.
- a autenticidade afirma a inclusão de informações suficientes que validem que o objeto não sofreu alterações, intencionais ou não, que não tenham sido documentadas;
- as atividades de preservação preconizam a documentação das ações tomadas ao longo do tempo para preservar o objeto digital e suas respectivas consequências ao longo do tempo;
- o ambiente técnico exige a descrição das dependências técnicas necessárias para a apresentação dos objetos digitais, ou seja, os hardwares, sistemas operacionais e *softwares* necessários para que eles possam ser lidos; e
- a gestão dos direitos registra tudo o que esteja submetido às questões de propriedade intelectual que possam vir a limitar ações de preservação, disseminação ou uso desses dados.

Para apresentar uma abordagem mais pragmática sobre os objetos digitais juntos aos seus metadados, o que o OAIS (2024) define como Informação de Conteúdo (IC) e suas Informações Descritivas de Preservação (PDI), respectivamente, traz-se um exemplo dado por Machado (2023)⁴⁰, conforme exposto abaixo na figura 16, que representa um *slide* apresentado pelo autor em uma aula proferida na Fundação Casa de Rui Barbosa (FCRB).

Na oportunidade, Machado (2023) demonstrou visualmente a prática aplicada para o empacotamento em SIP dos documentos audiovisuais da VideoSaúde Distribuidora (VSD), subordinada ao Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), da Fiocruz.

Figura 16 – Composição das Informações de Descrição de Preservação (PDI)



Fonte: Machado, 2023 (slide 71).⁴¹

A figura acima (16) representa um modelo visual do que está inserido na pasta *objects*, que contém os arquivos dos documentos em si, ou seja, as informações de conteúdo, e a pasta *metadata*, que constitui o que vai compor a PDI do pacote, onde se encontram os metadados descritivos, técnicos e de preservação. Após constituído esse

⁴⁰ Material enviado por e-mail à autora.

⁴¹ Esse slide consta da apresentação do curso “Planejamento para estruturação de Pacotes de Informações OAIS e o seu gerenciamento através do software de preservação para repositórios digitais”. Esse conteúdo foi enviado por e-mail, disponibilizado diretamente pelo autor João Guilherme Nogueira Machado.

dossiê, o mesmo é empacotado em padrão BagIt⁴², seguindo o RFC8493/IETF (Força Tarefa de Engenharia da Internet) para geração de *checksum* dos objetos digitais que compõem o dossiê.

Compreende-se, assim, que o uso exclusivo do modelo de referência OAIS não garante a confiabilidade no espectro integral. Dessa forma, faz-se necessário que outros instrumentos também sejam considerados para complementá-lo na prática efetiva da preservação digital.

4.1.2 Referenciais internacionais complementares ao OAIS

Nesse sentido, apresenta-se a Prática Recomendada da CCSDS 652.0-M-2 intitulada como “*Audit and certification of trustworthy digital repositories*” que deu origem à Norma ISO 16.363. Essa Norma ISO intitulada “*Space data and information transfer systems — Audit and certification of trustworthy digital repositories*” foi recentemente atualizada, em 2025, quando a Prática Recomendada da CCSDS passou por uma revisão.

A ISO 16.363:2025 estabelece um conjunto abrangente de requisitos que visam garantir a confiabilidade dos repositórios digitais na sua função de preservar informações digitais para o futuro. Assim, ao contrário de focar em formatos ou tecnologias específicas, ela define atributos que caracterizam um repositório digno de confiança. Isso permite que produtores de dados, usuários e financiadores tenham segurança na longevidade e acessibilidade dos acervos digitais.

De acordo com Corujo e Revez (2021, p. 249), ela “destina-se principalmente aos responsáveis pela auditoria de repositórios digitais e para aqueles que trabalham ou são responsáveis por repositórios digitais, que buscam uma medição objetiva da fiabilidade do seu repositório”.

Essa Norma (2025) aborda desde a estrutura organizacional e a sustentabilidade financeira da instituição responsável pelo repositório até a robustez da sua infraestrutura tecnológica. Além disso, indica a implementação de políticas claras para o gerenciamento

⁴²Segundo IBICT (2019, p.7), no documento “Estudo do BagIt”, o “BagIt é o formato padronizado de empacotamento utilizado pelo Archivematica, criado pela Biblioteca do Congresso Americano sob licença pública. Ele organiza de forma hierárquica os documentos digitais junto aos seus metadados”. Disponível em: [Estudo do Bagit – Hipátia](#). Acesso em: 10 out.2025.

dos objetos digitais ao longo do tempo. Dessa forma, a ISO 16.363 (2025) esclarece que um Repositório Digital Confiável (*Trusted Digital Repositories* - TDR) existe para preservar, de maneira compatível com o Modelo de Referência OAIS (2024), os recursos digitais confiados aos seus cuidados.

Cabe apontar que a ISO 16.363 (2025) se divide em três critérios: “infraestrutura organizacional”, “gerenciamento de objetos digitais” e “infraestrutura e gestão de riscos”. No entanto, para fins desta tese, opta-se por detalhar um desses pilares: o gerenciamento de objetos digitais (seção 4 da referida Norma).

O “gerenciamento dos objetos digitais”, é baseado no Modelo de referência OAIS que, conforme cita Santos e Flores (2020), normatiza a formação de pacotes de informação que podem ser admitidos no repositório. Esse tópico da Norma fornece requisitos sobre todo o processo de preservação, desde o recebimento e ingestão do SIP, a criação e preservação do AIP até o fornecimento de acesso do usuário aos recursos digitais preservados.

Isso inclui procedimentos para a ingestão segura dos dados, estratégias de armazenamento que garantam a sua integridade, ações de preservação para combater a obsolescência de formatos e mecanismos de acesso e disseminação que atendam às necessidades da comunidade designada.

Assim como vem sendo apontado ao longo da seção primária 4 desta tese, essa Norma (2025) também enfatiza a importância dos metadados que descrevem o histórico, o contexto e as características técnicas dos objetos digitais, reforçando que eles são essenciais para a sua compreensão e (re)uso futuro.

Por apresentar os requisitos necessários para que repositórios digitais sejam considerados confiáveis, a ISO 16.363 (2025) aponta que esses repositórios precisam gerenciar adequadamente as informações necessárias para garantia da autenticidade, integridade, acesso e usabilidade dos objetos digitais ao longo do tempo.

Para isso, aponta que um repositório confiável tem que ser capaz de:

- documentar a proveniência dos objetos digitais (quem os criou, histórico de alterações, cadeia de custódia),
- implementar mecanismos para garantir a integridade dos objetos digitais (fixidez),
- manter informações sobre o contexto dos objetos digitais (relações com outros objetos, ambiente de criação),

- gerenciar e documentar os direitos de propriedade intelectual e as condições de acesso e uso; e
- fornecer informações de representação que garantam a renderização e compreensão dos objetos digitais ao longo do tempo (formato, *software* necessário, etc).

Dessa forma,

[...] um repositório digital confiável, além de manter autênticos os objetos digitais, de preservá-los e prover acesso a eles pelo tempo necessário, contém algumas atribuições e responsabilidades específicas, baseados em normas e padrões de referências nacionais e internacionais (Flores; Pradebon; Cé, 2017, p. 77).

Considera-se, portanto, que o processo de auditoria e certificação de Repositórios Digitais Confiáveis, baseado nos critérios definidos por essa ISO visa avaliar a capacidade de repositórios manterem a autenticidade, a integridade e a acessibilidade dos seus objetos digitais a longo prazo. Essa certificação complementa as orientações de estruturação informacional do OAIS.

Ou seja, enquanto o OAIS (2024) traz questões conceituais para a preservação digital, a ISO 16.363:2025 apoia sua instrumentalização a partir do uso de repositórios que atendam ao recomendado pelo Modelo de Referência. O que os torna instrumentos indissociáveis para a efetivação da preservação digital.

Adicionalmente, pelo fato de os Princípios FAIR não abordarem questões sobre a confiabilidade dos dados de pesquisa, conforme demonstrado na subseção 3.1.3 desta tese, mesmo sendo essa uma qualidade fundamental tanto aos pesquisadores que compartilham dados quanto aos que pretendem (re)usá-los, considera-se relevante apontar a importância da integração de FAIR ao Modelo OAIS (2024).

Como exemplo dessa associação para a viabilização de possíveis (re)usos dos dados de pesquisa, toma-se um material de discussão elaborado no âmbito da *Consultative Committee for Space Data Systems* (CCSDS), em versão preliminar (draft, s.d.), que explora relações entre diferentes diretrizes fixadas em modelos, certificações, padrões e princípios utilizados de forma ampliada no cenário da preservação digital.

Para fins desta tese, no entanto, apresenta-se a seguir o extrato que demonstra as relações entre o “R” de FAIR e o Modelo OAIS e ISO 16.363.

Quadro 5 – Relações do FAIR com OAIS/ISO 16.363

FAIR	OAIS / ISO 16.363
R.1 (Meta)dados são descritos com uma pluralidade de atributos precisos e relevantes.	OAIS fornece requisitos mais detalhados para informações de representação e, em menor grau, para informações de proveniência, informações sobre fixidez, informações de referência e informações sobre direitos de acesso.
R.1.1 (Meta)dados devem ser disponibilizados com licenças de uso claras e acessíveis.	3.5.2 O repositório deve rastrear e gerenciar os direitos de propriedade intelectual e as restrições de uso do conteúdo do repositório conforme exigido pelo acordo de depósito, contrato ou licença. Para “metadados” também está implícito no OAIS e ISO 16363, onde os próprios “metadados” estão sendo preservados.
R.1.2 (Meta)dados devem estar associados à sua proveniência.	Dados precisam de Informação de Proveniência Para “metadados” também está implícito no OAIS e ISO 16363, onde os próprios “metadados” estão sendo preservados.
R.1.3 (Meta)dados devem estar alinhados com padrões relevantes do seu domínio.	Nenhuma restrição no OAIS/ISO 16363 sobre quais padrões usar. Em qualquer caso, os padrões mudarão com o tempo.

Fonte: Adaptado pela autora com base em material de discussão (draft, s.d.) da CCSDS.⁴³

O Quadro 5⁴⁴ demonstra como os critérios de preservação digital podem contribuir no (re)uso dos dados de pesquisa ao serem adotadas de maneira integrada aos Princípios FAIR.

Em complementação ao quadro acima, segundo disposto no site do Force11 (2021), para ser reusável, a descrição dos elementos de metadados essenciais, recomendados e opcionais deve ser processável por máquina e verificável. Além disso, reforça que seu uso deve ser fácil e os dados devem ser citáveis para sustentar o compartilhamento e o reconhecimento de seu valor.

Cabe frisar que o Princípio R de FAIR especifica ainda mais a essência dos outros princípios e traz a importância do detalhamento das características dos dados, inclusive sobre sua proveniência. Isso implica em “fornecer e publicar descrições de dados precisas e relevantes, licenças de acesso e uso, os padrões da comunidade que foram empregados

⁴³ Disponível em: [Relationship between OAIS, ISO 16363, CTS, FAIR and TRUST.docx \(live.com\)](https://www.force11.org/Relationship-between-OAIS-ISO-16363-CTS-FAIR-and-TRUST.docx). Acesso em: 20 ago. 2025.

⁴⁴ Cabe salientar que esse quadro tem caráter ilustrativo, derivado de material preliminar (*draft*) da CCSDS, não se tratando de um documento normativo formal, mas de um exercício de aproximação conceitual entre FAIR, OAIS e ISO 16363.

no processo, bem como a proveniência associada para cada conjunto de dados” (Boeckhout *et al.*, 2018, p. 932).

Traz-se, portanto, a compreensão de que essas descrições detalhadas devem ser arquivadas enquanto metadados e junto aos dados em uma ou mais categorias de informações que comporão a PDI do AIP a partir do que foi registrado enquanto PDI do SIP, cabe frisar.

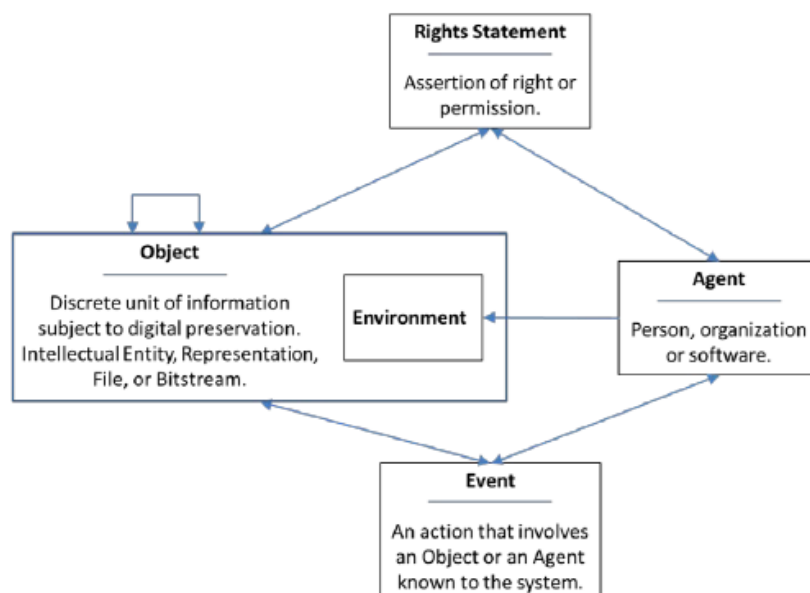
Tanto o OAIS quanto o FAIR são estruturas de alto nível que permitem que diversos domínios e comunidades implementem apropriadamente a preservação e o apoio necessário para o (re)uso dos dados em ambiente digital.

No entanto, em busca de lidar com aspectos práticos imbricados da adoção dos modelos internacionais inerentes à preservação digital, é preciso analisar o *PREMIS Data Dictionary*, produto do grupo de trabalho PREMIS (*Preservation Metadata: Implementation Strategies*), composto por especialistas internacionais e patrocinado pela OCLC e RLG desde junho de 2003.

Sua finalidade é definir metadados que suportem a viabilidade, renderização, compreensibilidade, autenticidade e identidade de objetos digitais em um contexto de preservação, representando informações que a maioria dos repositórios de preservação provavelmente precisam saber para amparar a preservação digital desses objetos a longo prazo (Premis Editorial Committee, 2015).

O Modelo de Dados PREMIS define unidades semânticas divididas em quatro entidades: Objeto Digital, Evento, Agente e Direito.

Figura 17 – Modelo de dados PREMIS



Fonte: PREMIS Data Dictionary (2015, p. 6).

De acordo com a Figura 17, a entidade Objetos⁴⁵ (*Objects*) é a unidade de informação sujeita à preservação digital, podendo ser composta por Entidade Intelectual, Representação, Arquivo ou *Bitstream*. Sendo cada uma delas, segundo PREMIS (2015):

- a) Entidade intelectual (*Intellectual entity*): Criação intelectual ou artística distinta, relevante para uma comunidade designada. Pode incluir outras Entidades Intelectuais e ter uma ou mais Representações.
- b) Representação (*Representation*): Conjunto de arquivos, incluindo metadados estruturais, necessários para uma renderização completa de uma Entidade Intelectual. Pode ser digital ou física.
- c) Arquivo (*File*): Sequência nomeada e ordenada de bytes conhecida por um sistema operacional.
- d) *Bitstream*: Dados contíguos ou não contíguos dentro de um arquivo que possuem propriedades comuns significativas para fins de preservação. Não pode ser transformado em um arquivo autônomo sem adição de estrutura de arquivo. Além disso, tem-se o *Filestream*, um *Bitstream* que contém todas as propriedades do arquivo (*file*).

⁴⁵ A versão 3.0 introduz o conceito de que um ambiente pode ser um Objeto.

Nessa versão, o Ambiente (*Environment*), ambiente técnico (*software, hardware, dependências*) é modelado como Objetos próprios e ligados a outros objetos por relações de dependência. Já os Eventos (*Events*) registram ações que envolvem ou afetam pelo menos um Objeto ou Agente conhecido pelo repositório de preservação.

Cabe reforçar que o registro das ações que modificam um Objeto é crítico para manter uma peça-chave para a autenticidade⁴⁶ e da proveniência digital. Essa proveniência é entendida como documentação de processos no ciclo de vida de um Objeto Digital, descrevendo Agentes, Objetos Digitais, Eventos e outras informações associadas à criação, gestão e preservação do Objeto Digital.

Os Agentes (*Agents*) são pessoa, organização ou programa/sistema de *software* associado a Eventos na vida de um Objeto, ou a Direitos anexados a um Objeto. Sendo os Direitos (*Rights*) as declarações de um ou mais direitos ou permissões referentes a um Objeto e/ou Agente. Seu objetivo principal é permitir que um repositório de preservação determine se tem o direito de realizar uma ação em modo automatizado.

Nesse dicionário de dados (2015), são apresentados e discutidos a importância dos conceitos de Fixidez, Integridade, Autenticidade. A Fixidez é verificada pela comparação de um *messageDigest (hash)* calculado em diferentes momentos. A Integridade se constitui pela verificação de que um objeto está de acordo com sua especificação, como validação de formato. Já a Autenticidade é reforçada enquanto uma qualidade de verificar que o objeto é o que ele afirma ser, dependendo, portanto, de evidências técnicas e processuais, como proveniência e assinaturas digitais.

Sendo assim, tanto o OAIS quanto o PREMIS *Data Dictionary* são padrões amplamente aceitos na comunidade que executa a preservação digital e ambos apontam os registros de informações sobre proveniência como recursos primários e relevantes na preservação (Li; Sugimoto, 2014). Enquanto o primeiro estabelece os conceitos fundamentais para descrever e organizar informações digitais para preservação, o segundo, que teve o OAIS como ponto de partida, usa esses conceitos para definir como os metadados de preservação devem ser estruturados, fornecendo um guia prático para sua implementação (www.loc.gov).

A ISO 16.363:2025, o PREMIS *Data Dictionary* e os princípios FAIR, em conexão ao OAIS (2024), apontam os registros de informações sobre proveniência como

⁴⁶ Nesse sentido, aplica-se o que se entende como autenticidade diplomática (Lacombe; Rondinelli, 2016), conforme exposto na subseção 2.3.3 desta tese.

metadados relevantes. No entanto, o foco desses instrumentos está predominantemente nas ações que ocorrem a partir da intenção de que esses objetos digitais sejam ingeridos em um repositório de preservação digital.

Considera-se, portanto, que eles não forneçam metodologia para garantir que as informações de proveniência, capazes de caracterizar os dados de pesquisa como evidência, tenham sido registradas nas etapas precedentes.

4.2 PREPARAÇÃO DOS DADOS DE PESQUISA: PAVIMENTANDO O CAMINHO PARA A PRESERVAÇÃO DIGITAL QUALITATIVA

Assim como foi demonstrado na subseção 4.1.1 desta tese, é necessário que ao longo do ciclo de vida dos dados sejam registradas as chamadas “Informações Adicionais”⁴⁷. Essas informações “devem acompanhar os dados para garantir que eles possam ser preservados e explorados, o que inclui as Informações de Representação e a PDI” (CCSDS 650.0-M-3, 2024, p. 7, tradução nossa).

Diante disso, o OAIS (2024) discorre sobre alguns desafios na coleta e registro dessas informações. São elas:

[...] o foco no hardware por aqueles envolvidos em estágios iniciais de um projeto, o que significa que eles podem nem sempre pensar em coletar e salvar informações sobre decisões de design e calibrações necessárias para a análise dos dados que o hardware coletará ou criará; **falta de compreensão de que deve haver uma alocação de orçamento para financiar a coleta dessas informações; incerteza sobre quais informações coletar** em vários estágios, o que muitas vezes significa que muito pouco é coletado; e **limites na coleta de dados de tal forma que as informações podem não ser coletadas** se não forem necessárias para o uso principal dos dados coletados, o que significa que usos alternativos são limitados (CCSDS 650.0-M-3, 2024, p. 1, tradução nossa, grifo nosso).

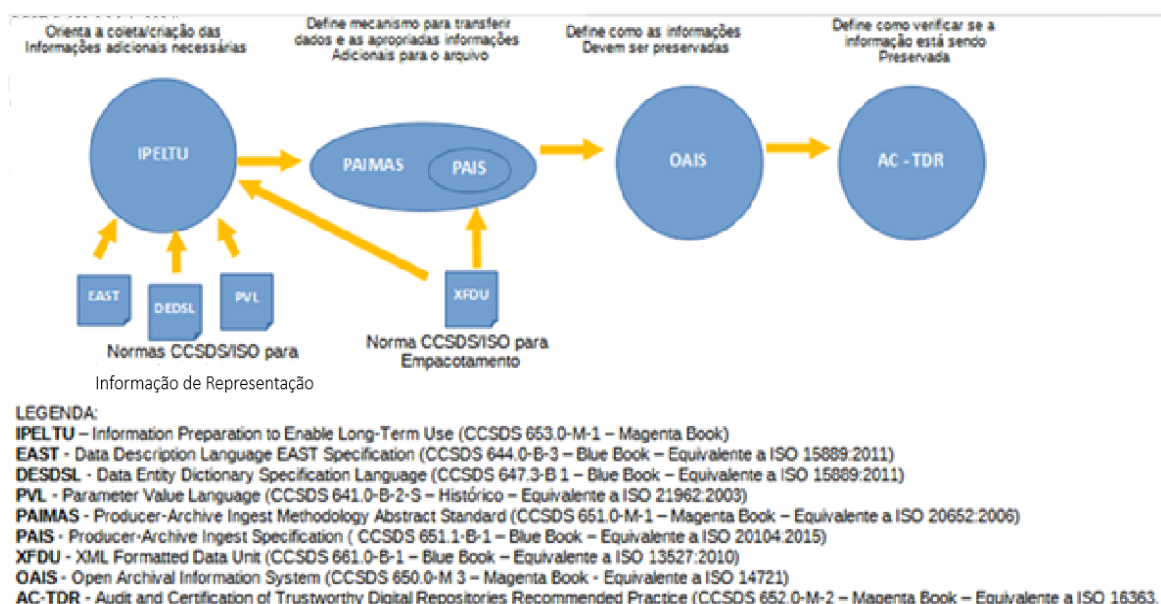
Observa-se que um dos problemas elencados pela CCSDS (2024) é a incerteza sobre quais informações devem ser coletadas ao longo do ciclo de vida dos objetos

⁴⁷ Em 2024, a CCSDS passou a utilizar o termo “informações adicionais” no lugar de “metadados”, em virtude de considerar que o uso de metadados é “inespecífico e pode causar confusão”. (CCSDS 650.0-M-1, 2024, p. 1-1, tradução nossa). No entanto, para fins desta tese, opta-se pela manutenção da utilização do termo metadados por ser um termo mais amplamente adotado, ao menos por enquanto, nos dias de hoje, no cenário da curadoria digital.

digitais, o que pode vir a implicar em uma baixa coleta que impacta na sua qualidade e, consequentemente, de sua preservação digital.

Em 2024, Comitê Consultivo da CCSDS clarificou a necessidade de interligação entre o OAIS a outras recomendações que se prestam a dar conta das fases precedentes ao seu uso. Essa costura integra orientações relativas ao que seriam todas as etapas do fluxo informacional no ciclo de vida dos dados de pesquisa, conforme é demonstrado a seguir.

Figura 18 - Relações entre os Padrões CCSDS



Fonte: Adaptado pela autora com base em CCSDS 653.0-M-1 (2024).

Para explicar a Figura 18, opta-se, propositalmente, por analisar os padrões da direita para a esquerda. Essa escolha marca a trajetória inversa do fluxo do ciclo de vida dos dados de pesquisa. No entanto, contribui no reforço da compreensão de que a preservação digital em si é fruto de um conjunto de tratamentos técnicos que foram realizados em estágios pretéritos.

Assim, tem-se os seguintes instrumentos:

- AC-TDR “fornece métricas para avaliação da confiabilidade de repositórios ou arquivos digitais” (apresentada na subseção 3.2.3 desta tese) (CCSDS 653.0-M-1, 2024, p. 1-3, tradução nossa);
- OAIS orienta como preservar informações e torná-las disponíveis para uma Comunidade Designada, a partir da organização de pessoas e sistemas

que aceitaram essa responsabilidade de preservar informações (CCSDS 653.0-M-1, 2024);

- PAIMAS “define uma metodologia para transferência de dados de um Produtor de informação para um Arquivo com base nas quatro fases seguinte: preliminar, definição formal, transferência e validação” (CCSDS 653.0-M-1, 2024, p. 1-3, tradução nossa);
- PAIS “fornece a sintaxe abstrata e uma implementação XML de descrições desses dados que serão transferidos” (CCSDS 653.0-M-1, 2024, p. 1-3, tradução nossa). Consiste, portanto, num Acordo entre as partes envolvidas (Produtor – Arquivo), fazendo uma ponte entre os dois ambientes, utilizando a Unidade de Dados Formatada em XML (*XML Formatted Data Unit* - XFDU) como padrão para a criação dos pacotes de informação. “Essa negociação é necessária para garantir que o nível de correto de detalhes de descrição seja produzido” (CCSDS 653.0-M-1, 2024, p. 1-3, tradução nossa);
- IPELTU promove orientação voltada aos estágios iniciais da pesquisa, para conscientizar os produtores de dados sobre a importância de coletar ou criar as informações adicionais que fornecem qualidades suficientes para que seus dados sejam usados agora e no futuro. Além disso, essas informações consubstanciam subsídios informacionais necessários para amparar sua preservação. A IPELTU indica a possibilidade de serem utilizados padrões como: Linguagem de Valor de Parâmetro (*Parameter Value Language* – PVL), Especificação de Dicionário de Entidade de Dados (*Data Entity Dictionary Specification Language* – DEDSL) e Linguagem de Descrição de Dados EAST (*Data Description Language EAST Specification* – EAST) para criar as Informações de Representação.

Importa explicar que todas essas recomendações visam dirimir impactos de um contexto tecnológico que deve ser identificado como um desafio contínuo. Isso porque o prazo de obsolescência dos recursos digitais, como *hardwares*, *softwares*, formatos de arquivos digitais e sistemas de armazenamento, se inicia, em média, a partir de 5 anos.

Conforme expõem Rocha e Silva (2007)

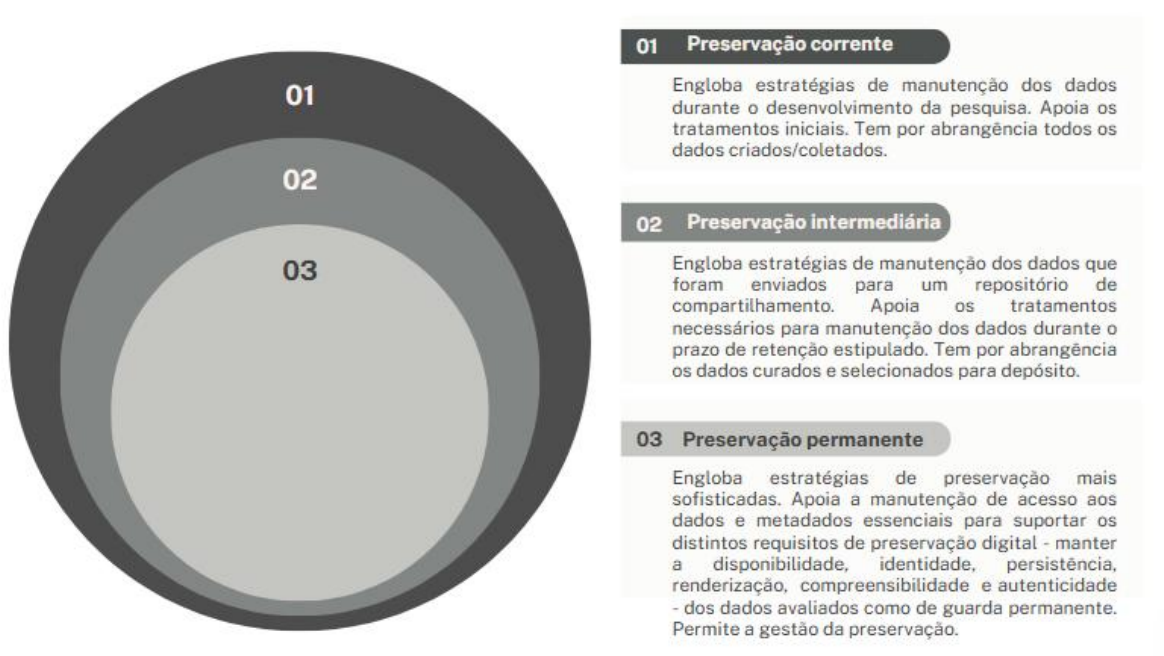
Além da obsolescência tecnológica, outro problema é a fragilidade do suporte digital que em poucos anos pode se danificar e impossibilitar a compreensão do documento. [...] no caso dos documentos digitais, um só bit danificado impossibilita a leitura de todo o documento (Rocha; Silva, 2007, p. 115-116).

Isso indica que a necessidade em preservar os recursos digitais se manifesta bem antes do momento em que estes são submetidos a um Repositório de Preservação Digital. De uma forma simplificada, então, considera-se que a preservação digital deve ser realizada de forma sistêmica e continuada pelas diversas etapas ao qual o objeto digital venha a percorrer. O que se aplica, também, aos dados de pesquisa.

Dessa maneira, de acordo com Silva (2019), a preservação digital se integra de forma natural e intrínseca à própria gestão de dados de pesquisa. Essa afirmativa sustenta a ideia de que a preservação não seja mais um procedimento restrito aos estágios finais do ciclo de vida.

Com base nisso, apresenta-se para esta tese o que esta pesquisadora chama de “camadas de preservação”. Essas camadas são definidas a partir de seu entendimento arquivístico que compreende que as demandas inerentes à preservação ao longo do ciclo de vida impõem à gestão de dados a necessidade de gerir, de forma detalhada e cumulativa, registros relativos a todas as ações preservacionistas implementadas em cada etapa.

Figura 19 - Camadas de preservação ao longo do ciclo de vida dos dados de pesquisa



Fonte: Elaborada pela autora (2025).

Ao observar a Figura 19, é relevante clarificar que essas camadas se tornam mais enxutas em termos de quantitativo de dados de pesquisa, englobados em cada uma delas. O que é inversamente proporcional ao quantitativo de metadados que devem ser registrados junto aos objetos digitais, ao perpassarem essas camadas. Isso porque o registro de informações significativas deve se dar de forma contínua e cumulativa ao longo do ciclo de vida dos dados de pesquisa.

A primeira camada, chamada de “Preservação corrente”, é aquela executada em ambiente sob ingerência do pesquisador até que seus dados sejam depositados em um repositório de gestão de dados de pesquisa, tratado na subseção 3.2.3 desta tese, que na figura 19 corresponde à segunda camada.

Na segunda camada, chamada “Preservação intermediária”, esses dados vão ser geridos e compartilhados. Ao passarem por um critério de avaliação para a manutenção de acesso a esses dados, os que forem considerados com valor para serem guardados a longo prazo ou permanentemente, serão submetidos à terceira camada.

Essa terceira camada, chamada de “Preservação permanente”, correspondente ao uso do repositório digital de preservação, os dados vão ser submetidos a um repositório de preservação digital, baseado no Modelo de Referência OAIS e em atendimento aos critérios da ISO 16.363:2025.

Aponta-se que progressão dos dados de pesquisa ao longo de seu ciclo de vida implica a passagem por essas camadas, onde cada qual possui seus requisitos e ações específicas de preservação digital.

Especificamente, a Preservação Corrente engloba estratégias de manutenção dos dados durante o desenvolvimento da pesquisa, apoiando os tratamentos iniciais e tendo por abrangência todos os dados criados/coletados.

A segunda camada, Preservação intermediária, engloba estratégias de manutenção dos dados que foram enviados para um repositório de gestão de dados de pesquisa, apoiando os tratamentos necessários para manutenção desses dados durante o prazo de retenção estipulado e tendo por abrangência os dados curados e selecionados para depósito.

A terceira camada, Preservação permanente, engloba estratégias de preservação mais sofisticadas, apoiando a manutenção de acesso aos dados e metadados essenciais para suportar os distintos requisitos de preservação digital (manter a disponibilidade,

identidade, persistência, renderização, compreensibilidade e autenticidade dos dados avaliados como de guarda permanente) e permitindo a gestão da preservação digital.

Observa-se, portanto, um aumento na complexidade e formalização das ações de preservação à medida que se avança da primeira para a terceira camada, mas cada uma delas possui papel fundamental na construção de uma Cadeia de Preservação, conforme descrito na subseção 2.3.2 desta tese. Isso reflete a importância da fase de gestão de dados para garantir os registros informacionais relativos aos objetos, eventos e agentes envolvidos nesses processos e a sua cumulatividade, persistência e transparência ao longo do tempo.

Quanto a isso, julga-se relevante apresentar em maior detalhamento o documento técnico de Práticas Recomendadas “Preparação de informações para permitir o uso a longo prazo (IPELTU)” (Magenta Book CCSDS 653.0-M-1), publicado em dezembro de 2024, pelo Comitê Consultivo do *Consultative Committee for Space Data Systems* (CCSDS). Documento este que se encontra no início do fluxo exposto anteriormente na figura 18 “Relações entre os Padrões CCSDS”.

Esse guia, com participação de agências de diversos países, incluindo o Instituto Nacional de Pesquisas Espaciais (INPE) e o Departamento de Ciência e Tecnologia Aeroespacial (DCTA), ambos brasileiros, visa “orientar projetos sobre a captura e retenção de “informações adicionais” (anteriormente referidas como metadados) para garantir a exploração dos dados de pesquisa a curto, médio e longo prazos” (CCSDS 653.0-M-1, 2024, p. 1, tradução nossa).

Conforme a IPELTU, dentre as informações que compõem as PDI, as de proveniência, provavelmente, são necessárias durante todo o ciclo de vida da pesquisa e além (tendo relevância para todos os seus resultados subsequentes). Esse fato justifica que sejam preservadas durante todo o desenvolvimento da pesquisa e após, devido a sua importância como evidência para a presunção da autenticidade e seu valor para reprodutibilidade.

Segundo essa Recomendação prática (2024), as demais informações das PDI (referência, fixidez, contexto e direitos de acesso) podem ser necessárias também nos estágios posteriores se forem relevantes para reforçar as informações de proveniência e se estiverem disponíveis.

O arquivo, que na IPELTU (2024) se refere ao que seria o arquivo permanente, deve criar o AIP, com todas as informações necessárias para usabilidade de longo prazo,

como parte do processo de preservação. Além disso, reforça, as informações “devem ser coletadas de forma oportuna ao longo do projeto de pesquisa, até porque muitos dos componentes que formam esse AIP podem ser conhecidos apenas pelos participantes do projeto” (CCSDS, 2024, p. 4-1, tradução nossa).

Assim sendo,

Dados coletados ou criados devem ter Informações Adicionais associadas a eles para que sejam **independentemente compreensíveis, utilizáveis e confiáveis como autênticos. A quantidade e o conteúdo de Informações Adicionais mudam ao longo do tempo**, conforme o hardware, o software, o ambiente geral e o conhecimento tácito dos usuários mudam (CCSDS 653.0-M-1, 2024, p. 1-4, tradução nossa, grifo nosso).

A IPELTU (2024) coloca, inclusive, que os Planos de Gestão de Dados (PGD) devem ser “elaborados, atualizados e monitorados” para contribuir no “aumento da eficácia das atividades de preservação e a exploração da informação” e “podem ser capturados como uma parte das Informações Adicionais” (CCSDS 653.0-M-1, 2024, p. 1-4, tradução nossa).

Tal afirmação, reforça a importância de que os registros que dão embasamento à construção das informações de proveniência são imprescindíveis para viabilização de (re)uso dos dados de pesquisa. Nessas informações devem estar incluídos registros sobre as origens e o histórico, como por exemplo: documentação específica da pesquisa⁴⁸, onde foi derivado (ou seja, dados coletados anteriormente), *software* de processamento, dados relacionados, custódia de dados, controle de versões, calibração e testes, processamento de hardware, *software* e logs, como a qualidade da informação pode ser verificada, gestão da migração, gestão de cópia dos dados, política de sincronização de cópias, defesa contra *hackers*, quais verificações de antivírus foram realizadas, funções de pessoas (que podem alterar/excluir) (CCSDS 653.0-M-1, 2024).

Em um exemplo apresentado pela IPELTU (2024), no qual são apresentadas as fases da pesquisa e os grupos de coleta (divididos em início, planejamento, execução e fechamento) de um grande projeto de pesquisa, é exposto em uma nota o seguinte:

As Informações Adicionais enviadas ao Arquivo **incluem informações de Representação**, como o formato, semântica e software de processamento dos dados, que foram criados por vários membros do consórcio. **No entanto**,

⁴⁸ Optou-se por utilizar o termo pesquisa por ser o escopo desta tese. Cabe apontar que o uso faz analogia ao disposto no CCSDS 653.0-M-1 (2024) que aponta projetos e missões em virtude de ser destinado, embora explicitamente não se limite, a sistemas de dados espaciais.

muitos membros do consórcio migraram para outros projetos ou se aposentaram e algumas informações relevantes foram perdidas. Isso pode reduzir substancialmente o valor dos dados a longo prazo pois limita quem poderá reutilizá-los e pode até mesmo levar a se questionarem sobre a autenticidade dos dados (CCSDS 653.0-M-1, 2024, p. E-5, tradução nossa, grifo nosso).

Essa nota clarifica que lacunas informacionais sobre os dados de pesquisa diminuem sua possibilidade de (re)uso, inclusive pela falta de segurança quanto à sua autenticidade.

4.2.1 O registro da proveniência dos dados de pesquisa

O conceito de “*provenance knowledge*”, apresentado por Nesmith (1982), aponta que conhecer o contexto é muito mais importante do que o conhecer o assunto do documento, devendo haver um foco no porquê e em como as pessoas criam os documentos ao invés de se debruçarem exclusivamente em seu assunto.

Assim, compreende-se que o uso de metadados, mesmo que descritivos, cujo objetivo se volta à recuperação e acesso aos conjuntos de dados, possam expressar, de alguma forma, informações que espelhem a proveniência, mas não são suficientes.

Isso converge com o entendimento de Millar (2015) que, ao tratar sobre que chama de “respeito à proveniência”, identifica a abrangência três componentes principais: a história do criador, a história do documento e a história custodial.

Segundo a autora (2015), o primeiro deve dar ênfase a quem criou, acumulou e/ou usou o documento ao longo do tempo. O segundo, tem por intuito conhecer como o documento foi criado, mantido e/ou usado, por quem e por que, seus trâmites, sua destinação (recolhimento para guarda permanente ou eliminação). Já o terceiro, tende a focar na explicação de como esse documento chegou onde está e quais cuidados lhes foram garantidos durante esse processo.

Millar (2015) entende que as razões que levaram um documento a ser criado, quem o produziu, sob qual intenção, e todo seu ciclo vital são fatores essenciais para que ele possa ser contextualizado na sociedade onde foi produzido e que o utiliza.

Esse entendimento encontra direção no Princípio FAIR “R1.2: (Meta)dados são associados à proveniência detalhada”, que nesse contexto, conforme Kohwalter (2021), pode ser chamada também de “linhagem de dados”. Essa linhagem, segundo o autor visa responder as seguintes perguntas: Por que os dados foram produzidos? Como os dados

foram produzidos? Onde os dados foram produzidos? Quando os dados foram produzidos? Por quem os dados foram produzidos?

Assim, para fins desta tese, considera-se que informações que reflitam contextualização e fixidez, segundo o disposto no OAIS (2024), também encorpam informações relevantes sobre proveniência dos dados de pesquisa, principalmente compondo evidências para a presunção da autenticidade desses recursos digitais.

Para tanto, traz-se a família PROV, desenvolvida pelo *World Wide Web Consortium* (W3C). Trata-se de um modelo padrão para representar a proveniência digital, integrando metadados de diferentes esquemas para registrar informações sobre entidades, agentes e atividades envolvidas na criação, modificação e uso de recursos digitais.

Estruturado em três classes principais: entidade (objetos digitais), atividade (processos que geram ou modificam entidades) e agente (responsáveis ou participantes das atividades, como pessoas, organizações ou *softwares*)⁴⁹, o PROV estabelece relações como atribuição, delegação, geração, uso e derivação, construindo uma cadeia detalhada e transparente da proveniência dos dados.

Segundo Li e Sugimoto (2014), apesar de sua abrangência, o PROV se destaca especialmente por garantir interoperabilidade e facilitar a troca de informações em ambientes heterogêneos, como a Web e sistemas de *linked data*⁵⁰, sendo ideal para rastrear a origem e trajetória dos dados em contextos como a gestão de dados de pesquisa. Além disso, conforme Haynes (2018), a proveniência registrada pelo PROV é crucial para a preservação digital, pois assegura que os dados não foram adulterados, garantindo sua autenticidade e integridade ao longo do tempo, sobretudo em ambientes complexos e descentralizados, como repositórios científicos.

A autenticidade dos objetos digitais é um requisito fundamental para a preservação a longo prazo, cuja validação depende do registro detalhado da proveniência. Varnienè-Janssen e Kuprienè (2018) definem a proveniência como o conjunto de informações sobre entidades, atividades e agentes envolvidos na criação, modificação e

⁴⁹ Para o aprofundamento específico sobre esta questão, vale a pena consultar Arakaki (2019, p. 119-122), que em sua tese de doutorado elaborou um quadro com a visão geral do crosswalk dos elementos do PROV-O com os padrões MARC, DublinCore, BIBFRAME, Schema.org e Premis.

⁵⁰ Para Bizer *et al.* (2008), o *Linked Data* é apresentado como um paradigma para tornar dados legíveis por máquina e interconectáveis na Web, através da aplicação de princípios de padronização, semântica e interligação, visando aumentar sua utilidade e valor em diversos domínios.

custódia dos objetos digitais, permitindo sua rastreabilidade e assegurando a confiança e integridade desses recursos.

Esse artigo (Varnienê-Janssen; Kuprienė, 2018) destaca a importância da documentação do histórico de transformações, migrações e intervenções técnicas para sustentar evidências de autenticidade e garantir a confiabilidade dos dados para uso futuro. As autoras também ressaltam que nenhum modelo isolado cobre todas as dimensões necessárias para a documentação completa da autenticidade e proveniência, recomendando a complementaridade entre PREMIS e PROV para apoiar funções essenciais de preservação, promover interoperabilidade institucional e agregar valor aos dados digitais, sobretudo em iniciativas colaborativas.

Dessa maneira, PROV e PREMIS atuam como modelos complementares na gestão da proveniência e preservação digital. Enquanto o PROV oferece uma estrutura flexível para modelar relações entre entidades, atividades e agentes, facilitando a interoperabilidade em ambientes heterogêneos, o PREMIS destaca-se por registrar eventos de preservação, agentes e direitos associados, focando na manutenção da integridade e identidade dos objetos digitais em repositórios institucionais.

A integração desses modelos é fundamental para cobrir todas as dimensões da documentação da autenticidade e proveniência, ampliando a confiança, reusabilidade e interoperabilidade dos dados digitais em contextos colaborativos e multidisciplinares.

Assim, a combinação desses modelos possibilita uma gestão mais robusta e completa dos metadados de pesquisa. O PROV pode registrar o histórico de criação e transformação dos dados, enquanto o PREMIS documenta as ações de preservação subsequentes.

Essa abordagem integrada favorece a gestão cumulativa e transparente dos metadados, alinhando-se à proposta desta tese em garantir o registro da proveniência dos dados desde os estágios iniciais da pesquisa até sua preservação digital, de forma que se dê qualitativamente para potencializar seu (re)uso futuro.

Para isso, traz-se o “*Metadata Encoding & Transmission Standard*” – METS (Padrão de codificação e transmissão de metadados) enquanto um esquema de metadados capaz de registrar a proveniência digital dos dados de pesquisa. Esse esquema, desenvolvido em 2001 e mantido pela *Library of Congress*, consiste em um “padrão para codificar metadados descritivos, administrativos e estruturais relativos a objetos dentro

de uma biblioteca digital e expressos usando o esquema de linguagem *World Wide Web Consortium* (W3C)” (site METS [s.d.], tradução nossa).

Em sua última versão, a METS *version 2*, são abarcados metadados para objetos digitais complexos, em texto, vídeo ou imagem e, dependendo do seu uso, pode ser usado para o estabelecimento de SIP, AIP e DIP dentro de um ambiente baseado no OAIS (2024).

Um documento METS, de acordo com o Tutorial METS *version 2* ([s.d.]) consiste em quatro seções principais: Cabeçalho, Seção de metadados (caracterizado por <md>), Seção de arquivo (<file><fileGrp>) e Seção estrutural (<structMap>).

Portanto, para fins desta tese, ressalta-se a Seção de metadados. Essa Seção reúne

[...] informações sobre metadados descritivos, metadados técnicos que descrevem como os arquivos foram criados e armazenados, direitos de propriedade intelectual, metadados relativos ao objeto original do qual deriva o objeto digital e informações sobre a proveniência dos arquivos que compõem o objeto digital (ou seja, relacionamentos de arquivos mestre/derivativos e informações de migração/transformação) (METS, [s.d.], tradução nossa).

Ou seja, essa seção contém todos os metadados relativos ao próprio objeto digital quanto a seus componentes, compondo os principais tipos de metadados fornecidos em um documento METS, que geralmente são: Metadados descritivos sobre o objeto digital; Metadados técnicos (informações sobre criação, formato e uso de arquivos características); Metadados de Direitos de Propriedade Intelectual (informações de direitos autorais e licença); Metadados de origem (metadados descritivos e administrativos relativos à fonte analógica da qual deriva um objeto digital); e Metadados de Proveniência Digital (informações sobre origem/destino relações entre arquivos, incluindo relações mestre/derivadas entre arquivos e informações sobre migrações/transformações empregadas em arquivos entre a digitalização original de um artefato e sua encarnação atual como um objeto digital) (METS, [s.d.], tradução nossa).

Desses cinco tipos de metadados, três podem ser constituídos por metadados PREMIS: os técnicos, de direitos de propriedade intelectual e os de proveniência digital. Os de proveniência digital são encontrados como “<digiprovMD>” no documento METS, conforme figuras abaixo.

Figura 20 – Extrato de um documento METS com exemplo da proveniência digital registrada

```

255 </mets:mdWrap>
256 </mets:techMD>
257 <mets:digitProvMD ID="digitProvMD_1">
258 <mets:mdWrap MDTYPE="PREMIS:EVENT">
259 <mets:xmlData>
260 <premis:event xmlns:premis="http://www.loc.gov/premis/v3" xsi:schemaLocation="http://www.loc.gov/premis/v3 http://www.loc.gov/standards/premis/v3/premis.xsd" version="3.0">
261 <premis:eventIdentifier>
262 <premis:eventIdentifierType>UUID</premis:eventIdentifierType>
263 <premis:eventIdentifierValue>c07a37a8-38af-412c-844f-20a1f0ach352</premis:eventIdentifierValue>
264 </premis:eventIdentifier>
265 <premis:eventType>Ingestion</premis:eventType>
266 <premis:eventDateTime>2025-04-16T13:51:03.943764+00:00</premis:eventDateTime>
267 <premis:eventDetailInformation>
268 <premis:eventDetail></premis:eventDetail>
269 </premis:eventDetailInformation>
270 <premis:eventOutcomeInformation>
271 <premis:eventOutcome></premis:eventOutcome>
272 <premis:eventOutcomeDetail>
273 <premis:eventOutcomeDetailNote></premis:eventOutcomeDetailNote>
274 </premis:eventOutcomeDetail>
275 </premis:eventOutcomeInformation>
276 <premis:linkingAgentIdentifier>
277 <premis:linkingAgentIdentifierType>preservation system</premis:linkingAgentIdentifierType>
278 <premis:linkingAgentIdentifierValue>Archivematica-1.13.2</premis:linkingAgentIdentifierValue>
279 </premis:linkingAgentIdentifier>
280 <premis:linkingAgentIdentifier>
281 <premis:linkingAgentIdentifierType>repository code</premis:linkingAgentIdentifierType>
282 <premis:linkingAgentIdentifierValue>Repositório Digital de Preservação (RDP)</premis:linkingAgentIdentifierValue>
283 </premis:linkingAgentIdentifier>
284 <premis:linkingAgentIdentifier>
285 <premis:linkingAgentIdentifierType>Archivematica user pk</premis:linkingAgentIdentifierType>
286 <premis:linkingAgentIdentifierValue>4</premis:linkingAgentIdentifierValue>
287 </premis:linkingAgentIdentifier>

```

Fonte: Elaborada pela autora (2025)

Figura 21 – Extrato de um documento METS com exemplo da proveniência digital registrada – Eventos e Agentes

```

397 <mets:digitiprov ID="digitiprovMD_5">
398 <mets:mdWrap MDTYPE="PREMIS:EVENT">
399 <mets:xmlData>
400 <premis:event xmlns:premis="http://www.loc.gov/premis/v3" xsi:schemaLocation="http://www.loc.gov/premis/v3 http://www.loc.gov/standards/premis/v3/premis.xsd" version="3.0">
401 <premis:eventIdentifier>
402 <premis:eventIdentifierType>UUID</premis:eventIdentifierType>
403 <premis:eventIdentifierValue>6e216cf4-bd5c-4a44-a814-dbec797fe3a2</premis:eventIdentifierValue>
404 </premis:eventIdentifier>
405 <premis:eventType>creation</premis:eventType>
406 <premis:eventDateIn>2024-06-26T00:00:00+00:00</premis:eventDateIn>
407 <premis:eventDetailInformation>
408 <premis:eventDetail>Pacote de Informações OAIS contendo representante(s) digital(is) oriundo(s) do processo de digitalização de documentos videográficos da VSD relacionados ao título VIII
409 Conferência Nacional de Saúde com reformatação de fita(s) magnética(s) de formato original U-MATIC.
410 </premis:eventDetail>
411 <premis:eventDetailInformation>
412 <premis:eventOutcomeInformation>
413 <premis:eventOutcome>success</premis:eventOutcome>
414 <premis:eventOutcomeDetail>
415 <premis:eventOutcomeDetailNote>Evento de preservação obtido com a digitalização antecedendo à ingestão ao repositório OAIS. Ambiente de digitalização e registro do histórico de processo
416 de digitalização descritos em premis:environment</premis:eventOutcomeDetailNote>
417 </premis:eventOutcomeDetail>
418 <premis:eventOutcomeInformation>
419 <premis:linkingAgentIdentifier>
420 <premis:linkingAgentIdentifierType>preservation system</premis:linkingAgentIdentifierType>
421 <premis:linkingAgentIdentifierValue>Archivematica-1.13.2</premis:linkingAgentIdentifierValue>
422 </premis:linkingAgentIdentifier>
423 <premis:linkingAgentIdentifier>
424 <premis:linkingAgentIdentifierType>organization</premis:linkingAgentIdentifierType>
425 <premis:linkingAgentIdentifierValue>VideoSaúde Distribuidora da Fundação Oswaldo Cruz</premis:linkingAgentIdentifierValue>
426 </premis:linkingAgentIdentifier>
427 </premis:event>
428 </mets:xmlData>
429 </mets:mdWrap>
430 </mets:digitiprov ID="digitiprovMD_5">
431 <mets:mdWrap MDTYPE="PREMIS:AGENT">
432 <mets:xmlData>
433 <premis:agent xmlns:premis="http://www.loc.gov/premis/v3" xsi:schemaLocation="http://www.loc.gov/premis/v3 http://www.loc.gov/standards/premis/v3/premis.xsd" version="3.0">
434 <premis:agentIdentifier>
435 <premis:agentIdentifierType>preservation system</premis:agentIdentifierType>
436 <premis:agentIdentifierValue>Archivematica-1.13.2</premis:agentIdentifierValue>
437 </premis:agentIdentifier>

```

Fonte: Elaborada pela autora (2025)

Essas duas figuras acima (20 e 21) exemplificam o Padrão de metadados PREMIS já embutidos na estrutura do METS, o que demonstra a possibilidade do registro detalhado, cumulativo e contínuo dos metadados gerados ao longo do ciclo de vida dos dados de pesquisa. Além disso, o documento METS permite que humanos e máquinas possam ter acesso de forma transparente e interpretar os metadados julgados relevantes para a presunção de sua autenticidade.

Isso posto, julga-se necessário expor que, embora a ação de preservação se encontre apenas na letra “e” (“Gerenciar, armazenar e preservar”) no Modelo do ciclo de vida dos dados de pesquisa da JISC, apresentado na subseção 3.1.1 desta tese, o estabelecimento de suas estratégias e metodologias devem fazer parte das aspirações inerentes à primeira fase “criação dos dados e depósito”, mais especificamente em seu primeiro tópico “Planejar e projetar” que engloba, inclusive, a elaboração do PGD.

Para compreender, no entanto, o que deve ser incorporado nessas fases iniciais ao ponto de garantir grau de qualidade da preservação digital dos dados de pesquisa, julga-se necessário expor o entendimento teórico-metodológico sob o escopo da Arquivologia e, então, suas possíveis contribuições para o tema em discussão.

5 A GESTÃO DE DOCUMENTOS: CONTRIBUTO DA ARQUIVOLOGIA PARA A QUALIFICAÇÃO DA PRESERVAÇÃO DIGITAL DE DADOS DE PESQUISA

Esta seção organiza-se de modo a refletir o percurso analítico e interpretativo que fundamenta a contribuição arquivística desta tese. Sua organização segue uma lógica de aprofundamento progressivo: inicia-se pela apropriação da Arquivologia como referencial teórico e, em seguida, aplica-se seus fundamentos à leitura crítica e propositiva do contexto contemporâneo da gestão de dados de pesquisa.

Para tanto, apresenta inicialmente um breve histórico da área, de modo a fundamentar o surgimento da gestão de documentos e sua aplicação prática, com ênfase no cenário brasileiro. Na sequência, são abordadas as principais funções e técnicas arquivísticas: classificação, avaliação, descrição e identificação, bem como a importância do acompanhamento especializado para sua efetivação.

Essa abordagem sustenta a compreensão teórica da necessidade dos processos de documentação e de registro de metadados para o amparo da preservação digital qualitativa dos dados de pesquisa. Além disso, evidencia a contribuição da Arquivologia como campo científico capaz de oferecer instrumentos conceituais e metodológicos que qualificam a gestão de dados de pesquisa, especialmente em seus estágios iniciais, para garantir a proveniência, organicidade, autenticidade e confiabilidade desses dados.

5.1 A ARQUIVOLOGIA E UM DE SEUS CONCEITOS BASILARES: A GESTÃO DE DOCUMENTOS.

A Arquivologia costuma situar seu marco moderno na Revolução Francesa (1789), baseada no princípio do acesso público, na responsabilidade estatal pela preservação e na institucionalização da administração arquivística (Schellenberg, 1974). No entanto, o desenvolvimento de reflexões do campo se volta, inicialmente, ao que Santos (2010, p. 67) chama de “arquivo instituição”, com pensamentos fundamentados no diálogo com a História.

Portanto, seus princípios e teorias tinham por premissa prover recursos que atendessem às necessidades daquelas que seriam as instituições custodiadoras do, propriamente dito, Arquivo (*Archive*). Nessa época, os primeiros estudos arquivísticos,

inclusive com a proposição de seus princípios basilares como o “Princípio de respeito à ordem original” e o “Princípio da proveniência”⁵¹⁵², dão início à construção da Arquivologia enquanto um campo do saber embasado por contornos científicos.

Durante o século XX, sobretudo no período da Segunda Guerra Mundial, a Arquivologia passa por uma reconfiguração conceitual, marcada por tensões entre sua dimensão histórica e administrativa. Nesse contexto, segundo Da Cunha Marques (2017), diferentes correntes teóricas e práticas passam a coexistir em função das distintas formas de produção e uso dos arquivos em cada realidade nacional, refletindo uma crise de identidade da disciplina e a ascensão de uma abordagem mais técnica.

Isso porque o volume exponencial de documentos gera a urgência de uma metodologia padronizada e eficiente para garantir a organização, preservação e acesso, sob o risco de uma perda informacional irreversível (Schellenberg, 1974). Surge, então, entre as décadas de 1940 e 1980, a chamada Arquivologia Moderna, ou, segundo Schmidt (2015), fase “paradigmática”⁵³, que tem como um de seus marcos a publicação, em 1956, de uma importante obra “Arquivos Modernos - Princípios e técnicas”, de Schellenberg, marcada pela preocupação com os documentos administrativos chamados pelo autor de “documentos modernos”.

Nesse cenário, ganha destaque a consolidação do *records management*, traduzido nesta tese como Gestão de documentos, cujo objetivo é permitir a intervenção arquivística desde a fase de produção e tramitação dos documentos. Schellenberg (1974, p. 46) afirma que uma administração eficiente exige atenção aos documentos “desde sua criação até o momento em que são transferidos para um arquivo de custódia permanente ou são eliminados”.

O surgimento da gestão de documentos revelou um panorama aberto para novas concepções, como a noção do “Ciclo vital dos documentos” e a “Teoria das Três Idades”,

⁵¹ A opção por citar somente esses dois princípios tem por base o estudo realizado por Rangel e Silva (2021) que, após análise dos dicionários brasileiros de terminologia arquivística, identificaram que somente os princípios da proveniência e do respeito à ordem original são princípios essencialmente arquivísticos “uma vez que somente a proveniência e o respeito à ordem original podem ser aplicados a qualquer conjunto documental independente da circunstância” (p. 12)

⁵² Schellenberg (1974) considerava que o valor de evidência dos documentos é protegido a partir da aplicação desses dois princípios.

⁵³ Esta fase se caracteriza “pelas contribuições de Schellenberg; pela institucionalização acadêmica da Disciplina; pela criação do Conselho Internacional de Arquivos (CIA); pelo desenvolvimento de teorias como a Teoria das Três Idades; e pelo entendimento do documento de arquivo como objeto indutor de reflexões no âmbito da comunidade científica arquivística, especialmente nas dimensões científica e administrativa, direcionadas à gestão” (Da Cunha Marques, 2016, p. 195)

ambas iniciadas nos Estados Unidos. Jardim (2015) aponta que esses termos são tão imbricados, em virtude da “Teoria das Três Idades” ter suas bases na concepção de “Ciclo vital de documentos”, que frequentemente são geradas confusões entre eles.

O chamado “Ciclo vital” (*Records Lifecycle*), compreende “as sucessivas fases por que passam os documentos de um arquivo, desde a produção até a guarda permanente ou eliminação, ou seja, produção, utilização e destinação final” (Arquivo Nacional, 2019, p. 21). O autor Cruz Mundet (2012) ressalta que mesmo tendo passado por diversas análises desde o final do século XX sobre a possibilidade, ou não, de comportar em sua versão original toda a complexidade inerente aos documentos eletrônicos, o Ciclo vital continua sendo válido para representar a vida e atividade dos documentos.

Já a Teoria das Três Idades, conforme disposto no Dicionário Brasileiro de Terminologia Arquivística, é entendida como “teoria segundo a qual os arquivos são considerados arquivos correntes, intermediários ou permanentes, de acordo com a frequência de uso por suas entidades produtoras e a identificação de seus valores primário e secundário” (Arquivo Nacional, 2005, p. 160).

Até aqui, mesmo sendo possível perceber que a tradição estadunidense tenha mantido sob sua base a influência europeia, trata-se do emprego de uma nova perspectiva para o campo do saber arquivístico. Adota-se, portanto, um olhar mais administrativo, em busca de eficiência e economia de recursos. Assim, surge a necessidade da realização de avaliações que embasem eliminações criteriosas de documentos considerados destituídos de valor ao longo de seu ciclo.

O conceito de *records management*, em virtude de ter emergido a partir de uma necessidade da administração pública e não da prática ou teoria dos arquivos (Jardim, 1987), adota metodologias e procedimentos ao longo do ciclo de vida, a partir de definições oriundas do estabelecimento claro do que seria uma divisão entre os ambientes do “*records management*” e “*archives*”. No *records management* (Gestão de documentos), as ações estão voltadas à organização, controle e manutenção de documentos que ainda se encontram junto ao seu produtor e realizadas pelos *records managers* (gestores de documentos). Já no *archives*, há a responsabilidade pela preservação daquilo que veio a se tornar o patrimônio documental⁵⁴, realizada pelos *archivists* (arquivistas).

⁵⁴ A autora Marcia Rodrigues em seu artigo “Patrimônio Documental Nacional” (2016) aponta que trata-se de um conceito ainda em construção e apresenta a trajetória legal e teórica que vem sendo utilizada para o

Essa separação gera preocupações de caráter teórico e fomenta o debate sobre o que deve ser considerado material de arquivo (Santos, 2010). Cumpre esclarecer que o uso do termo *records* ainda possui algumas interpretações. Nesta tese, adota-se a concepção de Yeo (2015), que o entende como “representações de ocorrências”.

Ocorrências são fenômenos que possuem final temporal. *Records*⁵⁵ as representam de uma forma duradoura; eles continuam disponíveis depois que as ocorrências acabaram. *Records* são usualmente associados às atividades de negócios e transações, mas nós podemos ter *records* em outras ocorrências: declarações feitas, questões perguntadas ou sentimentos expressos. Em cada caso, a ocorrência carece de uma persistência intrínseca, mas o *record* é a representação continuada disso (Yeo, 2015, p. 51).

Desde que surgiu a Gestão de documentos tem sido objeto de diversas interpretações e defesas sobre sua aplicação prática e a maneira como ela se insere na dinâmica arquivística. De toda forma, Caya (2004 *apud* Jardim, 2015) afirma ser preciso que o arquivista intervenha no contexto de produção dessa documentação⁵⁶.

Reforça-se essa necessidade ao compreender que, “somente com procedimentos de gestão arquivística é possível assegurar a autenticidade dos documentos arquivísticos digitais” (CONARQ, 2022, p. 25). Ainda mais porque a identidade e integridade, que a compõem, “são constatadas à luz do contexto (jurídico-administrativo, de proveniência, de procedimentos, documental e tecnológico) no qual o documento arquivístico foi produzido e usado ao longo do tempo” (CONARQ, 2022, p. 201).

Além disso, a presunção de autenticidade é uma inferência “feita a partir de fatos conhecidos sobre a maneira como aquele documento foi produzido e mantido” (CONARQ, 2022, p. 215).

Dessa forma, esta tese assume que os dados de pesquisa, por seu valor de evidência, devem ser geridos a partir de procedimentos que transpareçam sua autenticidade e potencializem sua confiabilidade. Julga-se oportuno, então, destacar e esclarecer duas das correntes teóricas que influenciam na execução prática da Gestão de documentos no cenário internacional.

uso desse termo no Brasil. Albuquerque e Silveira (2023) acrescentam que não é tema exclusivo da Arquivologia, sendo estudado também por pesquisadores da Biblioteconomia, História, Letras e outras áreas formação.

⁵⁵ Na tradução do texto de Yo (2015), as autoras Oliveira e Silva ressaltam que alguns termos não foram traduzidos do inglês por determinação do autor.

⁵⁶ É preciso esclarecer que no Brasil, conforme será apresentado na subseção 5.1.2, não são praticadas distinções entre arquivistas e gestores de documentos.

5.1.1 Correntes teóricas da gestão de documentos

A segmentação e linearidade adotadas na Teoria das Três Idades passou por críticas quanto à sua aplicabilidade junto aos documentos digitais. O autor canadense Atherton, reconhecido por renascer a ideia de “continuum” (1985), aponta que o Ciclo vital foi impactado pelo computador à medida que, com os dados digitais, seus estágios não poderiam mais ser separados, dada sua natureza e volatilidade (Atherton, 1987).

O autor reconhece que o ciclo de vida tem utilidade para promover “sentido de ordem, uma abordagem sistemática para a gestão global da informação registrada” (Atherton, 1987, p. 47), mas propõe para esse novo cenário a interligação de quatro fases (produção/recebimento, classificação, avaliação, manutenção e utilização da informação) com vistas a constituir uma gestão contínua das informações registradas e que envolvesse, em diferentes graus, gestores de documentos e arquivistas.

Em uma nota de rodapé do livro do professor arquivista Sérgio Conde de Albite Silva intitulado “A preservação da informação arquivística governamental nas políticas públicas do Brasil”, de 2008, o autor apresentou de maneira clara que, embora seja usual o uso da expressão “gestão documental”, outras formas de apresentá-la são amplamente utilizadas.

O Dicionário brasileiro de terminologia arquivística publicado pelo Arquivo Nacional em 2005 não inclui o verbete “gestão da informação”. No entanto, é cada vez mais frequente na literatura arquivística o uso de “gestão da informação arquivística”. Isso deriva das discussões sobre o objeto da Arquivologia ser os arquivos, os documentos de arquivo ou, mais recentemente, a informação arquivística. Autores como Jardim (1999, p. 29), Fonseca (2005, p. 59), Couture e Rosseau (1998, v. 30, p. 24, 64-67) e Thomassen (1999), entre outros, abordam o tema. O próprio CONARQ, na Carta para Preservação do Patrimônio Arquivístico Digital, utiliza tanto o termo “informação arquivística” como “documento arquivístico”. É usual ainda a expressão “gestão documental” (Silva, 2008, p. 25, grifo nosso).

O autor (2008) demonstra que um novo direcionamento vem sendo adotado para considerar o objeto da Arquivologia, do documento à informação, e o termo “gestão”, portanto, acompanhando essa tendência. Ao apresentar a visão de Malheiro da Silva (2000), Silva (2008, p. 44) aponta a “nítida predominância do conteúdo sobre o suporte”, e esclarece que o que Jardim (1999) chama de “informação arquivística”, Ampudia Mello

(1988) chama de “informação institucional” e Couture, Ducharme e Rousseau (1988) denominam como “informação orgânica registrada”.

Essas diversas reflexões se encontram em um cenário onde também estão em sistematização alternativas ao Modelo do Ciclo vital na gestão de documentos em meio digital. Há, portanto, a emergência de novas correntes arquivísticas enquanto possibilidades de adaptação do campo às novas demandas que brotam junto à evolução tecnológica constante. Nesse sentido, destacam-se as correntes australiana (*Records continuum*) e canadense (Arquivística Integrada).

5.1.1.1 Corrente australiana: *Records Continuum*

A teoria de *Records Continuum* teve sua divulgação ampliada mais recentemente, na década de 90, através do arquivista australiano Frank Upward. No entanto, suas ideias já são mais antigas. Há indícios de que Upward tenha “bebido na fonte” de reflexões anteriores declaradas por seu conterrâneo Ian Maclean que, nas décadas de 50 e 60, já entendia a interseção entre as profissões até então segmentadas do gestor de documentos e o arquivista.

Upward também teria sido influenciado pelo arquivista canadense Jay Atherton, que efetivou a abordagem teórica do *continuum* em seu artigo “*From life cycle to continuum: some thoughts on the records management-archives relationship*”, publicado em 1985. Nele o autor (1985) se opõe à estrutura do Modelo de Ciclo vital e propõe um mais enxuto e unificado que reflita as fases em um processo contínuo, por considerá-las interligadas e executáveis simultaneamente.

Ao descartar a ideia de um ciclo com demarcações de limites e separações evidentes, Atherton (1985) levanta como principal conceito abordado nessa visão a compreensão de que os registros são elementos dinâmicos que estão sempre em processo de formação e reinterpretação, influenciados pelo contexto e pelas transações nas quais estão inseridos e não objetos estáticos e isolados (McKemmish, 2005).

A gestão integral dos documentos, concebida pelos *Records Continuum*, que, segundo Upward (2005), compreende o entorno tecnológico e se apoia numa visão sincrônica de espaço-tempo (ou seja, que opera através do tempo e não no tempo) constituída de maneira multi-relacional, revela o entendimento de uma gestão documental unificada, não linear e contínua.

Essa visão direciona importância ao papel do contexto na criação, interpretação e uso dos registros, a partir do entendimento de que eles não são produtos de um único momento no tempo, ou seja, não são apenas evidências de eventos passados. Os registros estão sempre imersos e influenciados pelo contexto organizacional, legal, social, cultural, ideológico no qual são criados, usados e interpretados. Portanto, seu valor e sua função dependem da compreensão desses contextos, o que demonstra que sua interpretação não pode ser entendida como neutra.

A perda da fisicalidade que ocorre quando os documentos são capturados eletronicamente está forçando os arquivistas a reavaliarem entendimentos básicos sobre a natureza dos documentos de atividades sociais e organizacionais, e de suas qualidades como evidência. [...] A ordenação e localização física de tais documentos compreendem uma visão de suas relações contextuais e documentais, mas não apresentam os múltiplos entendimentos do que é uma realidade complexa. O papel custodial tradicional assume outra dimensão quando é aceito que o documento é apenas parcialmente manifesto no que está nas caixas das prateleiras do arquivo. A finalidade dos sistemas de arquivo é garantir que os documentos sejam preservados no contexto de sua criação e uso e que mantenham suas qualidades como evidência, de modo que, quando recuperados para um uso futuro, seu significado e importância possam ser entendidos (McKemmish, 1994, *apud* Costa Filho; De Sousa, 2017, p. 47, grifo nosso).

Costa Filho e De Sousa (2017) acreditam que o *records continuum* forneça subsídios teóricos para que a complexidade trazida pelos documentos digitais seja compreendida. Matlala e Maphoto (2020) apontam que esse conceito é apresentado em diversos estudos como o maior contraponto ao que já estava previamente estabelecido.

No entanto, pelo menos desde início dos anos 2000, debates internacionais sobre as práticas arquivísticas vêm sendo embasados por análise dessa corrente. Cabe identificar que a primeira norma internacional para gestão de documentos, a ISO 15489 “*Information and documentation - Records management*”, publicada em 2001⁵⁷, teve como gênese a Norma australiana AS 4390 (1996), elaborada em resposta à norma de qualidade ISO 9000, que aponta a importância da gestão de documentos e registros para garantir a qualidade e eficiência dos processos organizacionais (Sanjuan, 2015).

Em 2016, a ISO 15489 foi atualizada como “*Information and documentation – Records management – Part 1: Concepts and principles*” e passou a contemplar

⁵⁷ Na versão de 2001, a ISO 15489 tinha a seguinte composição: “*ISO 15489-1:2001 Information and documentation - Records management – Part 1: General*” e era acompanhada de um informe técnico: “*ISO/TR 15489-2:2001 Information and documentation - Records management – Part 1 – Parte 2: Guideline*”.

“acréscimos conceituais em que se pode mencionar, dentre outros, a ideia de documento como evidência (de origem do *Records Continuum*) e análise de riscos em sistemas de gestão de documentos” (Teixeira e Silva; Dos Santos, 2019, p. 12).

Nesse documento, a gestão de documentos é responsável pelo

[...] controle eficiente e sistemático da produção, recebimento, manutenção, uso e destinação de documentos de arquivo, incluindo processos para captura e manutenção de provas e informações sobre atividades e transações de negócio na forma de documentos de arquivo (ISO 15489, 2016, p. 3.15).

Portanto, sua execução ao longo da produção, captura e gerenciamento de documentos os mantém autênticos, confiáveis, íntegros e utilizáveis. Além disso, ao estabelecer que a gestão de documentos deve ser norteadada por cinco princípios fundamentais, destacam-se, para fins desta tese, que “o documento é prova oficial das atividades de negócio” e que “o documento consiste em conteúdo e metadados/contexto” (Teixeira e Silva; Dos Santos, 2019, p. 12).

Para isso, a partir desses princípios, são definidas seis classes de metadados que, em conjunto, registram proveniência e organicidade dos documentos e, portanto, reforçam a necessidade de que informações sobre identidade, descrição, utilização, plano de eventos, histórico e relações são elementos essenciais para a confiabilidade do documento.

Adicionalmente, ao incluir em seu processo a produção, captura, classificação e indexação, controle de acesso e permissões, armazenamento, uso e (re)uso, migração e conversão, e destinação, a ISO 15489:2016 indica as boas práticas para cada um, conforme o quadro a seguir:

Figura 22 – Boas práticas de gestão de documentos

Processo	Boas-práticas recomendadas
Produção de documentos	- Identificação de requisitos por meio de avaliação; - produção de conteúdo e metadados para registrar contexto.
Captura de documentos	- Atribuição de identificador único; - metadados no ponto de captura; - registro de relacionamentos lógicos.
Classificação e indexação	- Vinculação do documento à atividade documentada; - Plano de classificação baseado em funções, atividades e processos de trabalho; - possibilidade de reclassificação; - indexação por metadados para facilitar recuperação.
Controle de acesso e permissões	- Sistemas de permissão e restrição; - regras de acesso e permissão formalmente declaradas e mantidas atualizadas; - regras de acesso registradas em metadados.
Armazenamento de documentos	- Ambiente e mídia apropriados; - Materiais de proteção e procedimentos especiais de manuseio; - segurança física; - segurança da informação; - procedimento para recuperação em caso de desastres; - dados de armazenamento registrados em metadados; - avaliação rotineira dos sistemas de armazenamento.
Uso e reuso	- Reprografia ou conversão digital; - Migração; - Dependências técnicas do documento registradas em metadados; - Processo de conversão registrado em metadados; - Plano que assegura usabilidade contínua em caso de desastre.
Migração e conversão	- Migração; - Digitalização; - Conversão de suporte; - Desativação de sistema; - Processos de migração e/ou conversão registrados em metadados.
Destinação	- Eliminação sempre autorizada; - Eliminação por destruição completa; - Manutenção de documentos pertencentes a litígios judiciais pendentes; - Ações de destinação documentadas.

Fonte: David (2018, p. 37).

É importante trazer essa norma para análise pois, de uma forma geral, ao passo de ter sua origem fundamentada na dinâmica australiana de *records continuum*, ela “universaliza” as boas práticas da gestão de documentos que, nos dias atuais, perpassam o tratamento ao documento (suporte) e se destinam cada vez mais à informação (conteúdo).

Isso porque, diante de uma visão entendida como “pós-moderna” da Arquivologia, o documento, de acordo com Cook (2001), deixou de ser um objeto físico para ser um dado conceitual controlado por metadados que, de maneira virtual, combinam conteúdo, contexto e estrutura, no intuito de fornecer evidência da atividade ou a função do criador. Esse fato transforma o documento num agente ativo que desempenha constante papel nas vidas dos indivíduos, organizações e sociedade

Assim, é possível verificar que a Corrente australiana foca na crítica à linearidade do “Ciclo vital” no cenário digital e na defesa de uma gestão contínua na qual a relevância do registro de contexto e proveniência é essencial para a manutenção de seu potencial de evidência.

Todavia, conforme apontam Matlala e Maphoto (2020), sua literatura ainda carece de mais investigações empíricas sobre a aplicação de sua teoria na prática, tanto na visualização dos padrões estabelecidos quanto na avaliação dos seus possíveis benefícios, na influência que sua forma afeta a gestão de documentos e na extensão de seu impacto.

5.1.1.2 Corrente Canadense: Arquivística Integrada

Enquanto o *Records Continuum* propõe um arcabouço teórico para a superação do Ciclo vital no ambiente digital, a Arquivística Integrada (corrente canadense) se apresenta como o modelo metodológico adotado por esta tese devido à sua aplicação consolidada no Brasil.

Quanto à corrente canadense, Tognoli (2010) aponta em sua dissertação a Arquivística Integrada como uma das contribuições epistemológicas canadenses para a construção da Arquivística Contemporânea. Segundo a autora (2010), essa Arquivística se baseia na Arquivística Francesa (arquivos históricos e Princípio da proveniência) e na Gestão de documentos norte-americana, tendo a “informação registrada” como seu objeto de estudos.

Assim, a chamada Arquivística Integrada, que visa contemplar todo o ciclo vital dos documentos (Gagnon-Arguin, 1992), passa a promover uma “redescoberta” do “Princípio da proveniência” (defendida pelos autores Nesmith, Ketelaar, Bearman, Cook, Duranti) com base nos novos cenários, voltada agora à informação, ao conhecimento e ao contexto de produção dos documentos (Marques, 2013).

Isso “leva os arquivistas a entenderem a contextualização do documento, sua função, seu criador, sua forma, e seu papel na instituição que o criou” (Tognoli, 2010, p. 66), como forma de refletir sobre a necessidade de que a identidade das informações arquivísticas em meio eletrônico pudesse ser preservada.

[...] a partir dos anos 80, uma nova “corrente” surge no Canadá francês com a intenção de reintegrar a disciplina, trabalhando novamente com todo o ciclo

documental, dos documentos correntes até os permanentes. É a chamada Arquivística Integrada (Tognoli, 2010, p. 33).

A partir disso, em 1994, Luciana Duranti e Terry Eastwood passam a trabalhar no projeto “*The preservation of the Integrity of Electronic Records*” conhecido como “Projeto UBC”, que tem por objetivo identificar e definir, de maneira conceitual, a natureza de um documento eletrônico e os critérios para criar, manter e conservar confiáveis e autênticos os documentos eletrônicos enquanto estiverem sob a guarda de seu criador, em uma fase ativa ou semi-ativa (correspondendo, a grosso modo, às fases corrente e intermediária, respectivamente).

Sobre esse Projeto, Tognoli (2010) analisa que

De um ponto de vista teórico, esse trabalho resultou em uma pequena revisão nos conceitos, integrando a Arquivística a outras disciplinas, como a Diplomática. **Do ponto de vista metodológico, produziu regras para a criação e manutenção de um registro pelo criador em meio eletrônico** (Tognoli, 2010, p. 96, grifo nosso).

Cabe ressaltar que os estudos promovidos por Luciana Duranti, da Universidade de British Columbia, em Vancouver, são reconhecidos internacionalmente e tem um papel fundamental no desenvolvimento de trabalhos aprofundados sobre a aplicação dos princípios e métodos promulgados pela Diplomática Clássica⁵⁸ aos documentos contemporâneos.

A chamada “Diplomática Arquivística” conhecida atualmente como “tipologia documental” (Belloto, 2008) que, segundo Souza (2022), nasceu da concretização do que já vinha ocorrendo na Arquivologia, possibilita redescobrimento e inovações dos métodos diplomáticos, unificando os dois campos de estudo: Arquivística e Diplomática.

⁵⁸ A Diplomática clássica nasceu com a publicação do monge Beneditino Jean Mabillon, em 1681, como o “estabelecimento de regras e termos certos e precisos pelos quais os instrumentos autênticos podem ser distinguidos dos falsos, e instrumentos certos e originais dos incertos e suspeitos” (MABILLON, *apud* MACNEIL, 2000, p.20), em resposta a acusações apostas no segundo volume da *Acta Sanctorum* pelo padre jesuíta Papenbroek sobre a não autenticidade dos documentos guardados nos arquivos do Mosteiro de Saint Denis. Se historicamente a posição instrumental dos arquivos face à paleografia e à diplomática tenha sido alterada aos poucos e tenha permitido que a arquivística ensaiasse sua afirmação e autonomia àquelas disciplinas (Santos, 2010, p. 69), em 1987, a publicação da italiana Paolla Carucci intitulada “*Il documento contemporaneo: Diplomática e criteri di edizione*”, onde a autora aplica os conhecimentos da Diplomática aos documentos públicos administrativos produzidos por uma organização no curso normal das funções, a partir da análise do processo administrativo e não mais documental como era feito pelo diplomatas italianos tradicionais, redefinindo assim o conceito de documento diplomático (Tognoli, 2010), faz com que elas se cruzem novamente.

A partir do estudo da tipologia documental, o arquivista é capaz de identificar todo o contexto de criação do documento. Além das fontes usuais como os organogramas e regimentos, ele tem em mãos mais um instrumento. Com isso, quando essas fontes não forem suficientes o bastante para recriar esse contexto – elas nunca são, uma vez que a informação mais completa é aquela que o próprio documento, a peça documental, fornece – **o arquivista pode se valer do estudo da tipologia documental para recriar o contexto de produção** daquele conjunto documental (Tognoli, 2008, p. 834, grifo nosso).

Diante desse cenário, ao aprofundar estudos sobre a correlação entre a Diplomática e a Arquivística que lida, contemporaneamente, com o documento digital, Duranti (2002) compreende que a presunção de autenticidade deve ser assegurada e mantida ao longo do tempo. Para isso, a autora (2005) esclarece sobre a necessidade de que seja definida e conservada a identidade dos materiais junto à proteção de sua integridade. Ações atribuídas ao responsável pela manutenção dos objetos e passada posteriormente para o preservador.

Isso posto, reforça-se a importância da execução adequada da gestão de documentos e da integração entre os agentes envolvidos ao longo do ciclo de vida enquanto recurso basilar para viabilizar o seu uso (agora e no futuro) de maneira segura e confiável. Essa visão dilui as barreiras metodológicas e procedimentais entre os *records* e os *archives*.

É diante desse cenário que nasce o Projeto InterPares, cujo objetivo é

[...] desenvolver o conhecimento essencial para a preservação a longo prazo de registros autênticos criados e/ou mantidos em formato digital e fornecer a base para normas, políticas, estratégias e planos de ação capazes de garantir a longevidade de tal material e a capacidade de seus usuários confiarem em sua autenticidade (InterPares Project, [s.d.], tradução nossa).

Melo (2021) acrescenta que essa confiança na autenticidade, conforme pretensão do Projeto InterPares, se dá a partir da análise da proveniência dos conjuntos documentais, por meio da Diplomática (Melo, 2021, p. 152).

No Glossário do Projeto InterPares são definidas as diferenças entre *records*, *documents*, *information* e *data*. A saber:

- *Record*: um documento criado ou recebido no curso de uma atividade prática como um instrumento ou um subproduto de tal atividade, e separado para ação ou referência, ou seja, o que seria o documento de arquivo;

- *Document*: uma unidade indivisível de informação constituída por uma mensagem afixada a um meio (gravada) de forma sintática estável, tendo forma fixa e conteúdo estável;
- *Information*: um conjunto de dados destinados à comunicação através do espaço ou do tempo; e
- *Data*: a menor unidade significativa de uma informação (InterPares Project, [s.d.], tradução nossa).

A partir dessa compreensão, julga-se pertinente apontar que diante do que se compreende nessa corrente canadense, há a possibilidade de considerar os dados de pesquisa como objetos da Arquivologia. Além da questão dos dados de pesquisa enquanto uma evidência já direcionarem a Arquivologia como um campo do saber com potencial colaborativo para sua gestão, o fato dos dados fazerem parte do que Couture, Ducharme e Rousseau (1988) denominam como “informação orgânica registrada”, reforça esse entendimento.

Dessa forma, a possibilidade de que a metodologia aplicada à gestão de documentos arquivísticos seja aplicada ao contexto da gestão dos dados de pesquisa pode propiciar impacto positivo na manutenção de autenticidade desses dados.

Considera-se relevante apontar que, ao citar Jenkinson (1965), Melo (2021) aponta dois conceitos desenvolvidos pelo autor e que, nesse momento, podem ser utilizados para demonstrar as necessidades em comum dos dois objetos (documentos arquivísticos e dados de pesquisa): a imparcialidade e a autenticidade.

A imparcialidade, de acordo com o estudioso, caracteriza os documentos de arquivo como inerentemente verdadeiros (ao dizer ser o que eles se propõem a registrar). Assim, sua função é potencialmente seu **valor como prova inequívoca do passado**. Já a **autenticidade** define-se pela característica de o documento ser livre de suspeitas, ou seja, livre de adulterações, ao registrar seguramente os fatos e os atos passados e preservados. Ambas as características, ainda de acordo com o autor, **estão vinculadas à noção de custódia, como a responsabilidade dos produtores e seus legítimos sucessores em preservarem os documentos, garantindo, assim, que os arquivos não sofram corrupções ao longo do tempo**. A custódia, dessa forma, associa-se à forma pela qual os **documentos são sucessivamente passados e preservados de sucessor a sucessor** (Melo, 2021, p. 152, grifo nosso).

Assim, com base em expertises arquivísticas, reforça-se que a gestão de documentos é capaz de fornecer um campo seguro para a execução da gestão de dados de

pesquisa. No entanto, para isso, antes de desdobrar as possíveis interlocuções das áreas e as contribuições que o campo da Arquivologia pode trazer, faz-se necessário, então, compreender como seu princípio basilar vem sendo revisitado, e fortalecido, ao ponto de continuar provendo orientação essencial para que os documentos arquivísticos sejam registrados com informações que reforcem sua autenticidade.

Cumpre esclarecer que, toda aplicação deve levar em consideração as especificidades do contexto ao qual se vincula e que, diante disso, apresentam vantagens e limitações. Isso não é diferente do que pode ser reconhecido nas duas correntes apresentadas. Opta-se, portanto, para o desdobramento desta tese, pela intenção de apresentar uma abordagem que integra os fundamentos tradicionais com práticas contemporâneas, alinhadas às demandas da preservação digital qualitativa e às possibilidades extrínsecas e intrínsecas do campo no contexto brasileiro.

5.1.2 A gestão de documentos no cenário brasileiro

Ao considerar os mais diversos marcos da prática e dos pensamentos arquivísticos que delineiam correntes e abordagens, é preciso que sejam conhecidas as práticas arquivísticas nacionais que constituem o estofo para a construção de um caminho à promoção do tratamento técnico dos dados de pesquisa nas instituições brasileiras. Para tanto, nesta subseção são concentrados esforços na análise do que é estudado, discutido e praticado no cenário arquivístico brasileiro.

Conforme apresentado nesta tese, o *Records Continuum* (corrente australiana) fornece um arcabouço teórico que visa superar a linearidade do Ciclo vital. No entanto, a Arquivística Integrada (corrente canadense), representa o modelo metodológico adotado por esta tese sob a justificativa de sua aplicação e consonância com as políticas nacionais de gestão de documentos e preservação digital.

Ao considerar todos os vieses que atravessam a atmosfera teórica da Arquivologia, e conseqüentemente a sua prática, é possível afirmar que cada país foi constituindo seu campo arquivístico com formas particulares diante de seus contextos nacionais. São as “tradições arquivísticas” nacionais (Jardim, 2015) forjadas a partir de suas circunstâncias histórico-sociais desde o século XIX.

Essa construção nacional pode mesclar diferentes aspectos. No caso brasileiro, a construção dessa tradição se deu a partir de correntes, influências e colaborações

internacionais, com visitas, estudos, relatórios e cursos técnicos promovidos por profissionais e pesquisadores estrangeiros (Marques, 2013).

Isso quer dizer que ao mesmo tempo que as diversas correntes e abordagens arquivísticas refletem os pensamentos desenvolvidos e propagados ao redor do mundo, a adoção do *modus operandi* do fazer arquivístico precisa considerar questões específicas. As questões do campo, sejam extrínsecas (cultural, político e social, por exemplo) ou intrínsecas (capacidade técnica e recursos tecnológicos adequados à adoção de determinados procedimentos, por exemplo), em níveis nacionais, mas também regionais e institucionais, influenciam de maneira direta as reais possibilidades de implementação de uma ou outra proposta importada. O sucesso, ou não, da sua execução e atingimento de objetivos também são influenciados por essas questões.

Lopes (1998) aponta, por exemplo, que muitas das soluções adotadas em outras áreas do conhecimento ou em outros países não respondem aos problemas brasileiros. Assim sendo, embora seja reconhecido que a trajetória da Arquivologia, enquanto uma disciplina científica (Fonseca, 2005), precisa ser muito bem aprofundada para delinear fielmente todo o processo pelo qual atravessou até o presente momento, esclarece-se que, para esta tese, estão sendo trazidos apenas os aspectos considerados contributos ao tema em questão.

Ao citar Lopes (1997-1998), Marques (2013) aponta que o autor considera frágil e até questionável a configuração da tradição arquivística dos países latinos, ao mesmo tempo que, segundo ele, no pensamento arquivístico internacional, coexistem três correntes:

“1) A da **Arquivologia Tradicional**, que se preocupa predominantemente com os arquivos históricos e é representada pela França e outros países do continente europeu ocidental; 2) a da **records/information management**, voltada para os arquivos correntes e representada pelos Estados Unidos; 3) a da **Arquivística Integrada**, corrente que se consolidou no Canadá, sobretudo no Quebec.” (Marques, 2013, p. 96, grifo nosso).

Com base nessa visão trazida por Marques (2013), compreende-se que no Brasil, atualmente, há uma confluência da aplicação de diferentes abordagens arquivísticas, mas especificamente, das três citadas pelo autor. Todavia, considera-se que, embora distintas, podem ser entendidas como complementares e, não deixando de ser, uma amostra da construção de uma “tradição” baseada no encadeamento das correntes internacionais com o contexto e necessidades nacionais.

Por um lado, há uma forte influência da Arquivologia tradicional quando, por exemplo, valoriza-se os documentos com valor histórico e probatório e constitui-se um Arquivo Nacional. Esse Arquivo é uma instituição pública que centraliza a organização, preservação e acesso dos documentos públicos entendidos com valor para serem mantidos permanentemente, a fim de manter a memória das instituições e figuras públicas e a história do país (Arquivo Nacional, [s.d.]).

Por outro lado, observa-se a constituição de instrumentos de gestão de documentos que adotam o Ciclo de vida dos documentos e a Teoria das Três Idades, promovida pela corrente norte-americana.

Essa construção histórica das tradições arquivísticas brasileiras influencia diretamente as normas legais que estabelecem bases e regem a gestão documental atual, como a Lei 8.159/1991. É preciso reforçar que, embora possa haver críticas e contestações ao que é preconizado na Teoria da Três Idades, no cenário arquivístico brasileiro, há incentivos legais para sua prática mediante busca pela “racionalização e eficácia administrativa através da execução da gestão de documentos” (Rousseau; Couture, 1998).

Isso pode ser observado claramente na Lei de 8.159/1991, a chamada Lei de Arquivos, quando nos parágrafos de seu Art. 8º diferencia os documentos correntes, intermediários e permanentes.

§ 1º - Consideram-se documentos correntes aqueles em curso ou que, mesmo sem movimentação, constituam objeto de consultas freqüentes.

§ 2º - Consideram-se documentos intermediários aqueles que, não sendo de uso corrente nos órgãos produtores, por razões de interesse administrativo, aguardam a sua eliminação ou recolhimento para guarda permanente.

§ 3º - Consideram-se permanentes os conjuntos de documentos de valor histórico, probatório e informativo que devem ser definitivamente preservados (Brasil, 1991).

A forma de estruturar o fluxo e a instanciação⁵⁹ desses documentos nos órgãos e instituições públicas a partir da separação dessas 3 (três) fases do Ciclo vital dos documentos é nitidamente definido em termos temporais por instrumento técnico de gestão de documentos, como a Tabela de Temporalidade e Destinação de Documentos Relativos às Atividades-Meio/Suporte do Poder Executivo Federal⁶⁰.

⁵⁹ Para fins desta pesquisa, a instanciação diz respeito à definição de instâncias enquanto âmbito, esfera, setor ou área.

⁶⁰ Disponível em: [cod_classif_tab_temp_ativ_meio_atualizacao_2024.pdf](#) (www.gov.br)

Aprovada sua última atualização em 2024, pelo Arquivo Nacional, enquanto órgão central do Sistema de Gestão de Documentos e Arquivos (SIGA), essa Tabela de Temporalidade teve como uma de suas alterações os campos pertinentes aos prazos de guarda que passaram a ser identificados como “arquivo corrente” e “arquivo intermediário”.

Além disso, a compreensão da aplicação norte estadunidense no cenário arquivístico brasileiro é reforçada pelo esclarecimento de que, a partir de meados do século XX, o Arquivo Nacional⁶¹ passou a ter por embasamento teórico-metodológico o autor estadunidense Schellenberg. Essa fundamentação teórico-metodológica se dá pela aplicação de suas obras, pelos produtos gerados a partir de sua visita em 1960 para realização de cursos, palestras e a elaboração de um relatório onde autor aponta alguns problemas arquivísticos do Governo brasileiro (Marques, 2013).

A observância deste fato é relevante para compreender a influência da corrente norte-americana na adoção das práticas arquivísticas nacionais, assim como a própria construção do campo arquivístico no Brasil.

No Brasil ao contrário dos países onde a gestão de documentos se desenvolveu como teoria e prática, foram os arquivos públicos que, com vistas à sua modernização, dirigiram-se à administração pública, a partir dos anos 80, com projetos que objetivavam a adoção de componentes básicos da gestão de documentos (Jardim, 1987, *apud* Santos, 2010, p. 17).

Na década de 70, o governo federal brasileiro institui sistemas para a gestão dos arquivos de Administração Pública (como o Sistema de Serviços Gerais – SISG, aos quais estavam vinculados os arquivos correntes, e o Sistema Nacional de Arquivos – SINAR, cuja área de atuação era restrita aos arquivos intermediários e permanente), porém, não surtiu o efeito esperado (Santos, 2010, p. 112).

Em 1991, com a Lei 8.159, houve a “recriação”⁶² do SINAR, regulamentado pelos decretos nº 1.173, de 29 de junho de 1994 e 1.461, de 25 de abril de 1995, e consolidado

⁶¹ Criado em 1838, o Arquivo Nacional é o órgão central do Sistema de Gestão de Documentos e Arquivos-SIGA, da Administração Pública Federal, integrante atualmente da estrutura do Ministério da Justiça e Segurança Pública. Sua finalidade é “implementar e acompanhar a política nacional de arquivos, definida pelo Conselho Nacional de Arquivos - Conarq, por meio da gestão, do recolhimento, do tratamento técnico, da **preservação** e da **divulgação do patrimônio documental do país**, garantindo pleno acesso à informação, visando apoiar as decisões governamentais de caráter político-administrativo, o cidadão na defesa de seus direitos e de incentivar a produção de conhecimento científico e cultural” (site do AN, grifo nosso). Disponível em: <https://www.gov.br/pt-br/orgaos/arquivo-nacional>. Acesso em: 20 set. 2025.

⁶² Termo usado em <https://www.gov.br/conarq/pt-br/conexoes/sinar>. Acesso em: 20 out. 2025.

e revogado pelo Decreto nº 4.073, de 3 de janeiro de 2002. Seu objetivo passou a ser o desenvolvimento de atividades integradas com sistemas estaduais e municipais.

Em acréscimo, atualmente, tem-se também o Sistema de Gestão de Documentos e Arquivos (SIGA), criado em 2003 pelo Decreto nº 4.915, enquanto um dos sistemas estruturadores da administração pública federal previstos no Decreto-Lei nº 200/1967, com objetivo de

[...] estruturar, de forma sistêmica, a gestão de documentos e arquivos no âmbito dos órgãos e entidades da administração pública federal, tendo como principais finalidades racionalizar a produção e reduzir os custos operacionais e de armazenagem da documentação, fortalecer os serviços arquivísticos, preservar o patrimônio documental arquivístico e garantir o acesso aos documentos de forma ágil e segura (Site do SIGA [s.d.], grifo nosso).

A observância do incremento do SIGA com estruturação que integra o que seriam os dois ambientes basilares do ciclo de vida dos documentos arquivísticos (Gestão de documentos e Arquivo) permite compreender como a práxis da Teoria das Três Idades vem sendo desenvolvida no cenário brasileiro sob a influência da corrente canadense da Arquivística Integrada.

Quanto a isso, é preciso entender que

O ciclo de vida dos documentos e teoria das três idades dos arquivos são ideias distintas e envolvem processos diferentes. Além disso, as fases do ciclo de vida (produção, utilização e destinação) podem se desenvolver apenas no arquivo corrente, ou abrangendo o arquivo corrente e o intermediário (Arquivo Nacional, 2025, p. 18).

A Arquivística da corrente canadense, conforme tratado anteriormente, incorpora, segundo Tognoli, 2010, os arquivos históricos, o Princípio da proveniência e a gestão de documentos, entendendo por objeto da Arquivologia a informação registrada, e estabelece regras para a criação e manutenção de um registro pelo seu criador.

Sua implantação no Brasil é observada também pela participação de representantes do próprio Arquivo Nacional do Brasil no Projeto InterPares, coordenado por Luciana Duranti, e pela elaboração de publicações e orientações nacionais com base nos estudos desse Projeto. Esses fatos se conectam a essa percepção.

Segundo o Dicionário de Terminologia Arquivística, a gestão de documentos é o “conjunto de procedimentos e operações técnicas referentes à produção, tramitação, uso, avaliação e arquivamento de documentos em fase corrente e intermediária, visando sua

eliminação ou recolhimento” (Arquivo Nacional, 2005, p. 100). Seus principais objetivos, são:

- a) assegurar, de maneira eficiente, a produção, utilização e destinação final dos documentos;
- b) garantir que a informação esteja disponível quando e onde seja necessária;
- c) **contribuir para o acesso e conservação dos documentos de guarda permanente, por seu valor probatório, científico e histórico;**
- d) assegurar a eliminação dos documentos que não tenham valor administrativo, fiscal, legal ou para pesquisa;
- e) permitir o aproveitamento racional dos recursos humanos, materiais e financeiros⁶³ (Arquivo Nacional, 2019, p. 20, grifo nosso).

Dentre esses cinco objetivos elencados, destaca-se a contribuição da gestão de documentos para o acesso e conservação dos documentos de guarda permanente. Essa afirmação revela o quanto a gestão de documentos se correlaciona com a qualidade do que será disponibilizado futuramente no arquivo permanente. Isso evidencia a importância da integração entre o que, pela corrente americana, se compreende como “gestão de documentos” e “arquivo”.

Fica clara, portanto, a ligação entre os três arquivos: corrente, intermediário e permanente para que possa ser provida qualidade do que será preservado permanentemente. Para tanto, a adoção da Arquivística Integrada é um caminho mais alinhado.

Até porque, deve-se considerar que

Em um primeiro momento, as perdas de registros eletrônicos/digitais em suportes óticos/magnéticos se dão da mesma forma que as perdas dos registros nos documentos analógicos/convencionais. Como nos documentos convencionais, ocorrem no universo microscópico, em um processo lento, silencioso, sendo, no início, imperceptíveis. O diferencial agravante é que a preservação dos registros eletrônicos/digitais não contempla a possibilidade do acaso, pois depende de escolhas e decisões anteriores, que remetam ao momento em que a informação é produzida (Silva, 2008, p. 13).

Para tanto, no Brasil, compreende-se que a gestão de documentos, encontra-se dividida em três fases:

- Produção: na qual deve ser otimizada a produção dos documentos, com vistas a evitar a produção daqueles que não sejam essenciais, diminuindo o volume a ser manuseado, controlado, armazenado e eliminado e, assim,

⁶³ Disponível em https://www.gov.br/arquivonacional/pt-br/servicos/gestao-de-documentos/orientacao-tecnica-1/publicacoes-tecnicas-de-referencia/gestao_de_documentos.pdf. Acesso em: 10 out. 2025.

garantindo que os recursos de tecnologia da informação (reprografia e automação) sejam utilizados de forma adequada;

- Utilização: fase referente ao fluxo percorrido pelos documentos para o cumprimento de sua função administrativa, assim como de sua guarda, após cessar o trâmite; e
- Destinação: fase da qual devem participar arquivistas e gestores de documentos e que envolve as atividades de análise, seleção e fixação dos prazos de guarda dos documentos e, portanto, considerada muito importante. (Arquivo Nacional, 2025)

Na destinação, está implicada a ação de decidir quais são os documentos que devem ser conservados como testemunho do passado e quais devem ser eliminados, além de definir o tempo pelo qual devem ser mantidos por razões administrativas, legais ou fiscais (Arquivo Nacional, 2019). Aponta-se que, nessa fase, caso a destinação final seja o recolhimento, em virtude de seus valores secundários⁶⁴ (Schellenberg, 1974), essa documentação é encaminhada ao Arquivo Permanente, entidade que passa a ser responsável pela sua preservação.

Contudo, é relevante clarificar que conforme disposto no Guia de Gestão de Documentos para os Órgãos e Entidades do Poder Executivo Federal, “a preservação é uma função arquivística que se insere na gestão de documentos, e tem como objetivo a salvaguarda dos documentos que chegarão à idade permanente, após passarem pelas fases de produção, uso e destinação” (Arquivo Nacional, 2025, p. 38).

Assim, ao compreender que as teorias e abordagens arquivísticas estão entrelaçadas com a realidade e contexto do país em questão, reconhece-se que a gestão de documentos a ser utilizada nesta tese é conferida mediante uma “tradição arquivística” nacional.

Dessa forma, mesmo que haja a possibilidade de que a abordagem do *Records Continuum* embase o que seria a gestão de dados de pesquisa, ainda mais ao considerar o

⁶⁴ Schellenberg, desde 1974 já expunha o entendimento de que os documentos possuem dois valores: primários e secundários, valores estes atribuídos seguindo o uso pelo seu criador. Os primários são relativos aos valores que tem para a própria entidade onde se originam os documentos e se subdividem em valores: administrativos, fiscais, legais e executivos. Já os documentos com valores secundários “são preservados em arquivos por apresentarem valores que persistirão por muito tempo ainda depois de cessado seu uso corrente e porque os seus valores serão de interesse para outros que não os utilizadores iniciais” (Schellenberg, 1974, p. 153), e subdividem em valor evidencial e valor informativo.

contexto tecnológico atual na produção da maioria desses objetos digitais, traz-se para esta tese o entendimento de que sua efetivação necessita do estabelecimento de práticas que reflitam as teorias, normativas e legislações arquivísticas em vigor no contexto brasileiro.

A intenção aqui é costurar o que já se encontra consolidado no Brasil a fim de propor uma aplicação prática no universo dos dados de pesquisa, o que nos coloca a observar que o não desenvolvimento de estudos e tentativas empíricas para aplicação do *Records Continuum* no cenário arquivístico brasileiro (acadêmico e institucional) inviabiliza, ao menos imediatamente, que ele seja utilizado como base para essa proposição.

Essa compreensão reforça a adoção da gestão de documentos sob o ponto de vista da tradição arquivística brasileira para fundamentar a proposição da gestão de dados de pesquisa nesta tese.

5.2 A APLICAÇÃO PRÁTICA DO CONCEITO DE GESTÃO DE DOCUMENTOS NO BRASIL E SUAS CORRELAÇÕES PARA A GESTÃO DE DADOS DE PESQUISA

Como foi apresentado até aqui, novas tendências surgiram e impuseram a necessidade de revisitar fundamentos teóricos a fim de abarcar os desafios e complexidades oriundos das formas de produção documental. Quanto a isso, Couture (2015) reforça que a arquivística possui fundamentos teóricos que são eixos que podem ser revisitados para serem refinados, enriquecidos e melhorados, mas que é preciso distinguir evolução de revolução, nos lembrando que reexaminar não significa rejeitar, assim como visitar também não é demolir (Couture, 2015).

Nesse cenário, embora a informatização das administrações públicas e a ampliação do acesso à internet tenham transformado o ambiente informacional, tais mudanças não invalidam os fundamentos teóricos da Arquivologia. Pelo contrário, esses princípios vêm sendo desafiados, o que tem exigido sua constante revisão, refinamento e adaptação às novas realidades.

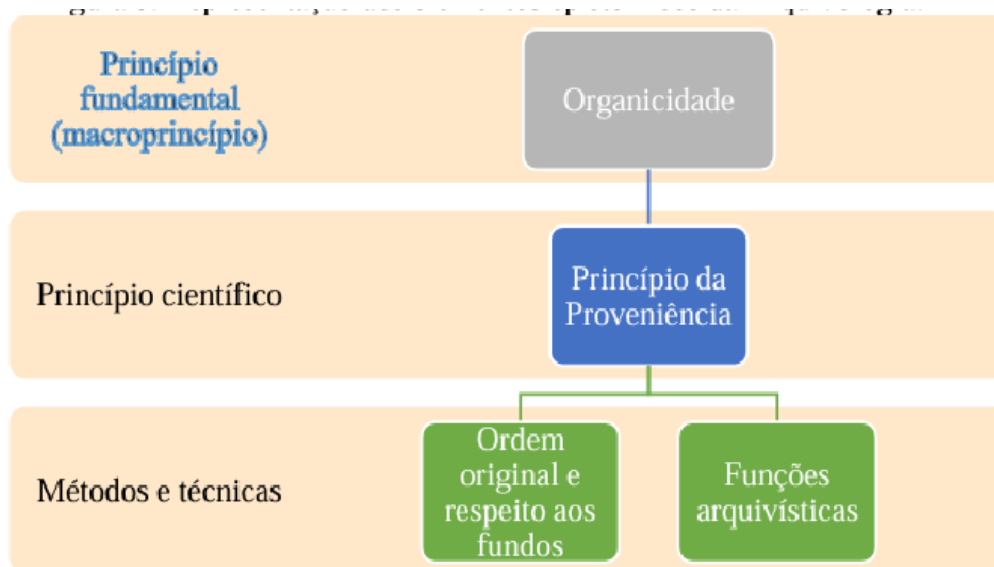
Um exemplo disso é a retomada do Princípio da proveniência, motivada pela necessidade de garantir a autenticidade e a integridade dos documentos em meio digital (Dollar, 1993). Assim, como observa Couture (2015), é fundamental que qualquer

atualização desses fundamentos seja conduzida com cautela e respeito às bases que historicamente conferem credibilidade e solidez à disciplina e à profissão arquivística.

Segundo o estudo apresentado por Ívina Flores Melo, em sua tese de doutorado, há um consenso na literatura internacional sobre o Princípio da proveniência ser o princípio fundador da Arquivologia, sendo o condutor na delimitação de seus métodos e técnicas. Sendo assim, a autora enriquece esse entendimento ao concluir que a organicidade, portanto, é coração da Arquivologia, o que a difere de todas as demais disciplinas nos campos científicos de Bourdieu (1983), considerando, então, se tratar de seu macroprincípio.

A correlação defendida por Melo, desde sua dissertação⁶⁵ (2016) é a seguinte:

Figura 23 – Representação dos elementos epistêmicos da Arquivologia



Fonte: Melo (2021, p. 36).

A Figura 23 demonstra que o Princípio da proveniência se desdobra da Organicidade. Já os demais princípios, como o de “respeito à ordem original” e o “de respeito aos fundos”, são métodos que se desdobram do Princípio da proveniência. Da mesma forma as funções arquivísticas são técnicas que operacionalizam esse Princípio (Melo, 2021).

⁶⁵ Dissertação de mestrado defendida na UNB, intitulada “Demarcações conceituais dos princípios científicos da Arquivologia e da Ciência da Informação: contribuições para a configuração científica das disciplinas no Campo da Informação”.

No contexto atual, ao considerar a aplicação dos princípios arquivísticos no ambiente digital, destaca-se a ampliação do conceito de proveniência, agora também reconhecido sob a forma de “proveniência digital”, conforme exposto na subseção 4.2.1 desta tese. Vale retomar que, segundo consta no *PREMIS Data Dictionary* (2015), a proveniência digital é a documentação de todos os processos do ciclo de vida de um objeto digital, a partir da descrição dos agentes responsáveis pela custódia e administração dos objetos digitais, os eventos-chave que ocorrem ao longo desse ciclo de vida e outras informações relativas à criação, gestão e preservação desse objeto.

Trata-se, portanto, de um registro detalhado, contínuo e cumulativo, da origem, do contexto de criação, das transformações e percursos pelos quais um objeto digital passa ao longo de seu ciclo de vida. O conceito de proveniência no ambiente digital é ampliado para abranger não apenas a identificação do produtor, mas também o acompanhamento de todas as ações, processos e agentes envolvidos na criação, manipulação, custódia, curadoria e administração do documento digital, por meio do registro de metadados que documentam eventos-chave como alterações, transferências de custódia e preservação (Rabelo; Schmidt, 2025).

Segundo o *World Wide Web Consortium* (W3C), a proveniência digital pode ser analisada sob três perspectivas complementares: o agente (quem criou ou modificou o documento), o recurso (a origem e o histórico do próprio documento) e o processo (as etapas e procedimentos realizados sobre o documento).

Essa abordagem é essencial para a preservação digital pois permite reconstruir a história do documento, verificar sua autenticidade e garantir que ele permaneça confiável e acessível em longo prazo (Rabelo; Schmidt, 2025). Assim, a proveniência digital constitui elemento central para a gestão de documentos digitais, assegurando a rastreabilidade, autenticidade e integridade dos registros, requisitos indispensáveis para a preservação e o uso confiável da informação digital ao longo do tempo (Rodrigues, 2008)

É importante destacar que, no âmbito arquivístico, considera-se que “o documento só é de arquivo se os vínculos de proveniência e organicidade se preservarem autênticos em seu conteúdo” (Rodrigues, 2008, p. 43). Isso significa que o Princípio da proveniência, mais do que dizer de onde o documento vem, possui um aspecto intrínseco, a configuração da relação orgânica.

A Arquivologia ao longo de sua trajetória passou por revisitações, aprimoramentos e adequações em suas teorias e práticas, principalmente ao se deparar

com novos marcos, que passaram a compor sua linha do tempo, de uma forma geral, em âmbito mundial ou regional. Contudo seu fundamento basilar, a organicidade, se mantém em vigor e, por isso é identificado como relevante para embasar esta tese.

Para tanto, inicia-se esse desdobramento reconhecendo que a atividade de pesquisa origina diversas informações sob as mais variadas possibilidades de registros documentais. Para além da produção de dados de pesquisa, objeto nuclear desta tese, que tratam, em sua maioria, de informações inerentes à execução e à análise da pesquisa, a atividade científica produz, recebe e acumula informações registradas sobre outros momentos da pesquisa. Quanto a isso observa-se a produção de documentos de planejamento (por ex: PGD, projetos de pesquisa, convênios e contratos, orçamento e cronograma de execução, submissões a comitês de ética, termos de consentimento) e de comunicação (por ex: artigos, teses, dissertações, apresentações, relatórios finais).

Figura 24 – Outras informações registradas na atividade de pesquisa



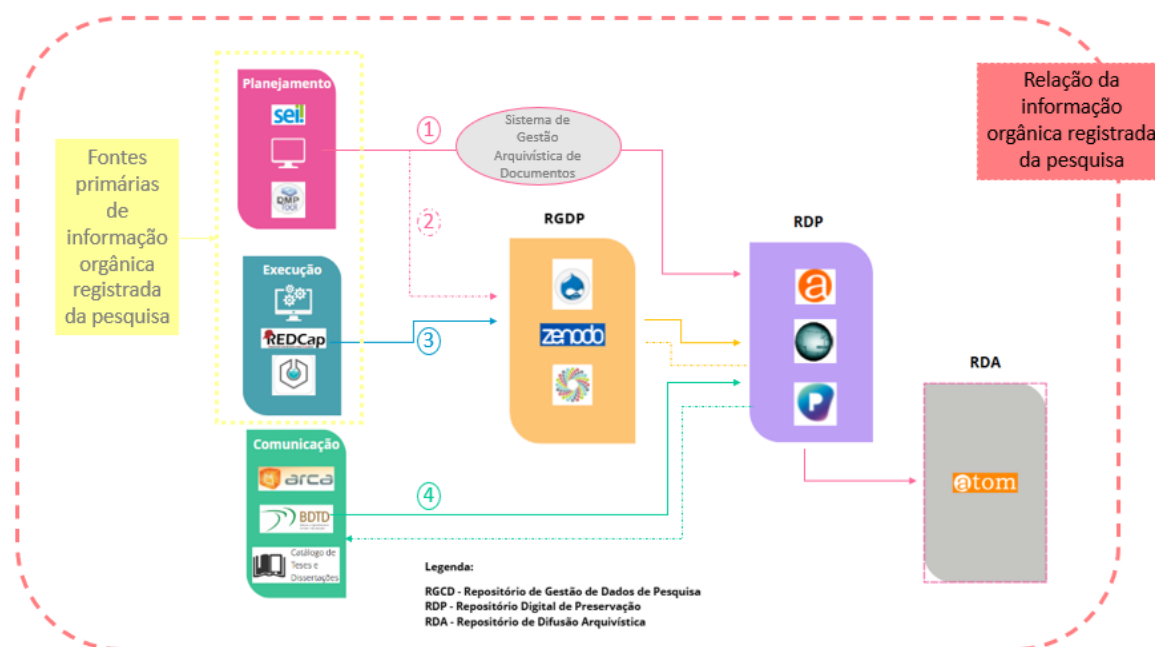
Fonte: Elaborada pela autora (2025).

Cada um desses documentos, produzidos no curso da atividade de pesquisa, constitui sua organicidade, ou seja, é um elo na construção da relação orgânica estabelecida na atividade de pesquisa.

No entanto, é preciso esclarecer que, atualmente, esses documentos, embora façam parte de um todo, são produzidos e/ou mantidos em sistemas de negócio específicos. Dessa maneira, encontram-se pulverizados em diferentes *softwares* ou aplicativos de acordo com seus fins e, conseqüentemente, os fluxos de seu ciclo de vida.

É possível encontrar, por exemplo, documentos administrativos, que apoiam o planejamento e mediam a realização da pesquisa, em sistemas de negócios que objetivam produzir, gerenciar e dinamizar trâmites processuais. Enquanto documentos que concretizam a comunicação da pesquisa podem ser encontrados em repositórios institucionais que objetivam a abertura e difusão do conhecimento.

Figura 25 - Relação da informação orgânica registrada da pesquisa



Fonte: Elaborada pela autora (2025).

Na Figura 25⁶⁶ são exemplificados 4 (quatro) possíveis fluxos para as informações registradas ao longo do percurso da pesquisa, ao qual opta-se por denominar de acordo com Couture, Ducharme e Rousseau (1988), como “informação orgânica registrada” (tratado anteriormente na subseção 5.1.1) e, portanto, escopo abrangido pela Arquivologia.

No **fluxo 1**, tem-se a produção de documentos que apoiam a realização da pesquisa, com um cunho mais gerencial e administrativo, com o uso, por exemplo, do Sistema Eletrônico de Informação (SEI)⁶⁷ para produção e gestão de processos

⁶⁶ É importante esclarecer que esses fluxos não são realizados de maneira automática, devendo, portanto, contar com a padronização de atividades e uso de recursos que os efetive.

⁶⁷ Mais informações em <https://www.gov.br/servicoscompartilhados/pt-br/assuntos/gestao-documental/sistema-eletronico-de-informacoes-sei>

administrativos, o DMP Tool⁶⁸ para a elaboração de PGD e outros programas, aplicativos específicos, como Word, Excel, por exemplo, que sejam utilizados nos computadores institucionais ou dos próprios pesquisadores para esse fim.

Esses documentos, seguindo uma Cadeia de preservação (conforme tratado na subseção 2.3.2) passam, ou ao menos recomenda-se que passem, por um Sistema de Gestão Arquivística de Documentos (SIGAD)⁶⁹ para tratamento técnico a partir do uso de funcionalidades que possibilitem a aplicação das funções arquivísticas para, então, em tempo previsto em Tabela de Temporalidade e Destinação (o que será melhor exposto mais adiante, na subseção 7.2) serem eliminados ou preservados permanentemente em Repositório de preservação digital, aderente ao Modelo OAIS (2024), conforme tratado na subseção 3.2 (como o Archivematica⁷⁰, RODA⁷¹, Preservica⁷², por exemplo). Esse RDP preserva os AIP dos objetos digitais admitidos e pode gerar DIP para que seu acesso seja franqueado em um RDA como os que utilizam Atom⁷³, por exemplo.

O **fluxo 2** ainda diz respeito a informações oriundas de sistemas de negócio administrativos. No entanto, pelo fato de poderem ser utilizadas para contextualização dos dados de pesquisa disponibilizados em RGDP (como, por exemplo, o Dataverse Project, Zenodo⁷⁴ e Figshare⁷⁵), podem também integrar os conjuntos de dados aos quais fazem parte⁷⁶. Ou seja, se encaixam como administrativos e/ou contextuais.

Tem-se, por exemplo, o Plano de Gestão de Dados (PGD). Primeiramente, ele é produzido como um documento administrativo que dá início ao processo de financiamento. Ao passo que também pode passar por versionamentos ao longo da pesquisa, já que precisa ser atualizado à medida que a execução do processo científico apresenta novas demandas e realidades.

⁶⁸ Mais informações em <https://dmptool.org/>

⁶⁹ De acordo com o Modelo de requisitos do e-Arq (2022, p. 20, grifo nosso), um SIGAD “é uma solução informatizada que visa o **controle do ciclo de vida** dos documentos, desde a produção até a destinação final, **seguindo os princípios da gestão arquivística de documentos**. Pode compreender um software particular ou um determinado número de softwares integrados, adquiridos ou desenvolvidos por encomenda. **Um SIGAD tem que ser capaz de manter a relação orgânica** entre os documentos e de **garantir a confiabilidade, a autenticidade e o acesso**, ao longo do tempo, aos documentos arquivísticos, ou seja, seu **valor como fonte de prova das atividades do órgão produtor**”.

⁷⁰ Mais informações em <https://www.archivematica.org/pt-br/>

⁷¹ Mais informações em <https://dglab.gov.pt/roda-repositorio-objectos-digitais-autenticos/>

⁷² Mais informações em <https://preservica.com/>

⁷³ Mais informações em <https://www.accesstomemory.org/pt-br/>

⁷⁴ Mais informações em <https://zenodo.org/>

⁷⁵ Mais informações em <https://figshare.com/>

⁷⁶ Cabe apontar que, inclusive, essa análise fez parte do levantamento empírico realizado, culminando na recomendação que esses documentos fossem compartilhados dessa maneira para fortalecer o grau de transparência e maturidade da gestão dos dados de pesquisa, conforme subseção 5.3.1.

Assim, O PGD inicial, fixo, pode ser encontrado junto aos demais documentos administrativos do contexto da pesquisa, no fluxo 1. Todavia, esse PGD inicial, junto a todas as suas versões, também deve compor as Informações Adicionais dos dados de pesquisa a serem preservados, pelo fato de fornecer informações que enriquecem a contextualização da pesquisa.

Ao serem submetidos a esse repositório, devem ser permanentemente preservados em um RDP que siga os requisitos demonstrados na seção primária 4 desta tese. Importante indicar que o RDP apoia a sua preservação, mas o acesso aos dados de pesquisa se dá por intermédio do próprio RGDP.

Adiante, tem-se o **fluxo 3**, onde dados de pesquisa são produzidos em *softwares* ou aplicativos executados em sistemas ou aplicativos, os quais viabilizam o registro das informações e sua materialização em formatos de arquivo. São vários os sistemas de negócio para tal, como por exemplo o Libre Office⁷⁷, o RedCap⁷⁸ e o eLabFTW⁷⁹, os quais produzem dados de pesquisa que podem vir a ser compartilhados em Repositórios de gestão de dados.

Assim como no fluxo 2, todas informações submetidas a esse Repositório, devem ser admitidos, em algum momento a ser estabelecido, no RDP. Nesse caso também, o RDP apoia a sua preservação, mas o acesso aos dados se dá por intermédio do próprio RGDP.

Já o **fluxo 4** se “origina” nos Repositórios Institucionais como, por exemplo, o Arca da Fiocruz⁸⁰, a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) do IBICT⁸¹ e o Catálogo de Teses e Dissertações da CAPES⁸². Esses repositórios tem por finalidade disponibilizar os produtos finais da pesquisa como teses, dissertações, artigos, relatórios finais, apresentações, entre outros, para ampliar seu acesso.

Por sua vez, esses repositórios devem ser integrados a um RDP para garantir que seus documentos sejam preservados a longo prazo, assim como ocorre com o RGDP. Da mesma forma, o RDP apoia a sua preservação, mas o acesso aos dados se dá por intermédio do próprio Repositório Institucional.

⁷⁷ Mais informações em <https://pt-br.libreoffice.org/>

⁷⁸ Mais informações em <https://project-redcap.org/>

⁷⁹ Mais informações em <https://www.elabftw.net/>

⁸⁰ Mais informações em <https://arca.fiocruz.br/>

⁸¹ Mais informações em <https://bdtd.ibict.br/vufind/>

⁸² Mais informações em <https://catalogodeteses.capes.gov.br/catalogo-teses/#!/>

Quanto a esses fluxos, compreende-se, que a organicidade deve ser expressa mesmo que os documentos estejam “fisicamente” separados, o que amplifica e aprofunda a proveniência sob a perspectiva tradicional. Essa proveniência “intelectual” destaca que, além das relações orgânicas, deve-se considerar a “verdadeira” proveniência, correspondente ao contexto de criação (Melo, 2021).

Logo, é possível entender que a análise da organicidade “integral” correspondente ao contexto sob o qual os dados de pesquisa se encontram abarca a totalidade desses fluxos. Todavia, para fins desta pesquisa, a intenção é tratar, especificamente, do ciclo de vida dos dados de pesquisa sob a ótica do **fluxo 3**, o que, mesmo sendo uma parte do fluxograma exposto anteriormente, demanda o registro adequado de sua proveniência para o estabelecimento das relações orgânicas entre eles e entre o conjunto integral das informações registradas sobre essa atividade.

Conforme discutido no artigo “Dados de pesquisa na Arquivologia: uma reflexão” (Gava *et al.* 2023), os dados de pesquisa devem ser compreendidos como documentos arquivísticos considerando que

As instituições também começaram a **produzir cada vez mais dados, no decorrer de suas atividades, onde destacam-se as atividades de pesquisa, uma vez que os dados são produzidos no decorrer de atividades institucionais de pesquisa**, muitas vezes com financiamento público, é importante que a Arquivologia olhe para tais conjuntos de dados como um tipo de documento arquivístico, que precisa ser inserido no contexto das políticas de gestão, preservação e acesso (Gava *et al.*, 2023, conclusões, grifo nosso).

Portanto, conforme expõem os autores, exigem tratamento específico para garantir sua autenticidade, integridade e reusabilidade ao longo do tempo, assegurando sua rastreabilidade e o controle documental desde sua criação, permitindo o reconhecimento do contexto institucional e das atividades que os originaram.

Dessa forma, para demonstrar como os fundamentos arquivísticos podem ser integrados ao contexto atual da gestão de dados de pesquisa, essa subseção traz as possibilidades de aplicação prática de funções arquivísticas, mais especificamente aquelas praticadas na gestão de documentos no Brasil.

Para isso, considerando que não há um consenso estabelecido na literatura da área sobre quais são essas funções arquivísticas, optou-se por escolher a classificação, a avaliação e a descrição, complementando esse rol com a técnica de identificação. Essa

técnica que, embora aqui esteja exposta posteriormente às funções, deve ser considerada na prática como precedente a elas.

Essa escolha se deu de maneira a fluidificar a interligação de três das funções arquivísticas clássicas de Jean-Yves Rousseau e Carol Couture (1998), a partir da visão de Melo (2021) que, ao entendê-las como “técnicas, ou seja, a manifestação e a operacionalização do Princípio da Proveniência, espelho da organicidade” (p.166) em sua análise sobre o *habitus* da Arquivologia, seleciona as funções: classificação, avaliação e descrição. Para tanto, julga-se oportuno trazê-las à luz desta tese e integrá-las à identificação de Rodrigues (2013), em virtude da relevância técnica dessa atividade para que a gestão de documentos seja adequadamente iniciada.

Importante ressaltar, para tanto, a relevância do acompanhamento especializado para efetivação dessa gestão, visto que trata-se de uma atividade que, em partes, ou seja, nos estágios iniciais, deve ser conduzida ainda no ambiente do produtor. Agente este que, muitas vezes, não é conhecedor das nuances pragmáticas e da relevância dessa ação. Assim, essa questão também é abarcada ao final da subseção 5.2.

5.2.1 As funções arquivísticas na gestão de dados de pesquisa: Classificação, avaliação e descrição

As funções arquivísticas básicas desempenham papel central na gestão de documentos e são amplamente reconhecidas na literatura arquivística como essenciais para a organização, controle e preservação dos documentos. De forma geral, os arquivistas canadenses Jean-Yves Rousseau e Carol Couture (1998) apontam sete funções arquivísticas: produção, avaliação, aquisição, conservação, classificação, descrição e difusão.

No entanto, Melo (2021), a partir do Princípio da proveniência, que considera o vínculo orgânico e a origem dos documentos como fundamentos epistemológicos centrais da Arquivologia, analisa enquanto funções arquivísticas clássicas a classificação, a avaliação e a descrição. Para a autora, essas funções são técnicas que derivam desse

princípio, configurando-se como práticas institucionais que refletem o *habitus*⁸³ da disciplina.

Cabe destacar que as funções apresentadas a seguir são interdependentes, complementares e não estáticas. Elas formam o núcleo da prática arquivística, sustentam a gestão documental integrada e evoluem conforme o contexto histórico, social e tecnológico, sempre orientadas pelo Princípio da proveniência.

5.2.1.1 A Classificação arquivística

Segundo Marilena Leite Paes (2004), a classificação de documentos é um processo que ocorre conforme as características, forma e conteúdo próprios dos documentos e pode ser realizada segundo o gênero e a natureza do assunto. Para a autora, a classificação é essencialmente lógica pois parte da análise do organismo produtor dos documentos para criar categorias e classes genéricas que refletem as funções e atividades dessa instituição ou órgão. Essas categorias são organizadas hierarquicamente em classes, subclasses e séries tipológicas, que agrupam os documentos conforme as atividades que os geram. Sua aplicação facilita a compreensão do contexto de produção e a recuperação da informação arquivística.

Adicionalmente, Paes (2004) destaca que a classificação deve respeitar os princípios arquivísticos fundamentais, como o Princípio da proveniência, que assegura que os documentos sejam organizados segundo a competência e as funções do produtor. Isso preserva sua individualidade e evita a mistura com documentos de outras origens. A autora, portanto, a considera uma função central para a gestão de documentos pelo fato de subsidiar outras atividades arquivísticas, como a avaliação. A classificação garante que os arquivos sejam organizados de forma coerente e funcional, e atendam às necessidades administrativas e históricas da instituição produtora (Paes, 2004).

Em complemento, Gonçalves (1998) expõe que a classificação visa estabelecer um sistema predeterminado de arquivamento para as séries documentais, o que propicia visibilidade às funções institucionais e facilita a recuperação dos documentos. Para isso, o Arquivo Nacional (2001) reforça que essa classificação deve ser realizada por

⁸³ Melo (2021) especifica esse *habitus* enquanto o conjunto de disposições e saberes que definem a identidade e o funcionamento da Arquivologia enquanto campo científico.

profissionais qualificados, por meio do estudo detalhado dos documentos e da atribuição de códigos que refletem a hierarquia funcional do órgão produtor.

Barros e Moraes (2013, p. 79) entendem que “a classificação é uma atividade gerencial e de planejamento não só para os arquivos, mas para as administrações que geraram os documentos” e que deve contemplar

[...] as relações entre os documentos, mas não só essas, também, as relações entre os documentos e as pessoas, assim, o sistema de classificação se tornaria uma ferramenta útil e facilitadora da descrição e avaliação dos documentos (Barros; Moraes, 2013, p. 79).

Ao reforçar a ligação entre as atividades de classificação, avaliação e descrição, os autores revelam o papel fundamental da classificação arquivística: contribuir para a atividade de avaliação e facilitar o acesso aos documentos (Schellenberg, 1974).

No Brasil, o sistema de classificação mais utilizado é o funcional, que reflete as funções e atividades do órgão. Todavia, ainda é possível encontrar alguns sistemas de classificação que utilizam a lógica estrutural, ou seja, que reflete a estrutura hierárquica interna dos órgãos (Barros; Moraes, 2013).

Ao integrar essa função arquivística, que faz parte da gestão de documentos, ao contexto da gestão dos dados de pesquisa, é possível considerar que

Na Gestão de dados, assim como na Arquivologia, uma classificação malfeita (atribuição de metadados inadequados) pode influenciar diretamente a recuperação e a compreensão dos dados, prejudicando a descoberta e dificultando (ou mesmo impedindo) o seu reuso, um dos principais objetivos da Gestão de Dados de Pesquisa (Corrêa; De Souza, 2022, p. 447).

Isso demonstra que, não obstante ao fato de que, até o presente momento, a classificação de dados de pesquisa estar se consolidando por parâmetros distintos dos fundamentados pela Arquivologia, essa atividade, por propiciar a recuperação e a compreensão desses dados, é relevante para efetivação do (re)uso.

Segundo Sayão e Sales (2019), as classificações dos dados de pesquisa refletem o propósito para o qual foram construídas e um contexto específico. Assim, os autores, ao elaborarem uma taxonomia para categorização dos tipos de dados de pesquisa adotam oito características divididas quanto à/ao: origem, grau de processamento, natureza, abordagem da pesquisa, nível de sensibilidade, materialidade, perenidade e abertura.

Quanto à origem, Sayão e Sales (2019) destacam que esse é o princípio mais relevante, pois define a profundidade e abrangência do tratamento necessário. “A

diferença de criticidade conferida pela possibilidade, ou impossibilidade, de replicação desses dados implicarão adoção de estratégias distintas de gestão e curadoria” (Sayão; Sales, 2020, p. 32). Mediante análise da origem, os dados de pesquisa podem ser:

- Dados observacionais quando são únicos e não podem ser coletados novamente. Portanto, devem ser arquivados para sempre com fidedignidade, autenticidade e integridade em repositórios confiáveis.
- Dados experimentais que, mesmo provenientes de situações controladas que viabilizem sua reprodução, nem sempre é possível refazer o experimento. Além disso, geralmente, o custo de geri-los é menor do que sua reprodução.
- Dados computacionais, gerados por modelos e simulações, requerem arquivamento com metadados detalhados, incluindo *hardware*, *software* e parâmetros de entrada.

Ao analisá-los sob uma visão arquivística, os dados observacionais podem ser entendidos logo em seu planejamento, como dados permanentes que devem ser preservados ao longo de seu ciclo até que sejam admitidos em um Repositório Digital de Preservação. Já os dados experimentais e os computacionais precisam passar por uma avaliação que reconheça caso a caso.

Quanto ao grau de processamento, os autores (2019) expõem que os dados raramente são usados diretamente após a coleta, necessitando de diferentes níveis de processamento conforme o fluxo da pesquisa.

Já a abordagem da pesquisa influencia o tratamento, especialmente para dados qualitativos de ciências sociais, humanidades, arte e cultura, que exigem infraestruturas e práticas distintas. Enquanto a natureza dos dados revela grande diversidade em formatos, mídias, suportes e expressões (Sayão; Sales, 2019).

O nível de sensibilidade define os níveis de abertura e compartilhamento, exigindo tratamentos adequados para dados pessoais e confidenciais, como anonimização e armazenamento seguro (Sayão; Sales, 2019).

A materialidade diferencia dados digitais, frágeis e sujeitos à obsolescência, dos dados físicos, como amostras biológicas e mapas, que demandam cuidados específicos, mas também dependem de metadados para recuperação (Sayão; Sales, 2019).

Quanto à perenidade, as coleções são gerenciadas segundo sua relevância científica, desde coleções internacionais permanentes até coleções institucionais e comunitárias com valores transitórios. A abertura dos dados é definida por critérios como sensibilidade, propriedade intelectual e interesses comerciais, sendo regulada por licenças que determinam o grau de (re)uso (Sayão; Sales, 2019).

Por fim, os autores defendem que a gestão dos dados deve preceder a tecnologia e o compartilhamento, incluindo o mapeamento dos tipos de dados, padrões de metadados e arcabouços tecnológicos necessários para uma curadoria adequada e recuperação precisa. Adicionalmente, “a ausência de procedimentos de classificação afetará diretamente nos procedimentos simultâneos e posteriores como avaliação, descrição, preservação e acesso” (Santos; Flores, 2016, p. 171).

De maneira geral, os dados de pesquisa podem ser classificados sob diversos vieses, conforme demonstrado acima. Contudo, mesmo com especificidades a serem consideradas nesse contexto, no sentido da classificação estar atrelada à observação das características do objeto em si, é possível identificar que esses dados são fruto de atividades e funções da instituição. Além disso, para fins de embasamento da temporalidade de seu ciclo de vida, as categorias trazidas de Sayão e Sales (2019) são capazes de apoiar sua avaliação.

5.2.1.2 A Avaliação arquivística

Ao considerar que o foco das ações preservacionistas tem sido cada vez mais voltadas para a prevenção planejada e integrada (Nascimento, 2012), é importante que sejam empregadas durante todo o ciclo de vida. Caso contrário, ao serem iniciadas somente depois de anos de possível desatenção, pode ser que não haja tempo e recursos hábeis para neutralizar os efeitos degenerativos ou reparar os danos já sofridos (Praxedes, 2015), seja no objeto em si ou na sua capacidade de ser (re)usado.

No contexto da gestão de dados de pesquisa, portanto, compreende-se que uma adequada gestão é o que pode viabilizar o suporte necessário para incentivar de forma consciente e responsável todo investimento destinado à preservação dos dados de pesquisa.

Cabe destacar que

Para alguns tipos de dados, a administração requer preservação indefinida; para outros, a administração requer ciclos regulares de descarte de registros. No entanto, dada a natureza dinâmica dessas coletas de dados, é improvável que as abordagens arquivísticas tradicionais para manter o acesso a recursos estáticos sejam suficientes. Em uma "era de algoritmos" onde os conjuntos de dados estão em fluxo constante e podem ser desagregados e reagregados continuamente para múltiplos propósitos analíticos, novas abordagens são extremamente necessárias (Borgman, 2017).

São identificados, então, três fatores: 1) a complexidade da preservação digital sob os mais diversos espectros, 2) a importância de uma abordagem multidisciplinar e 3) a compreensão de que trata-se de um empreendimento caro, cujos custos precisam ser reconhecidos e apoiados de maneira consciente e responsável pelas instituições que produzem e coletam dados de pesquisa (Borgman, 2018).

Tal contexto implica na impossibilidade, e até mesmo dispensabilidade, de se preservar tudo o que é produzido e, portanto, a execução da preservação digital deve ser fruto de escolhas e decisões entre o que espera-se manter e o que pode, ou poderá, ser descartado.

Essa análise é feita levando-se em consideração o processo de avaliação, que é seguido da seleção do(s) conjunto(s) que será(ão) objeto da preservação. Segundo Oliveira (2012), esses processos devem ser criteriosos e cunhados por embasamentos técnico-científicos ou legais, e não pelo argumento da necessidade de espaço, tendo em vista também a possibilidade de uso pela sociedade.

A função de avaliar os documentos consiste na análise crítica de seus valores para determinar sua destinação final. Segundo Rousseau e Couture (1998), a avaliação é uma atividade fundamental para a eficiência da gestão documental e para a preservação do patrimônio arquivístico e deve considerar os aspectos administrativos, legais, históricos e informativos dos documentos.

Alguns dados possuem baixo potencial para futuros interesses ou não têm valor a longo prazo e, mesmo que tenham um ou outro, podem não ter informações de apoio suficientes para permitir uma análise secundária (Gutmann *et al.*, 2004).

Dessa forma, com base na metodologia arquivística, entende-se a importância da avaliação para assegurar que os dados de pesquisa relevantes sejam preservados, enquanto os dispensáveis sejam descartados de forma controlada.

Conforme expõem Sá e Araújo (2024)⁸⁴

⁸⁴ Em processo de aprovação.

Realizar o arquivamento de dados de pesquisa sem informações sobre contexto, formas de coleta ou produção, se houve manipulação etc., provavelmente apenas ocupará espaço no repositório, sem ter novos usos (Sá; Araújo, 2024, p.15).

A visão das autoras aponta para uma avaliação que vai além da análise das características do dado de pesquisa em si. Elas demonstram a importância de que essa avaliação seja balizada também a partir da identificação, ou não, de informações sobre o que nesta tese está sendo tratada como sua proveniência digital.

Jardim (2012) expõe que a quantidade, variedade de suportes e formatos documentais, de um lado, e a escassez de recursos arquivísticos, de outro, implicam que os documentos produzidos e acumulados por uma organização devam, inevitavelmente, passar por um processo de avaliação. Essa atividade, portanto, segundo o autor, deve ser norteada pelos valores “primário” e “secundário”.

O “valor primário” se refere aos aspectos gerenciais do documento e à demanda de uso que este recebe por conta da administração que o produziu (considera suas dimensões gerenciais, fiscais, legais e financeiras). Já o “valor secundário”, se refere às possibilidades de utilização do documento por usuários que o procuram por razões distintas e posteriores àquelas do seu produtor (considera seu valor probatório e informativo) (Jardim, 2012).

Essa mesma ideia de atribuição de valores também está presente no ciclo de vida dos dados de pesquisa. Tem-se, por exemplo, o caso dos dados sísmicos. Ao serem recolhidos por instrumental científico geram experimentos originais e a eles pode-se atribuir o valor primário. Porém, ao serem reprocessados e integrados pela engenharia para fundamentar a definição de fundações de uma usina nuclear ou de uma barragem, desponta-se um valor posteriormente atribuído a esses dados. Em virtude desse conjunto de dados não ter sido gerado para essas aplicações, identifica-se o valor secundário.

Assim, compreende-se que a implementação de critérios de avaliação contribui tanto na gestão dos dados de pesquisa quanto na sua preservação. A definição de prazos de guarda e a seleção criteriosa dos dados de pesquisa que devem ser preservados indeterminadamente, distinguindo-os dos passíveis de eliminação viabilizam transparência nos processos de gestão e contribuem para sua confiança.

No entanto,

Arquivos para diferentes tipos de dados (por exemplo, observacionais versus experimentais, ciências físicas versus ciências biológicas, seres humanos ou não) podem ter procedimentos e métricas díspares para a qualidade dos dados e critérios diferentes para avaliar o valor e selecionar dados para preservação (Esanu *et al.*, 2004, p. 227).

Isso indica a relevância da avaliação criteriosa baseada nas especificidades dos dados de pesquisa e seus contextos. Quanto a essa demanda, é oportuno indicar a expertise em âmbito nacional que vem sendo consolidada pelas arquivistas Ivone Sá e Renata Araújo no âmbito da Fiocruz. As autoras elaboraram orientações gerais para seleção, avaliação e retenção de dados de pesquisa naquela Fundação, a partir de um estudo com mapeamento seguido de análises sobre os critérios e procedimentos que vem sendo aplicados em instituições internacionais.

A partir de uma vasta coleta de referências, Sá e Araújo (2024) listam os seguintes critérios para embasar a seleção e a avaliação dos dados de pesquisa na Fiocruz:

- a) Requisitos legais, éticos ou outras razões regulatórias;
- b) Requisitos do instituto, agência de fomento ou editor;
- c) Importância dos dados;
- d) Unidade dos dados;
- e) Valor dos dados;
- f) Significativo interesse público;
- g) Questões econômicas e financeiras;
- h) Singularidade;
- i) Potencial para reuso; e
- j) Critérios de seleção secundários.

Tais critérios podem ser replicados para outras instituições, conforme exposto logo no início do próprio documento. Além disso, essa avaliação pode ser realizada ainda na fase de planejamento da pesquisa. O Plano de Gestão de Dados (PGD) pode ser utilizado como subsídio para verificação de, ao menos, algumas informações desse rol.

A avaliação dos dados de pesquisa antes do seu depósito em repositórios de gestão (RGDP) configura-se como uma etapa estratégica fundamental para garantir a efetividade e a sustentabilidade⁸⁵ da curadoria digital de dados de pesquisa ao longo do seu ciclo de

⁸⁵ No programa Tecnopolítica, no episódio #255, onde o professor Sérgio Amadeu conversa com os pesquisadores do LAPIN (Laboratório de Políticas Públicas e Internet): Camila Cristinada Silva e Marcelo Aparecido de Faria Júnior, o tema “Data centers em tensão com a justiça social”, que aborda os impactos

vida. Essa avaliação prévia permite definir critérios claros de seleção e retenção, alinhados aos conceitos de valor primário e secundário.

Essa ação, quando antecipadamente executada, assegura o compartilhamento responsável dos dados de pesquisa, com identificação da intenção ou não de abrí-los e por quanto tempo, além de possibilitar o encaminhamento posterior ao Repositório Digital de Preservação daqueles considerados relevantes, autênticos e com potencial de (re)uso.

Dessa forma, evita-se o acúmulo indiscriminado de dados (não necessariamente dados de pesquisa propriamente ditos), que além de sobrecarregar os sistemas, dificulta a recuperação e consome recursos institucionais desnecessariamente. Além disso, a avaliação antecipada dos dados de pesquisa possibilita o atendimento rigoroso a requisitos legais, éticos e institucionais desde o início do processo de curadoria. Esse fato é essencial para prevenir riscos relacionados à privacidade, confidencialidade e conformidade regulatória, sobretudo em pesquisas que envolvem informações sensíveis.

Ao apoiar a elaboração de Planos de Gestão de Dados (PGD) mais precisos e aderentes à realidade da pesquisa, a avaliação prévia facilita a tomada de decisões sobre prazos de guarda, eliminação e preservação permanente, o que promove transparência e rastreabilidade na gestão dos dados de pesquisa. Outro aspecto relevante é que a avaliação prévia permite que diferentes tipos de dados recebam tratamentos específicos conforme suas características e potencialidades. Isso assegura que critérios de qualidade, valor científico e possibilidade de (re)uso sejam considerados de forma adequada, conforme destacado por Esanu *et al.* (2004).

A experiência da Fiocruz em sistematizar critérios para avaliação e seleção de dados de pesquisa, demonstra que a possibilidade de que tal metodologia embasada por procedimentos claros e fundamentados possa ser adotada antes do depósito e que essa ação pode ser replicada em outras instituições.

Em síntese, a avaliação dos dados de pesquisa antes do depósito em repositórios de gestão (RGDP) contribui na promoção de uma curadoria mais eficiente, ética e

socioambientais da instalação de data centers, a ausência de regulação ambiental específica, as operações de *greenwashing* e o gasto massivo de energia e água no treinamento e inferência da Inteligência Artificial. Essa visão pode ser utilizada para apoiar a demonstração da necessidade de que sejam discutidas as questões sobre a sustentabilidade da ciência e, nesse sentido, reforçar a importância de que avaliações sejam realizadas no intuito, também, de possibilitar a minimização de gastos em gestão e preservação de dados de pesquisa passíveis de serem eliminados.

alinhada às melhores práticas internacionais. Por isso, trata-se de uma condição indispensável para que a preservação seja realizada de modo eficaz, sustentável e orientada ao interesse público, à Ciência Aberta e à preservação do patrimônio científico.

Sendo assim, essa prática garante que os dados selecionados para preservação atendam aos critérios necessários para sua autenticidade, integridade e valor probatório, o que contribui para a construção de um acervo científico confiável e acessível a longo prazo.

Cabe destacar que as operações intelectuais de classificação e avaliação são essencialmente atividades descritivas, e que a descrição deve começar já no processo de classificação, aprofundando-se nas etapas subsequentes (Lopes, 1996). Essa visão explora a importância de cada uma das funções arquivísticas mas aponta a relevância da adoção conjunta dessas atividades como ações complementares.

5.2.1.3 A descrição arquivística

A descrição arquivística, como o próprio termo já indica, se caracteriza pela descrição e representação de informações. Trata-se do

[...] ato de descrever e representar informações contidas em documentos e/ou fundos de arquivo gerando instrumentos de pesquisa (inventários, guias, catálogos etc.), os quais explicam os documentos de arquivo quanto a sua localização, identificação e gestão, além de situar o pesquisador quanto ao contexto e os sistemas de arquivo que os gerou. As atividades de descrição são importantes em um arquivo porque **garantem a compreensão do acervo arquivístico** (De Moura Souza *et al.*, 2006, p. 41, grifo nosso).

Pelo fato de inicialmente ser considerada como um balizador da elaboração de instrumentos de pesquisa que auxiliam a identificação, o rastreamento, a localização e o uso das informações (Bellotto, 2008) com fins de viabilizar e promover o acesso aos documentos arquivísticos. Para Bellotto, a descrição é uma atividade típica dos arquivos permanentes. No entanto, de acordo como Norma Geral Internacional de Descrição Arquivística: ISAD (G), trata-se de uma atividade que

[...] abrange todo elemento de informação, não importando em que estágio de gestão ele é identificado ou estabelecido. Em qualquer estágio, a informação sobre os documentos permanece dinâmica e pode ser submetida a alterações à luz de maior conhecimento de seu conteúdo ou do contexto de sua criação (Conselho Internacional de Arquivos, 2000, p. 1).

De toda forma, embora idealmente deva ser aplicada em todo o ciclo de vida dos documentos, dificilmente é realizada nas práticas arquivísticas (Maurel; Champagne, 1999, *apud* Rodrigues, 2003). No que tange aos documentos registrados em formatos analógicos, a descrição é uma atividade ainda mais amplamente realizada nos arquivos permanentes.

Por outro lado, na perspectiva dos documentos digitais, a descrição se relaciona com as outras funções arquivísticas por “inserir informações referentes às possíveis transformações do documento, o que irá corroborar na sua presunção de autenticidade” (Santos; Flores, 2016, p. 170).

Quando o documento digital é produzido já deve ser classificado, e assim, o código de classificação se transforma em um registro nos metadados. Na avaliação, verifica-se o código de classificação e define-se a sua temporalidade e destinação com base na tabela de temporalidade, logo, esta informação descritiva será inserida nos metadados (Santos; Flores, 2016, p. 171, grifo nosso).

Isso significa que o registro de metadados é uma atividade descritiva que perpassa por todas as funções arquivísticas, logo, por todo o ciclo de vida dos documentos arquivísticos.

Assim, a função arquivística de descrição é imprescindível para registro de todas as informações relevantes à atribuição de valor e seleção de dados de pesquisa para preservação. Ao registrar não apenas informações relativas ao dado e seu conteúdo, a descrição registra informação sobre sua proveniência e, portanto, ampara evidências de presunção de autenticidade.

Dessa forma, considera-se que os registros de metadados inerentes às informações relevantes sobre o documento, o contexto e as ações sobre ele executadas, inclusive desde os estágios iniciais da gestão de dados de pesquisa, fazem parte de uma descrição contínua e cumulativa, por intermédio dos metadados, que não podem ser dissociados das demais funções.

Portanto, para fins desta tese, a descrição arquivística aplicada ao contexto dos dados de pesquisa abarca o registro de todas as informações adicionais, conforme disposto no OAIS (2024), ao longo de todo o seu ciclo de vida. Isso posto, compreende-se que o entendimento sobre sua importância e aos aspectos inerentes a sua execução desdobraram-se de maneira minuciosa na seção primária 4 desta pesquisa.

5.2.1.4 A técnica arquivística de identificação

Na Arquivologia, o processo de identificação é uma ferramenta metodológica baseada na análise de documentos a fim de propor soluções para os problemas decorrentes de sua gestão. A partir um diagnóstico desenvolvido sobre o documento de arquivo e seu produtor, é possível planejar estratégias para as questões apresentadas e facilitar o dia a dia da gestão de documentos (Rodrigues, 2013, p. 64-80).

Fundamentada na Diplomática Contemporânea, a identificação arquivística transcende a simples descrição documental, sobretudo no contexto digital, em que a quantidade e a diversidade dos documentos demandam métodos rigorosos para assegurar sua autenticidade, integridade e valor probatório. Rodrigues (2008) destaca que a identificação envolve uma pesquisa crítica sobre a gênese do documento, com o objetivo de determinar sua identidade e seu vínculo com o órgão produtor. Essa função assegura o reconhecimento e a caracterização dos documentos em seu contexto de produção, condição indispensável para garantir sua autenticidade e confiabilidade ao longo do tempo (Duranti, 2001).

Rodrigues (2008) destaca, ainda, a identificação como tarefa essencial do tratamento documental, determinando a identidade do documento e seu vínculo com o órgão produtor. Essa prática fornece base técnica para a presunção de autenticidade, indispensável à confiabilidade dos documentos arquivísticos digitais ao longo do tempo.

No Brasil, o termo vínculo arquivístico tem sido interpretado como tradução do conceito de “*archival bond*” de Duranti (1997). “O vínculo arquivístico é o elemento essencial que garante a autenticidade e a confiabilidade dos documentos ao longo do tempo” (Duranti, 1997, p. 87). Todavia, Claudia Lacombe Rocha (2011, p.85) expõe que “alguns arquivistas se referem ao conceito de *archival bond* como vínculo arquivístico” mas que os arquivistas brasileiros tradicionalmente utilizam o termo relação orgânica, ou ainda organicidade, com o mesmo significado desse vínculo.

Ao definir o conceito de vínculo arquivístico, Duranti (1997), porém, adverte que não deve ser confundido com o contexto de produção documental, pois “o contexto é, por definição, extrínseco ao documento, [...] é uma interpretação”, já o vínculo constitui “parte essencial do documento” (Duranti, 1997, p. 217, tradução nossa). E, conforme afirma Rangel (2015, p. 83) “o vínculo arquivístico não só determina a estrutura dos

fundos de arquivo, como também configura o componente principal de cada documento, ou seja, é ele quem determina o significado do documento”.

No ambiente digital, essa função deve abranger não apenas o conteúdo e a forma dos documentos, mas também os metadados que registram as ações realizadas sobre eles, garantindo a rastreabilidade e contribuindo para a preservação digital qualitativa⁸⁶. Rabelo e Schmidt (2025) destacam que a identificação incorpora uma análise aprofundada da origem, do contexto de produção e do vínculo orgânico dos documentos com as atividades que os geraram.

Embora normas internacionais ofereçam diretrizes relevantes para a gestão de documentos, elas frequentemente não contemplam a complexidade do contexto brasileiro e dos sistemas digitais contemporâneos. Nesse sentido, a identificação arquivística apresenta-se como uma metodologia detalhada e replicável, capaz de superar tais limitações e orientar a elaboração de instrumentos de gestão e preservação digital adequados à realidade institucional e tecnológica (Rabelo; Schmidt, 2025; Rodrigues, 2013).

Para se formular programas e planos institucionais adequados cujo objetivo seja a efetivação da gestão é importante que seja observado o planejamento dessa ação. Para tanto, planejar implica identificar o contexto técnico, administrativo e político. Assim, a aplicação dessa metodologia permite uma caracterização detalhada, considerando sua proveniência e contexto institucional (Silva, 2017).

A partir dessa caracterização inicial, tornam-se possíveis estudos e análises complementares voltados à definição de estratégias adequadas para cada caso, de modo a minimizar os efeitos da possível obsolescência dos formatos e a desfragmentação dos dados, além de garantir a conservação dos que ainda não sofreram danos.

No contexto desta tese, a identificação busca reconhecer também os dados de pesquisa, seus formatos (.pdf, .doc, .odf, entre outros) e seus respectivos volumes, além de averiguar o tratamento que vem sendo dado aos dados e à documentação associada.

⁸⁶ Em estudo recentemente apresentado Feliciati e Duranti (2025) ao trazerem “O uso responsável da Inteligência artificial em arquivos por meio do uso do paradados” apresentam um novo termo “paradados”, enquanto “dados estruturados e formalizados que documentam as metodologias, processos (inclusive tecnológicos) e pessoas (físicas e jurídicas) associadas à produção e gestão de registros” (p. 1), e concluem que os perfis de aplicação e a necessidade de metadados permanecem um campo de estudo em aberto na maioria das disciplinas interessadas na criação e gestão de dados. Ao mesmo tempo, as relações dos metadados com os metadados e a perspectiva de incorporar informações relativas ao processamento de dados na forma de descrições estruturadas estão atualmente em fase de experimentação por vários grupos de pesquisa.” (p.7)

Cabe expor que, mesmo sendo apresentada após as funções arquivísticas de classificação, avaliação e descrição, a identificação deve ser realizada antes de todas elas. Essa etapa inicial permite analisar e caracterizar os dados de pesquisa em seu contexto de produção, assegurar sua individualização e vínculo com as atividades e processos que os originaram e definir padrões e estratégias de gestão que garantam o acesso contínuo e a confiabilidade dos dados.

Além disso, identificar propicia reconhecer a escala de crescimento anual do volume documental do setor e planejar a capacidade de armazenamento necessária para abarcar todo o ciclo de vida dos dados de pesquisa.

Essa identificação permite garantir a rastreabilidade, autenticidade e integridade dos dados de pesquisa ao longo do tempo, por meio da atribuição de metadados padronizados, códigos de classificação e a aplicação dos prazos previstos para retenção. Ou seja, é um processo essencial para facilitar a classificação, a avaliação, a destinação e o acesso adequado desses dados, o que, juntamente com as demais funções já citadas na subseção 5.2.1, promove uma gestão de dados eficiente.

Nesse sentido, a identificação atua como base metodológica para a caracterização detalhada dos dados, considerando não apenas seu conteúdo, mas também os metadados que registram as ações e transformações sofridas durante seu ciclo de vida. Essa abordagem é fundamental para superar os desafios da preservação digital qualitativa e para viabilizar o (re)uso confiável dos dados, alinhando-se aos princípios da Ciência Aberta e às demandas contemporâneas de transparência e colaboração científica

Ademais, a identificação possibilita o controle intelectual e físico dos dados de pesquisa desde sua criação, facilitando a realização das etapas subsequentes como classificação e avaliação. Da mesma forma, apoia a elaboração, e/ou aprimoramento, de Planos de Gestão de Dados (PGD), alinhados às reais necessidades e possibilidades da pesquisa em questão. Rodrigues (2012, p. 78) enfatiza que “a utilização de instrumentos como fichas de identificação documental sistematiza o processo de reconhecimento, promovendo transparência e confiabilidade nos arquivos”.

Como parte de uma metodologia desenvolvida para contribuir na identificação de aspectos relevantes ao estabelecimento de estratégias de preservação dos dados de pesquisa, a Fiocruz (Anexo B) elaborou uma ficha para embasar essa atividade. Essa ficha foi concebida e aplicada institucionalmente por profissionais da Casa de Oswaldo Cruz

(COC/Fiocruz)⁸⁷, em um grupo de trabalho interdisciplinar no âmbito do projeto-piloto de gestão de dados de pesquisa coordenado pela Vice-Presidência de Ensino, Informação e Comunicação da Fiocruz. A proposta visa complementar as informações necessárias à composição de um PGD detalhado e efetivo.

Essa metodologia foi testada em quatro grupos da Fiocruz, selecionados como pilotos por desenvolverem pesquisas transdisciplinares que envolvem a saúde coletiva, psicologia, história e comunicação, entre outras áreas. As perguntas elencadas tiveram por finalidade identificar especificidades contextuais e técnicas dos dados de pesquisa em seus ambientes de produção, como forma de iniciar o desenvolvimento de propostas para uma gestão e preservação efetivas.

Cabe destacar que tal metodologia de identificação de tipos de dados de pesquisa tem por base a tese de doutorado de Sá (2019) que apresenta diretrizes comuns de metadados para dados de pesquisa que atendem a diferentes domínios do conhecimento.

Ressalta-se que

Após a aplicação da metodologia de identificação de tipos de dados de pesquisa da Fiocruz nos quatro pilotos, concluiu-se que os resultados obtidos foram satisfatórios, considerando sua aplicação válida em qualquer tipo de pesquisa, podendo ainda, em cada caso, ser avaliada a necessidade de sua aplicação na íntegra ou de forma parcial (Sá *et al.*, 2021, p. 13).

A partir da identificação, portanto, é possível diferenciar os dados de pesquisa (gerados ou coletados como resultado direto do processo científico) dos dados para pesquisa (aqueles que apoiam o desenvolvimento científico, mas não são gerados pelo pesquisador) e dos dados não pertencentes à pesquisa (apenas utilizados como referência).

Identificar essa distinção é fundamental para que cada tipo de dado receba o tratamento técnico adequado. Os dados de pesquisa devem ser objeto da gestão e preservação iniciadas pelo pesquisador e mantidas pela instituição à qual a pesquisa se vincula. Os dados para pesquisa podem passar por tratamento similar, desde que identificados como documentação de apoio à contextualização e proveniência dos dados de pesquisa. Já os dados de referência não precisam, necessariamente, ser objeto do mesmo tratamento técnico, devendo apenas estar vinculados por meio de *links* ou

⁸⁷ Disponível em <https://revista.arquivonacional.gov.br/index.php/revistaacervo/article/view/1764>. Acesso em: 01 out. 2025.

metadados relacionais, cabendo sua gestão e preservação aos responsáveis originais por sua geração ou coleta.

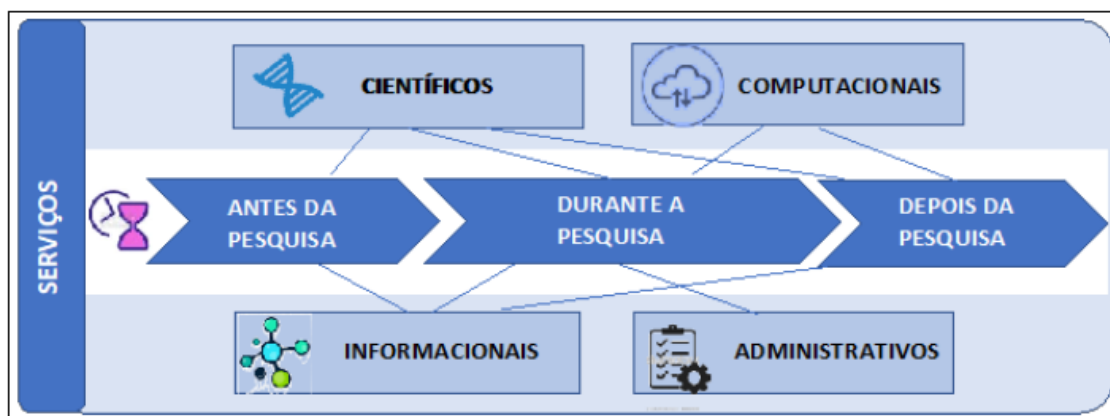
Portanto, a técnica de identificação possibilita que os esforços de tratamento e os recursos sejam aplicados de forma precisa por aqueles que detêm a responsabilidade de realizá-los. Assim, a identificação consolida-se como a base para estruturar a gestão e a preservação digital qualitativa dos dados de pesquisa, garantindo autenticidade, rastreabilidade e transparência em consonância com os princípios da Ciência Aberta.

5.2.2 A relevância do acompanhamento especializado para a efetivação da gestão de dados de pesquisa

Ao examinar a proposta de reintegração de todo o ciclo de vida dos documentos arquivísticos, preconizada pela Arquivística Integrada, bem como os serviços que podem ser disponibilizados durante o desenvolvimento das pesquisas para apoiar e fomentar o compartilhamento e o (re)uso de dados, Sayão e Sales (2022) identificam que a implementação das atividades de gestão de dados permanece inconsistente e, em grande parte, restrita a estágios iniciais, especialmente em instituições de pesquisa de caráter intensivo.

Os autores destacam, então, a necessidade de serviços específicos que envolvam a colaboração de diferentes áreas e expertises, com infraestruturas, práticas e metodologias integradas. Nesse contexto, consideram que a gestão de dados apresenta um espectro contínuo e variável (disciplinar, cultural, epistemológico, institucional e político) e depende das bases tecnológicas disponíveis. Embora possam existir sobreposições importantes e até mesmo desejáveis entre essas dimensões, os autores propõem quatro tipos de serviços: científicos, informacionais, computacionais e administrativos, que devem ser oferecidos ao longo das três fases da pesquisa: antes, durante e depois, conforme figura a seguir.

Figura 26 - Serviços de gestão de dados de pesquisa



Fonte: Sales; Sayão (2022, p. 15).

Em consonância com os objetivos desta pesquisa e considerando o conjunto de serviços informacionais apontados pelos autores na Figura 26, reconhecem-se de forma mais clara as interfaces e os potenciais benefícios de uma integração precoce entre o pesquisador e o profissional da informação. Tal integração se assemelha àquela observada no contexto da gestão de documentos, na qual se estabelece uma relação direta entre o produtor e as orientações formuladas pelo gestor de documentos.

No artigo de opinião “Ciência Aberta: uma visão desapassionada”, Débora Menezes e Ricardo Galvão destacam que, embora a proposta da Ciência Aberta seja meritória, ela envolve desafios de natureza prática que exigem planejamento, capacitação e recursos adequados (Menezes; Galvão, 2025).

Os autores também apontam a necessidade de que sejam consideradas, no caso do Brasil, as especificidades inerentes ao desenvolvimento de pesquisas nas diferentes áreas de conhecimento e de que essa discussão leve em conta a realidade das condições de trabalho dos pesquisadores brasileiros, a devida avaliação dos custos envolvidos na sua implementação e o estabelecimento de regulamentação adequada em âmbito federal.

Ainda que essas questões já tenham sido discutidas na seção 2 e que o referido artigo tenha sido objeto de análise crítica em carta-resposta elaborada por Medeiros *et al.* (2025), sua retomada aqui se justifica. Um dos pontos enfatizados pelos autores é justamente a “falta de ajuda profissional competente para auxiliar os pesquisadores a estruturar os dados da pesquisa e guardá-los”, aspecto que reforça a relevância da mediação técnica especializada (Menezes; Galvão, 2025).

Diante dessa afirmação tem-se a corroboração da visão de que esse é um aspecto pragmático que não pode ser ignorado e, por isso, é um tema que cabe ser detalhado para que possa ser implementado de maneira consistente, mesmo que gradualmente.

Assim, em virtude de todas as variáveis que vêm sendo apresentadas e que devem ser consideradas para o planejamento da pesquisa e do rol de informações que precisam estar bem registradas na composição do PGD, é importante que o pesquisador tenha acesso a orientações devidas ou mesmo assessoramento direto para que essa elaboração realmente reflita as intenções e demandas inerentes ao processo científico que será desdobrado.

Cabe destacar que “as informações adicionais necessárias para garantir a preservação serão coletadas ou criadas por indivíduos e equipes não necessariamente vinculados ao arquivo responsável pela preservação” (CCSDS, 2024, p. 1, tradução nossa). Nesse sentido, a insuficiência na criação ou coleta dessas informações compromete a capacidade de preservação a longo prazo (CCSDS, 2024)

Reforça-se o caráter contraproducente de postergar a mediação de um profissional da área de informação que auxilie o pesquisador em todas as fases da pesquisa, desde o planejamento e elaboração do PGD, perpassando também pela criação e/ou coleta dos dados de pesquisa, identificação, classificação, avaliação e descrição, e assim por diante. A ausência dessa intermediação tende a resultar na omissão de informações essenciais no PGD, comprometendo a identificação de recursos e ações necessárias à adequada gestão dos dados. Tal lacuna pode converter o plano em um documento meramente formal, limitado ao cumprimento de exigências institucionais ou de agências de fomento, sem efetiva aplicação prática.

Isso reafirma a ideia de que não é possível aguardar a iniciativa de depósito desses dados em um repositório que promova sua gestão e compartilhamento para se pensar em como gerir esses dados e seus metadados até porque, embora altamente recomendado, esta ação ainda não é mandatória no cenário brasileiro, o que permite pressupor um tempo alargado entre o momento presente e o qual o pesquisador seja sensibilizado ou institucionalmente induzido a depositar os dados de sua pesquisa.

Nesse contexto, impõe-se questionar de que modo esses dados vêm sendo coletados, tratados, organizados e armazenados. Ao ser considerada a vulnerabilidade dos formatos digitais, indaga-se se tais dados permanecerão localizáveis, legíveis e confiáveis

ao longo do tempo. Ademais, é pertinente refletir se as anotações sobre o processo de pesquisa: métodos, insumos e recursos empregados, estão sendo devidamente registradas.

Se o planejamento fornece o mapa ao qual o ciclo de vida dos dados de pesquisa deve prosseguir para viabilizar a pesquisa e, em adição, agregar qualidades de possibilitem usos diversos e em diferentes tempos, é nele que as atenções dos profissionais de informação devem se voltar à priori para que, juntamente com o pesquisador, sejam constituídas estratégias e traçadas as rotas.

Isso porque as possibilidades são inúmeras, e a capacidade de esses dados serem (re)usados vem constituindo uma de suas faces mais valiosas. No entanto, a prática de os pesquisadores tornarem os dados de pesquisa disponíveis ainda não é comum ou cultural (Chao, 2015), já que muitos deles não possuem treinamento formal em gerenciamento de dados, fazendo com que se torne difícil que eles os disponibilizem de maneira a torná-los visíveis (Federer, 2013).

Nessa perspectiva, importa salientar que o reconhecimento da necessidade de articulação entre o pesquisador, suas produções científicas e a Arquivologia não constitui uma discussão recente. No cenário brasileiro, alguns autores se tornaram referência nesse debate. Em 2002, Verônica Martins de Brito dissertou sobre “A preservação da memória científica da Fiocruz: a visão de quem faz a ciência”, abordando a importância de preservar a memória institucional a partir do olhar dos próprios cientistas. Posteriormente, em 2007, Maria Celina Soares de Mello Silva, em sua tese “Visitando laboratórios: o cientista e a preservação de documentos”, analisou a relação que os pesquisadores mantêm com os documentos produzidos em ambientes científicos e tecnológicos, ressaltando o papel da documentação na dinâmica da produção do conhecimento.

Em continuidade a essas reflexões, em 2008, Paulo Roberto Elian Santos, na tese “A arquivística no laboratório: história, teoria e métodos de uma disciplina”, demonstrou a relevância dos arquivos para a prática científica, tanto como elementos fundamentais à produção e validação do conhecimento quanto para a continuidade e transparência das pesquisas. As contribuições desses autores consolidam um pilar teórico sobre as interlocuções entre os fundamentos arquivísticos e as práticas científicas, promovendo uma melhor compreensão do desenvolvimento da ciência, do papel dos cientistas e da evolução de suas práticas ao longo do tempo.

Entre alguns pontos em comum nas teses dos autores supracitados, destaca-se a análise sobre os limites entre o pessoal e o institucional no contexto científico em

instituições de pesquisa. É relevante apontar que o fato de ignorar que esses dados são institucionais permite que o pesquisador o admita como parte integrante de seu acervo pessoal. Essa apropriação leva a que sejam omitidas as necessidades de serem produzidos e tratados de acordo com as orientações estabelecidas institucionalmente.

Em novembro de 1997, Terry Cook, em palestras proferidas no “Seminário Internacional de Arquivos Pessoais”, patrocinado pela CPDOC-FGV e pelo IEB-USP, já apontava novos desafios para os quais os arquivistas precisavam estar preparados, diante dos novos cenários de produção documental que “seguem funções, são criados como produto do trabalho em várias atividades ou transações” (Cook, 1997, p. 136). Ao refletir sobre o uso cada vez mais intenso dos formatos eletrônicos, o autor observa a necessidade de que os princípios arquivísticos fossem revisitados, a fim de responder às transformações trazidas pelas novas formas de criação, circulação e preservação dos documentos digitais.

Essa necessidade decorre, principalmente, do fato de que a construção metodológica da Arquivologia teve seu maior impulso a partir do século XIX, quando os primeiros esforços teóricos surgiram de profissionais⁸⁸ que, empiricamente, lidavam com questões voltadas à organização e descrição de documentos produzidos pelo Estado, e cuja função era servir aos interesses administrativos públicos. Por essa razão, a teoria, os princípios e as metodologias arquivísticas foram desenvolvidos a partir desse contexto e, posteriormente, difundidos e adotados ao redor do mundo, moldando práticas que hoje precisam ser reinterpretadas à luz dos desafios contemporâneos.

Cabe ressaltar que não se trata de uma desconsideração quanto à importância dessa divisa teórico-metodológica para a construção da área arquivística enquanto uma disciplina científica desde então, o que deve ser compreendido é que a modernização constante da própria humanidade impõe novas demandas que vem impactando a forma de produção, transmissão, uso, tratamento, manutenção dos documentos, gerando novos desafios para os quais os profissionais da informação precisam estar flexíveis e dispostos a atender com propostas e soluções.

⁸⁸ Nesse momento Cook cita o clássico da área arquivística: *Manual dos Arquivistas Holandeses* (1898) e os autores dos posteriores livros marco da primeira metade do século XX: Sir Hilary Jenkinson, Eugenio Casanova e Theodore Schellenberg.

Tal cenário pode ser observado, por exemplo, na inserção de uma nova personagem no ciclo de vida da curadoria dos dados, chamada de *data steward*⁸⁹, que atua de maneira diferente, embora complementar, do *data management* ou gestor de dados. Segundo Mons (2018), a definição de gestão de dados inclui o cuidado a longo prazo dos dados de pesquisa, já o *data stewardship* é definido como “o processo e a atitude que faz com que alguém lide de forma responsável com os próprios dados e os de outras pessoas durante e após o ciclo inicial de criação e descoberta científica” (Mons, 2018, p. 36).

A gestão de dados é um conceito que já está em uso há mais tempo. A maioria dos financiadores e pesquisadores interpretam isso como cuidar bem dos dados durante o processo de pesquisa. Obviamente, isso é muito importante, mas na era da ciência aberta e orientada por dados, muitos conjuntos de dados têm uma vida útil muito além do projeto que os gerou originalmente. Portanto, o *Data stewardship*⁹⁰, aqui definido grosso modo como: tratar os dados e os objetos de pesquisa associados com o máximo cuidado, com o objetivo de torná-los reusáveis para descoberta, desde que sejam válidos, não é uma tarefa chata e inevitável para pessoas empoeiradas e conservadoras, mas uma profissão em rápido desenvolvimento no coração da ciência moderna (Mons, 2018, p. 24, tradução nossa, grifo nosso).

Outra definição para esse termo o entende como uma ação responsável e comprometida no gerenciamento e supervisão de ativos de dados organizacionais, exercida por pessoa ou grupo, e que envolve cuidados adequados, uso dos Princípios FAIR e a manutenção da propriedade e da regulamentação no fornecimento de dados de alta qualidade (incluindo metadados), combinando confiança e prática ética (CODATA, 2023)⁹¹.

Para Marín Arraiza e Vidotti (2019, p. 20), no contexto acadêmico, os *data stewards*, traduzido pelas autoras como “administradores de dados” atuam em cada faculdade atendendo às necessidades de uma área do conhecimento específica na qual é especialista, tendo um amplo papel ao participar desde o desenvolvimento dos processos de governança de dados que devem estar em consonância com as políticas de dados vigentes (no nível de faculdade, universidade, estadual ou nacional) até a custódia dos

⁸⁹ Segundo Marín-Arraiza e Vidotti (2019), algumas vezes o *Data Steward* também é denominado como *Data Custodian*, em português “custodiador de dados” (p.262).

⁹⁰ Optou-se pela não tradução de “*data stewardship*” em virtude de ainda não haver um termo em português consolidado na literatura da área. Atualmente, qualquer tradução poderia implicar em uma compreensão equivocada.

⁹¹ Disponível em: <https://codata.org/initiatives/data-science-and-stewardship/rdm-terminology-wg/>. Acesso em: 04 ago. 2025.

dados e adequação aos padrões estabelecidos em cada área do conhecimento ou instituição.

Segundo Wilkinson *et al.* (2016), uma boa gestão de dados (RDM) e o *data stewardship* não são um objetivo em si mas são as principais vias que levam à descoberta e à subsequente integração e (re)uso de dados e conhecimento pela comunidade após o processo de publicação de dados, podendo ser considerados uma pré-condição para apoiar também a inovação do conhecimento. Portanto, entende-se que conforme a curadoria vai sendo aplicada, também vai sendo aprimorada e, provavelmente, daqui a alguns anos novas percepções e concepções surgirão em vistas a sua efetiva implementação no pragmatismo científico que também se adapta aos cenários de tempo e espaço nos quais se insere.

Dessa forma, observa-se que o desenvolvimento prático das atividades inerentes à gestão e curadoria de dados de pesquisa nos contextos acadêmicos e científicos vem se consolidando a partir da construção teórica e metodológica advinda de um repertório de conhecimento de profissionais da área de Biblioteconomia e Tecnologia da Informação com participação colaborativa de seções de apoio das instituições sob os quais as pesquisas se vinculam, como o jurídico e os de pesquisa.

Assim, por consequência positiva, tem-se que os serviços de gestão e curadoria de dados estão se tornando cada vez mais comuns entre as bibliotecas de pesquisa, as quais vem se reinventando para atender à crescente demanda de gerenciar, armazenar, curar e preservar esses dados, principalmente, quando produzidos por pesquisas financiadas com recurso público (Soehner, Steeves e Ward, 2010).

A análise dos perfis de profissionais que trabalham mais vinculados à gestão de dados no cenário acadêmico⁹², Marín-Arraiza e Vidotti (2019) apontam três figuras: o bibliotecário acadêmico de dados, o administrador de dados e o gestor de dados. Na avaliação das autoras, enquanto o “bibliotecário de dados” atua como uma figura central dentro da biblioteca e se encarrega da supervisão geral dos dados de pesquisa e dos sistemas da informação disponíveis na instituição, o “administrador de dados” encontra-se em cada faculdade ou centro de pesquisa, assessorando os pesquisadores de determinada área do conhecimento nas tarefas de gestão desses dados, devendo, portanto,

⁹² As autoras apontam que fora da academia outros profissionais de dados foram observados, como: o jornalista de dados, o analista de inteligência de negócio ou o analista de dados.

ter o título de doutor, ou prestes a alcançá-lo, já que precisa ser um especialista na área em que atua. Já o “gestor de dados” trabalha dentro do grupo de pesquisa e se encarrega da análise e preparação final dos dados seguindo os Princípios FAIR e os padrões comuns da área.

Ao destrinchar o levantamento sobre as competências e tarefas desses três perfis a partir de diversos autores e da descrição de vagas de emprego para gestores de dados, as autoras apresentam o seguinte quadro:

Quadro 6 – Extrato do quadro comparativo de competências dos bibliotecários acadêmicos de dados, administradores de dados e gestores de dados

	Bibliotecário acadêmico de dados	Administrador de dados	Gestor de dados
Competências e tarefas	Desenho e desenvolvimento conceitual de sistemas (conhecimento de tecnologia, metadados, protocolos, infraestruturas digitais de pesquisa, governança, curadoria digital, etc.) (RICE; SOUTHALL, 2016)	Assessoramento específico aos pesquisadores sobre GDP (VERHEUL et al., 2019)	Concepção de planos de gestão de dados específicos para cada projeto.
	Domínio do ecossistema de pesquisa (KENNAN, 2016; XIA; WANG, 2014)	Desenvolvimento de políticas de gestão de dados no nível de faculdade (VERHEUL et al., 2019)	Proposta de soluções para a adaptação de dados existentes aos princípios FAIR.
	Conhecimento e aplicação de fluxos de trabalho em infraestruturas digitais (RICE; SOUTHALL, 2016; XIA; WANG, 2014)	Informação e formação aos investigadores sobre GDP (VERHEUL et al., 2019)	Contribuição na estratégia geral em torno dos planos de gestão de dados para o departamento.
	Conhecimento de tecnologias da informação e familiaridade com as linguagens de programação para dados (R e Python) (BRADLEY-RIDOUT, 2018; FEDERER, 2018; KENNAN, 2016)	Aconselhamento sobre a escrita de um plano de GDP (JANSEN et al., 2019; VERHEUL et al., 2019; MONS, 2018)	Concepção de modelos de metadados para os dados do projeto com base nas ontologias e normas existentes.
	Conhecimento de métricas para infraestruturas de pesquisa (FEDERER, 2018; RICE; SOUTHALL, 2016)	Conhecimento dos padrões de metadados e ontologias presentes na área do conhecimento (VERHEUL et al., 2019; MONS, 2018)	Co-desenvolvimento de software e ferramentas para o projeto.
	Conhecimento de GDP (FEDERER, 2018; SNIPES, 2018; KENNAN, 2016; RICE; SOUTHALL, 2016; XIA; WANG, 2014; TENOPIR et al., 2014)	Conhecimento dos aspectos legais e éticos da gestão de dados (JANSEN et al., 2019; DIJKERS, 2019; MONS, 2018)	Análise dos dados dentro do projeto.
	Conhecimento de catálogos de sistemas e dados (qualidade de metadados, semântica e common standards) (KENNAN, 2016)	Trabalho direto com novos fluxos de trabalho e novos desenvolvimentos de infraestrutura para a área (DIJKERS, 2019; VERHEUL et al., 2019)	
	Habilidades comunicativas e didáticas (KENNAN, 2016)		

Fonte: Adaptada pela autora com base em Marín-Arraiza e Vidotti (2019, p. 264).

Como é possível observar no Quadro 6, embora haja diferentes perfis com especificidades distintas, eles se destinam a ações que se complementam ao longo do fluxo da pesquisa, o que implica que esses profissionais especialistas estejam integrados

em uma unidade de informação destinada à gestão de dados no âmbito institucional, conforme reforçam as autoras. Cabe ressaltar que o quadro em questão também reflete o cenário em que a gestão de dados vem sendo pauta de discussões e campo apoiado pelas bibliotecas acadêmicas há quase duas décadas porém, apesar de ser um serviço que se constituiu, predominantemente, sob o viés da biblioteconomia, a participação de outras áreas do conhecimento voltadas à informação também se mostra oportuna, reconhecendo que o desenvolvimento de políticas de gestão de dados de pesquisa é um processo multissetorial que demanda uma colaboração multifacetada (Cox *et al.*, 2017).

Nesse sentido, ao reconhecer a experiência dos bibliotecários na catalogação e registro de metadados descritivos para recuperação da informação e, no presente caso, dos dados de pesquisa, frisa-se que a garantia de acesso requer ainda outras ações como a manutenção da capacidade de confiabilidade e renderização desses artefatos digitais, o que demanda registros mais detalhados e específicos. Assim fica nítido considerar que trata-se de um domínio complexo, caracterizado por múltiplas especificidades e desafios técnicos que devem ser vistos como um campo fértil para a colaboração de variadas expertises. Portanto, tem sido trazida para reflexão nesta tese, como suporte incremental à gestão de dados de pesquisa, os benefícios que podem ser alcançados ao incorporar o estofo teórico-metodológico da Arquivologia nesse cenário.

Ao analisar o trabalho que foi desenvolvido em 2005 por uma Comissão do CNPq responsável por estudar a reestruturação das áreas de conhecimento no Brasil, Angelica da Cunha Marques (2016) destrincha os aspectos envolvidos na proposta de alocação da Arquivologia como uma área de conhecimento sob o mesmo nível hierárquico classificatório da Ciência da Informação e não mais como a ela subordinada (isso porque, até então, a Arquivologia encontra-se inserida como sua subárea).

Nesse artigo onde o papel da Arquivologia enquanto uma disciplina científica em formação desde o século XVI, conforme reforça a autora ao referenciar o estudo de Fonseca (2004), que vem passando por diversos escrutínios e obstáculos quanto a sua conformação e possíveis encaixes no cenário científico brasileiro, é possível encontrar um bom argumento para empreender a colaboração arquivística nas discussões sobre a gestão de dados de pesquisa, inclusive pela ratificação do objeto de estudo da Arquivologia: a informação orgânica registrada.

Objeto este que, segundo Thomassen (1999), se consolida a partir do que o autor entende como um novo paradigma da Arquivologia (pós-custodial), o qual deslocou do

“arquivo” para a “informação arquivística” ou “informação orgânica registrada” (Couture; Ducharme; Rosseau, 1988). Enquanto o primeiro tem foco no conteúdo semântico do documento, o segundo destaca o conteúdo diplomático do documento (Fonseca, 2005), o que Silva (2008, p. 44) aponta como sendo a pressuposição da “nítida predominância do conteúdo sobre o suporte”.

Conforme da Cunha Marques aponta desde sua dissertação, em 2007, a Arquivologia vem estabelecendo sua formação e configuração a partir de relações com outras disciplinas e áreas do conhecimento, sobretudo a Ciência da Informação, o que a autora afirma não afetar sua identidade como uma disciplina científica, ao contrário, subsidiam-na e que a adesão da Arquivologia à CI (ao menos aos seus discursos) se deu para que ela se acomodasse no campo da informação, sugere a autora (Marques, 2013).

Reforça-se, então, que no Brasil, conforme descrito no site da Universidade Federal do Estado do Rio de Janeiro (UNIRIO)⁹³ os graduados no curso de Arquivologia, portanto arquivistas, conforme Lei 6.546, de 1978, regulamentada pelo Decreto 82.590, do mesmo ano, são profissionais polivalentes capacitados a acompanhar todo o ciclo vital dos documentos produzidos por um órgão ou entidade no desempenho de suas funções e atividades, sendo qualificado enquanto gestor de documentos e arquivos.

Sua atuação abarca o gerenciamento da informação produzida pelas inúmeras atividades da sociedade a partir de um conhecimento técnico-científico que deve capacitá-los a coordenar e controlar a produção, o fluxo e a difusão dessa informação em qualquer tempo e lugar. Essa perspectiva é corroborada ao se verificar o “Quadro de oferta de disciplinas do curso de Arquivologia” dessa mesma Universidade⁹⁴ que, desde 2013, tem como algumas de suas disciplinas obrigatórias relevantes a serem destacadas dentro deste contexto:

- a) Gestão da informação arquivística, cuja ementa inclui: Arquivo Corrente - conceito, função, rotinas, serviço de protocolo, tramitação, gerenciamento de processos. Arquivo Intermediário - conceito, função, administração da eliminação e do recolhimento. Arquivo Permanente - conceito, função, recolhimento, custódia, acesso. A sociedade, a pesquisa e os Arquivos Permanentes;

⁹³ Disponível em: <https://www.unirio.br/arquivologia/a-profissao-de-arquivista>. Acesso em: 11 set. 2025.

⁹⁴ Disponível em: <https://www.unirio.br/arquivologia/arquivos/Proposta%20Matriz%20Arquivologia%20versao%202013%20COMPLETA.pdf>. Acesso em: 04 ago. 2025.

- b) Classificação de documentos arquivísticos com: Noções de lógica; Teoria da Classificação; a classificação como processo intelectual; estrutura de sistemas de classificação arquivística; métodos de classificação em sua ementa;
- c) Avaliação de documentos arquivísticos, com a seguinte ementa: Aspectos políticos e ideológicos da avaliação; avaliação, temporalidade e destinação; os valores arquivísticos dos documentos; registro esquemático da avaliação - Tabela de Temporalidade
- d) Gestão de documentos arquivísticos, incluindo na ementa: Princípios da teoria da gestão de documentos - conceituação, objetivos e campo de aplicação; o ciclo vital dos documentos - produção, utilização e destinação; planejamento e estruturação de sistema de arquivos.

E como disciplinas optativas, destacam-se as seguintes:

- e) Gestão arquivística de documentos eletrônicos, com reflexões e aplicações da gestão arquivística em sistemas e ambientes de produção eletrônica de documentos); e
- f) Documentação audiovisual e digital, com estudo dos vários arquivos audiovisuais; os métodos utilizados na organização e recuperação da informação em suporte não convencional; e a importância e a metodologia de preservação destes documentos.

Assim, à luz das discussões apresentadas, reconhece-se que a cooperação entre profissionais arquivistas e os demais especialistas envolvidos na gestão de dados é indispensável. Tal colaboração revela-se essencial porque o escopo dos “dados de pesquisa” abrange múltiplas tipologias documentais e requer análises, soluções e intervenções oriundas de diferentes domínios da informação.

Observa-se que, nos estudos que tratam da execução da gestão de dados de pesquisa, a natureza multifacetada e multidisciplinar desse processo é amplamente reconhecida. Entretanto, a contribuição da Arquivologia ainda se revela incipiente, sobretudo diante da escassa produção de trabalhos em que arquivistas se posicionem explicitamente sobre o tema e apresentem seus potenciais contributos teórico-metodológicos.

O aporte conceitual e prático da Arquivologia, ao fundamentar a definição de padrões e fluxos de gestão de dados, constitui um elemento agregador nesse campo ainda

em consolidação. Tal contribuição deve ser compreendida como complementar e convergente, e não como concorrente, em relação às demais áreas que participam, ou venham a participar, desse esforço interdisciplinar.

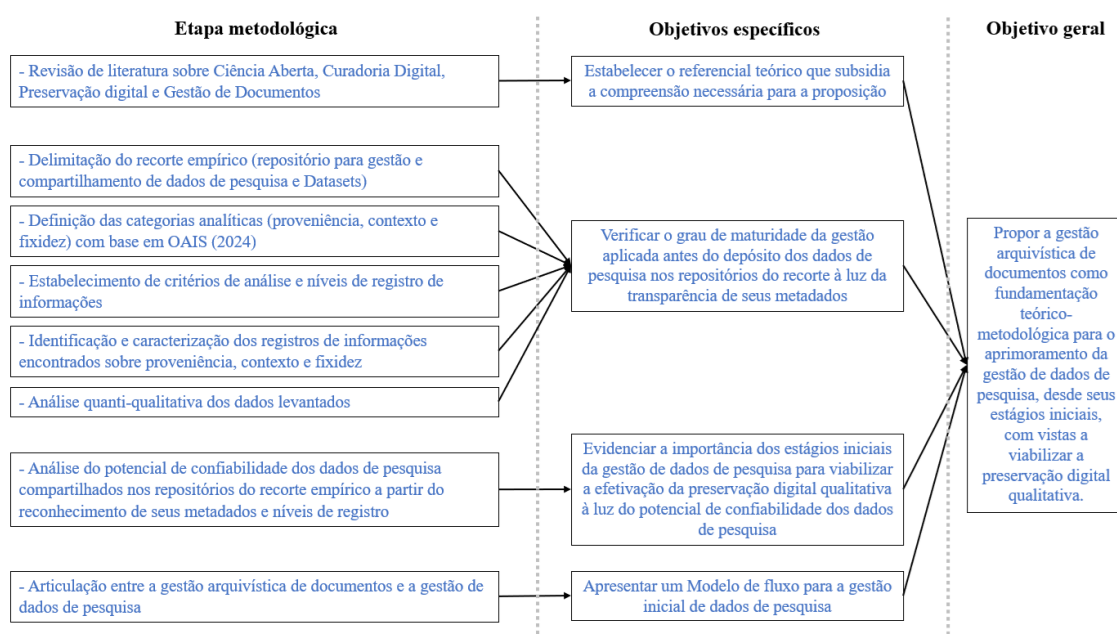
A partir das reflexões apresentadas, evidencia-se que a efetivação da gestão de dados de pesquisa requer não apenas o compromisso individual do pesquisador, mas o acompanhamento especializado de profissionais da informação, em especial arquivistas, bibliotecários e demais perfis técnicos envolvidos na curadoria, descrição, preservação e disponibilização de dados em ambientes institucionais. A consolidação desse acompanhamento, ancorado em fundamentos teórico-metodológicos da Arquivologia, da Ciência da Informação e das práticas emergentes de *data stewardship*, mostra-se condição estruturante para garantir autenticidade, rastreabilidade, acessibilidade e (re)uso responsável dos dados de pesquisa, em consonância com os princípios da Ciência Aberta.

Nesse contexto, torna-se pertinente verificar em que medida tais pressupostos vêm sendo incorporados na prática. Para tanto, na seção seguinte, apresenta-se a metodologia adotada para a análise do grau de maturidade e transparência dos metadados dos dados de pesquisa disponibilizados nos *datasets* selecionados como recorte empírico desta tese, de modo a analisar o referencial teórico discutido com a realidade observada nos ambientes de compartilhamento de dados.

6 PERCURSO DA PESQUISA DE CAMPO

Para conduzir esta pesquisa, considerou-se necessária a realização de etapas metodológicas que viabilizassem o alcance dos objetivos específicos e, conseqüentemente, do objetivo geral. Para tanto, foram realizadas oito etapas metodológicas, conforme demonstrado na Figura 27, que apresenta a correlação entre essas etapas, os objetivos específicos e o objetivo geral:

Figura 27 – Correlação entre etapas metodológicas, objetivos específicos e objetivo geral da pesquisa.



Fonte: Elaborada pela autora (2025).

O estabelecimento do referencial teórico necessário para fundamentar a proposição do Modelo foi realizado a partir da revisão de literatura sobre Ciência Aberta, Curadoria Digital, Preservação Digital e Gestão de Documentos, conforme os parâmetros apresentados na subseção 1.4. Esse referencial embasou o desenvolvimento das seções primárias 2, 3, 4 e 5 desta tese.

A verificação do grau de maturidade da gestão de dados de pesquisa aplicada aos conjuntos de dados selecionados antes do depósito nos repositórios do recorte à luz da transparência de seus metadados foi conduzida por meio da delimitação do recorte empírico, da definição das categorias analíticas: proveniência, contexto e fixidez, com base no OAIS (2024), do estabelecimento de critérios de análise e níveis de registro de

informações, da identificação e caracterização dos registros encontrados e da análise quanti-qualitativa dos dados levantados. Essas etapas são apresentadas e detalhadas ao longo desta seção primária 6.

Com o objetivo de evidenciar a importância dos estágios iniciais da gestão de dados de pesquisa para viabilizar a efetivação da preservação digital qualitativa à luz do potencial de confiabilidade, foi realizada a análise desse potencial nos dados compartilhados nos repositórios do recorte empírico, mediante a verificação de seus metadados e níveis de registro, conforme exposto na subseção 6.4.

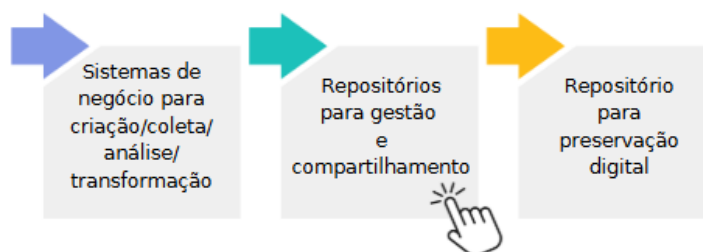
Para o desenvolvimento de um Modelo de fluxo para a gestão inicial de dados de pesquisa, apresentado na subseção 7.2 desta tese, foi realizada uma articulação entre a gestão arquivística de documentos e a gestão de dados de pesquisa ao longo da seção primária 7.

6.1 DELIMITAÇÃO DO RECORTE EMPÍRICO

Para conhecer quais ações devem ser realizadas durante a gestão de dados de pesquisa com vistas a propiciar o registro de metadados capazes de demonstrar a proveniência dos dados, é necessário compreender melhor como essa dinâmica pode ser observada na prática e, portanto, optou-se pela realização de análise empírica.

Ao identificar que os repositórios de gestão de dados de pesquisa encontram-se como elo entre o ambiente onde esses dados são criados, coletados, analisados e/ou transformados e o ambiente destinado à sua preservação digital permanente, para fins desta tese, eles estão sendo considerados um ambiente intermediário no fluxo, conforme Figura 28.

Figura 28 – Fluxo de ciclo de vida dos dados de pesquisa com ênfase no uso de ferramentas (pela abordagem da Cadeia de Preservação)



Fonte: Elaborada pela autora (2025).

Dessa forma, considera-se que a observação de como os dados de pesquisa estão disponibilizados para compartilhamento na ferramenta utilizada nessa fase intermediária foi um caminho viável para a análise de como esses dados, caso sejam considerados com relevância perene, poderão ser encaminhados, em nível qualitativo informacional, para preservação permanente, numa etapa subsequente.

Buscou-se, então, identificar se a proveniência tem sido evidenciada nos dados de pesquisa e de que forma isso pode ser observado nos dados já disponibilizados em repositórios voltados à sua gestão. Parte-se da consciência de que as dissonâncias no grau de maturidade das ações relacionadas à gestão de dados aderente às premissas da Ciência Aberta decorrem tanto de possibilidades quanto de limitações específicas (orçamentárias, técnicas e institucionais, em níveis nacional, regional e local). Diante disso, entendeu-se que o recorte empírico dessa análise deveria ser definido em camadas.

São elas:

- 1) País de origem do repositório;
- 2) Repositório disponível no FAIRSharing;
- 3) Status do repositório;
- 4) *Software* utilizado; e
- 5) Assunto no qual o repositório é classificado.

A primeira camada foi a escolha do país. Considerando que no Brasil já existem iniciativas e instituições atuantes para o desenvolvimento de pesquisas científicas abertas, considerou-se interessante o espectro pretendido sob o contorno nacional. Nesse sentido, definiu-se a seleção de repositórios para gestão de dados administrados por instituições de pesquisa brasileiras. Diante dessa escolha preliminar, optou-se por identificar o rol desses repositórios no site DataCite (datacite.org)⁹⁵, porém durante alguns dias (23, 24 e 25 de março de 2025) todas as tentativas de acesso foram frustradas, aparentemente, pela indisponibilidade do site.

Diante disso, optou-se por realizar a pretendida busca a partir do site do FAIRSharing (fairsharing.org). Entende-se que essa segunda opção tenha feito ainda mais

⁹⁵ Segundo disponível no site, o DataCite é uma “comunidade global que compartilha um interesse comum: garantir que os resultados e recursos de pesquisa estejam abertamente disponíveis e conectados para que sua reutilização possa avançar o conhecimento entre as disciplinas, agora e no futuro” (tradução nossa)

sentido visto o escopo global da presente tese. Isso porque esse site é um recurso informativo e educacional que tem por objetivo coletar, descrever e interligar informações necessárias para garantir que padrões, bancos de dados, repositórios e políticas de dados voltados para a comunidade de pesquisa estejam alinhados com os Princípios de dados FAIR (Sansone *et al.*, 2019). Assim, pressupondo que os repositórios ali cadastrados estejam alinhados com os Princípios FAIR e, portanto, demonstrem atenção à promoção de registros de metadados ricos, considerou-se uma escolha adequada para propiciar a amostragem desejada.

Ao acessar o FAIRSharing, foi utilizado o campo de “busca avançada” onde foram aplicados os seguintes filtros: (1) “Registro”: Banco de Dados; (2) “País”: Brasil (passo 2), resultando em 12 registros; (3) “Tipo de registro”: Repositório, o que reduziu a 7 registros.

Figura 29 – Passos para identificação de repositórios brasileiros para gestão de dados cadastrados no FAIRSharing



Fonte: Elaborada pela autora (2025).

Com essa busca demonstrada na Figura 29, foi possível conhecer os 7 (sete) repositórios brasileiros cadastrados no FAIRSharing, de acordo com a filtragem realizada. Dentre as informações disponíveis, para esta pesquisa, foram destacados no quadro abaixo, o nome, a descrição e o ano de criação de cada um deles.

Figura 30 - Repositórios brasileiros para gestão de dados cadastrados no FAIRSharing

Nome		Descrição	Ano de criação
1	SciELO Data	Repositório multidisciplinar para deposição, preservação e disseminação de dados de pesquisa de artigos submetidos e aprovados para publicação, já publicados em periódicos da Rede SciELO ou depositados em SciELO Preprints.	2020
2	Consórcio para Caracterização Clínica da COVID-19 por EHR (4CE)	Consórcio internacional para estudos baseados em dados de registros eletrônicos de saúde (EHR) da pandemia de COVID-19. O objetivo desse esforço - liderado pelo grupo de usuários acadêmicos internacionais i2b2 - é informar médicos, epidemiologistas e o público sobre pacientes com COVID-19 com dados adquiridos por meio do processo de assistência médica.	2020
3	Repositório de Dados de Pesquisa da EMBRAPA (REDAPE)	Repositório Institucional de Dados de Pesquisa hospedado pela Empresa Brasileira de Pesquisa Agropecuária que permite a organização, gestão e publicação de dados de acordo com os princípios que orientam a gestão de dados científicos em todo o mundo, incluindo acessibilidade, interoperabilidade, reprodutibilidade e reutilização.	2021
4	ARCA DADOS	Repositório oficial de dados de pesquisa da Fiocruz. Sua missão é ser a plataforma de arquivamento, publicação, divulgação, preservação e compartilhamento de dados de pesquisa digital produzidos pela comunidade da Fiocruz e/ou em parceria com outros institutos ou órgãos de pesquisa, seguindo os princípios do FAIR.	2019
5	SpeciesLink	Também conhecido como Sistema de Informações Distribuídas para Coleções Biológicas: Integrando Analista de Espécies e SinBiota (FAPESP)) implementa um sistema de informação distribuído para recuperar dados primários de biodiversidade de coleções dentro do estado de São Paulo (Brasil), bem como dados integrados de outras redes e dados de observação registrados no banco de dados SinBiota.	2002
6	Brazilian Flora 2020	O projeto Flora Brasileira 2020 faz parte do Programa Reflora e está sendo conduzido com o apoio do Sistema de Informação sobre a Biodiversidade Brasileira (SiBBR) e os resultados da pesquisa nesta página incluem informações sobre espécies ameaçadas de extinção (graças à cooperação com o Centro Nacional de Conservação da Flora) e permitem o acesso ao Index Herbariorum (devido à colaboração do The New York Botanical Garden). Os usuários podem acessar imagens de espécimes de herbário, incluindo tipos nomenclaturais, tanto do Herbário Virtual Reflora quanto do Herbário Virtual INCT de Flora e Fungos; bem como imagens de plantas vivas e ilustrações científicas com todas as imagens incluídas pelos especialistas de cada grupo.	2010
7	FluxData	Um portal de dados para a comunidade Fluxnet em todo o mundo que suporta o acesso e o compartilhamento de dados sobre as medições de fluxo de covariância parasita de carbono, vapor de água e troca de energia.	1997

Fonte: Elaborada pela autora (2025).

Diante dessa listagem da Figura 30, foi aplicada a seguinte metodologia para a escolha do campo empírico:

Figura 31 – Metodologia aplicada para escolha do campo empírico



Fonte: Elaborada pela autora (2025).

Assim, dos 7 (sete) repositórios, descartou-se o repositório *Brazilian Flora 2020*, por estar cadastrado com o status “em desenvolvimento”. Dos 6 (seis) repositórios restantes, identificados como “prontos”, observou-se uma considerável diversidade quanto às suas naturezas, formas de apresentação e *softwares* utilizados. Essa heterogeneidade poderia, por um lado, ampliar a compreensão sobre diferentes usos de repositórios em um contexto mais amplo; por outro, limitaria o aprofundamento necessário para reconhecer os aspectos informacionais específicos pretendidos, podendo gerar imprecisões nas análises comparativas. Diante disso, optou-se por estabelecer uma nova camada de recorte.

Com o objetivo de viabilizar a realização de uma análise minuciosa dos conjuntos de dados que viriam a constituir o campo empírico desta pesquisa, adotou-se como critério de seleção dos repositórios a utilização do mesmo *software*. Com base nessa definição, observou-se que 3 (três) repositórios⁹⁶ utilizam o Dataverse Project.

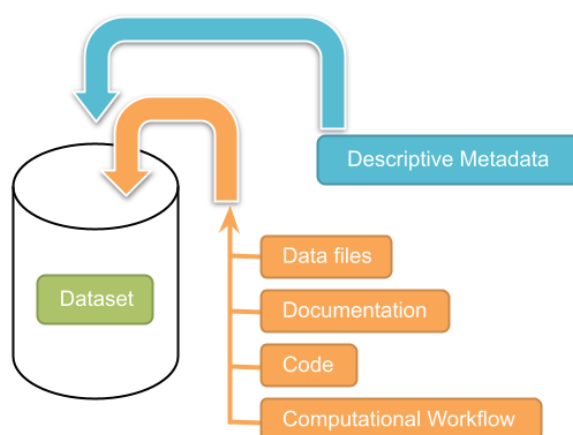
Trata-se de um *software* de código aberto desenvolvido pelo Instituto de Ciências Sociais Quantitativas (IQSS) da Universidade de Harvard, em parceria com colaboradores internacionais. Sua ampla adoção por instituições de diferentes países,

⁹⁶ Entendeu-se que a análise do que se tem como metade do quantitativo total, ou seja, esses 3 repositórios, poderia consubstanciar um bom panorama qualitativo.

associada à familiaridade prévia desta autora com sua estrutura e funcionalidades, reforçou a escolha desse sistema para compor o recorte empírico.

Cumpra esclarecer que um repositório do Dataverse⁹⁷ é a instalação do software, que hospeda vários arquivos virtuais chamados de coleções do *dataverse*. Cada uma dessas coleções do *dataverse* contém conjuntos de dados (*datasets*) e cada conjunto de dados contém os arquivos de dados (incluindo documentação, código e workflow que acompanham os dados) e seus metadados descritivos. Como um método de organização, as coleções do *dataverse* também podem conter outras coleções do *dataverse*⁹⁸.

Figura 32 – Diagrama esquemático do *dataverse*



Fonte: Guia do usuário Dataverse⁹⁹.

Ao selecionar os repositórios com distintas naturezas, mas com uso comum do Dataverse Project, foi possível observar que até o momento os 3 (três) repositórios disponibilizavam, quando somados, 203.517 arquivos de dados em 1.794 *datasets* de 164 *dataverses*. Esses *dataverses* abordam os assuntos: Medicina, Saúde e Ciências da vida, Ciências sociais, Ciências da Terra e Meio ambiente, Engenharia, Ciências Agrárias, Negócio e Gestão, Computação e Ciência da Informação, Artes e Humanidades, Direito, Química, Ciências Matemáticas, Física e Astronomia e Astrofísica. Sendo os sete primeiros comuns aos 3 (três) repositórios. Além disso, é relevante indicar que o

⁹⁷ Para fins desta tese, o software Dataverse Project está sendo exposto como um nome próprio, com Letra inicial maiúscula, enquanto as coleções, também denominadas *dataverses*, estão sendo grifada em fonte minúscula e itálica.

⁹⁸ Disponível em: <https://dataverse.org/about>. Acesso em: 20 out. 2025.

⁹⁹ Disponível em <https://guides.dataverse.org/en/6.2/user/dataset-management.html>. Acesso em: 20 out. 2025.

somatório de *downloads* realizados, no período de 2021 até a data de coleta, já ultrapassava a marca de 798.000.

Em seguida, para viabilizar a verificação adequada em tempo hábil, dentre os diversos “assuntos” em comum, foi escolhido “Medicina, saúde e ciências da vida”, pelo fato de ser o que possui maior quantitativo de dados em 2 (dois) dos 3 (três) repositórios escolhidos. A partir da escolha do assunto, foi verificada a existência de 93 (noventa e três) *dataverses*.

Visto que a pretensão é, exclusivamente, avaliar o conteúdo informacional disponível na amostragem dos *datasets* dos repositórios selecionados, julgou-se inoportuna a sua identificação. Dessa forma, os repositórios serão apresentados como Repositório 1, Repositório 2 e Repositório 3. Pelo mesmo motivo, os *datasets* também receberam uma codificação numérica para fins de apontamento das especificações levantadas nesta oportunidade.

Dessa forma, desses 93 (noventa e três) *dataverses*, 71 (setenta e um) são do Repositório 1, 4 (quatro) são do Repositório 2 e 19 (dezenove) são do Repositório 3, compostos a partir da seguinte distribuição anual:

Quadro 7 – Visualização dos *dataverses* do recorte por ano

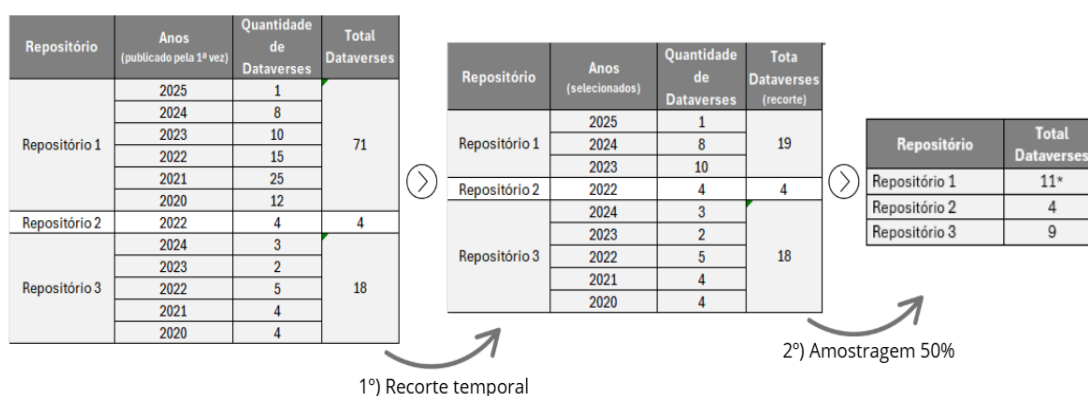
Repositório	Anos (publicado pela 1ª vez)	Quantidade de <i>dataverses</i>	Total <i>dataverses</i>
Repositório 1	2025	1	71
	2024	8	
	2023	10	
	2022	15	
	2021	25	
	2020	12	
Repositório 2	2022	4	4
Repositório 3	2024	3	18
	2023	2	
	2022	5	
	2021	4	
	2020	4	

Fonte: Elaborada pela autora (2025).

Considerando esse montante elevado para viabilização de uma verificação pormenorizada, considerando sua respectiva divisão por anos, buscou-se estabelecer os seguintes critérios para a seleção de uma amostragem:

- 1º) Recorte temporal. Fundamentado no quantitativo dos repositórios, foi reconhecida a necessidade de recortar o escopo do Repositório 1 para os *dataverses* dos anos mais recentes (2025, 2024 e 2023) que, no entendimento desta pesquisa, são capazes de demonstrar o conteúdo informacional desses conjuntos de dados de forma concomitantemente significativa e atual. Esse recorte não foi considerado necessário para os demais repositórios em virtude de seus satisfatórios quantitativos distribuídos pelos seus respectivos anos.
- 2º) Amostragem de 50% dos *dataverses* do montante resultado do recorte temporal. Esse critério não foi utilizado para o montante do Repositório 2, em virtude de sua quantitativo já ser limitado.

Figura 33 – Passo a passo dos critérios para amostragem



Fonte: Elaborada pela autora (2025).

A amostragem de 50% dos *dataverses* do recorte do Repositório 1 foi estabelecida por 50% do total dos anos de 2024 e 2023 ($19:2 = 9,5$, que se aproximou para 10) mais a soma do *dataverse* de 2025, visto ser o único exemplar desse ano corrente até o momento, não cabendo ser computado para fins de divisão, totalizando então 11 (onze) *dataverses* para o recorte desse repositório. A partir disso, foi definido aprofundar a análise em 1 *dataset* de cada um dos *dataverses* do recorte final, os quais foram escolhidos randomicamente, totalizando 24 (vinte e quatro) *datasets* para a verificação pretendida, dentre os quais 11 (onze) são do Repositório 1, 4 (quatro) do Repositório 2 e 9 (nove) do Repositório 3.

Com o recorte estabelecido, procedeu-se à verificação de registros que indicassem elementos informacionais essenciais à presunção de autenticidade dos dados de pesquisa disponibilizados. À luz do referencial teórico desta tese, essa análise centra-se na verificação da proveniência, articulada às categorias de PDI do modelo OAIS: proveniência, contexto e fixidez.

6.1.1 Categorias analíticas e Níveis de registro informacional

Verifica-se, em estudo recentemente publicado por Da Silva e Arakaki (2025), que dentre os 3.302 (três mil, trezentos e dois) repositórios provenientes de diversas instituições internacionais identificados no site Re3Data, apenas 8 (oito) utilizam o padrão de metadados da família PROV, sendo que nenhum deles é brasileiro.

Considerando o recorte empírico desta pesquisa, o estudo indica que a análise da confiabilidade dos dados dificilmente se consolidaria com base exclusiva em metadados específicos de registro de proveniência, em virtude de não serem utilizados nos repositórios brasileiros. Os autores reforçam que a “baixa adoção de tais padrões aponta um desafio na gestão de dados de pesquisa, uma vez que a proveniência inadequada pode comprometer a integridade dos dados e a sua utilidade futura” (Da Silva; Arakaki, 2025, p. 102).

Essa preocupação já havia sido destacada por Arakaki e Santos (2021), ao afirmarem que o processo de identificação da proveniência se beneficia do uso de metadados enquanto uma solução que permite que os usuários avaliem com mais precisão os registros sobre um recurso informacional específico, garantindo a ele qualidade e veracidade dos dados. Da mesma forma, já era sabido por esta pesquisadora que, muito provavelmente, os metadados em Premis também não estariam sendo registrados e/ou disponibilizados nos repositórios analisados.

Compreendendo que o conhecimento da proveniência deve priorizar o contexto de criação (Nesmith, 1982), complementado pela história do criador, do documento e de sua custódia (Millar, 2015), o que converge com a noção de “proveniência detalhada” dos Princípios FAIR (Kohwalter, 2021), buscou-se identificar registros que refletissem essas dimensões.

Diante disso, com vistas a categorizar as possíveis informações encontradas nos repositórios adotou-se por referencial o OAIS (CCSDS, 2024, p. 1-15), especificamente

suas categorias de PDI: “Informação de Proveniência”, “Informação de Contexto” e “Informação de Fixidez”.

Embora o modelo OAIS não seja prescritivo nem de uso obrigatório, sua ampla adoção por instituições de preservação digital o torna uma referência adequada para observar o grau de confiabilidade dos dados disponibilizados. Parte-se do princípio de que os dados passíveis de preservação no futuro são criados no presente, portanto, a qualidade da preservação futura depende das ações realizadas no momento atual.

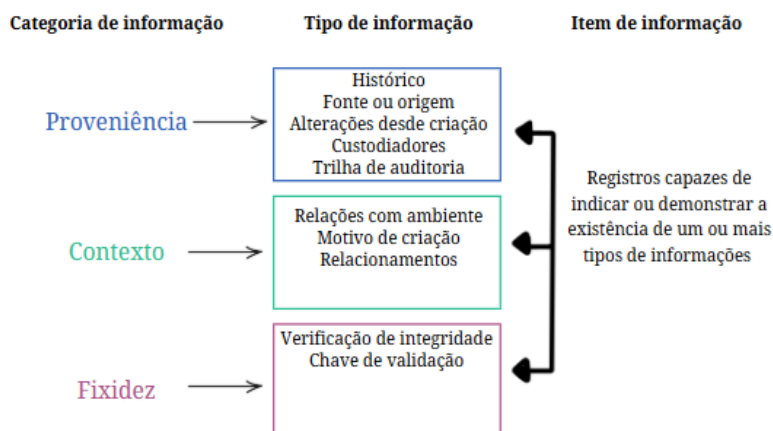
Nessa análise, não foram consideradas as categorias “Informações de Referência” e “Informações de Direitos de Acesso”, ainda que reconhecidamente relevantes à preservação digital. Optou-se por concentrar a análise nas categorias de Proveniência, Contexto e Fixidez (OAIS), por compreender que estas sustentam a aferição da autenticidade e, conseqüentemente, da confiabilidade dos dados de pesquisa digitais.

Isso porque o grau de confiabilidade que pode ser apoiado a partir da aferição da presunção de autenticidade, composta por identidade e integridade do objeto digital, que pode ser acrescida de evidências a partir de registros inerentes à proveniência, em uma abrangência ampliada por contextualizações e mecanismos de manutenção de sua integridade, portanto, acrescida do que se entende no OAIS englobado como informações de contexto e fixidez.

Com o objetivo de estabelecer parâmetros métricos para a análise quanti-qualitativa, elaborou-se uma hierarquização composta por três níveis informacionais: (1) Categoria de informação; (2) Tipo de informação; e (3) Item de informação. O primeiro nível, denominado “Categoria”, corresponde às três categorias selecionadas da PDI do modelo OAIS (Proveniência, Contexto e Fixidez). O segundo, denominada “Tipo”, abrange os termos-chave identificados na definição OAIS de cada categoria. O terceiro, “Item”, representa a menor unidade divisível: os metadados que evidenciem determinado tipo de informação.

A adoção das denominações “Tipo” e “Item” visa aproximar a estrutura analítica dos conceitos arquivísticos de tipo documental, enquanto conjunto de documentos com características comuns de conteúdo e forma, e item documental, entendido como a menor unidade intelectual indivisível (Arquivo Nacional, 2005)

Figura 34 – Hierarquização das informações a serem buscadas



Fonte: Elaborada pela autora (2025).

Tendo a compreensão desse encadeamento da Figura 34 e com vistas a coletar os insumos informacionais pretendidos, foram elaborados instrumentos para a coleta de dados. Primeiramente, foi construída uma tabela-base para nortear a identificação e possibilitar quantificação de tipos informacionais encontrados para uma somatória de verificação.

Quadro 8 – Quadro-base para quantificação de tipos informacionais

Categorias de Informações			
Registram os tipos	Proveniência	Contexto	Fixidez
	Histórico	Relações com ambiente	Verificação de integridade ou Chaves de validação
	Fonte ou origem	Motivo de criação	
	Alterações desde sua criação	Relacionamentos	
	Custodiadores		
	Trilhas de auditoria		
Tipos básicos	5	3	1

Fonte: Elaborada pela autora (2025).

Conforme disposto no Quadro 8, para verificar o registro de proveniência, há 5 (cinco) tipos de informação a serem considerados, já o registro de contexto pode estar composto por 3 (três) tipos e o de fixidez, 1 (um) tipo que, mesmo não necessariamente estejam intitulados dessa forma, apresentem informações sobre essas naturezas. Aponta-se que nos casos em que seja identificado metainformação/item informacional com

natureza encaixável em mais de um tipo, estes devem ser considerados a partir de sua aplicação/quantificação em todas as alternativas possíveis.

Assim, com base na definição dos tipos de informação a serem analisados sob os parâmetros acima estabelecidos, foi criada um instrumento de coleta para registro dos itens de informação identificados, se constituindo um instrumento a ser aplicado na verificação de cada um dos *datasets*.

Quadro 9 – Modelo de instrumento de coleta para registro de itens de informação identificados por *dataset*

Categoria de informações	Quantitativo de tipos observados		
	Ausente	Parcial	Integral
Proveniência			
Contexto			
Fixidez			

Fonte: Elaborado pela autora (2025).

Caso houvesse a identificação de ao menos um item que corresponda a determinado tipo de informação, por exemplo, há no *dataset* o registro do produtor (item), que indica fonte ou origem (tipo) e, por consequência, contribui na análise da proveniência (categoria), esse item será computado e, assim, será identificado que 1 tipo está presente, daí por diante.

Para compor o critério para análise qualitativa, foi desenvolvido um esquema de nivelamento para aposição de status por escala no qual a identificação das informações buscadas nos *datasets* foi qualificada em: **Ausente (A)**, **Parcial (P)** ou **Integral (I)**, com base na seguinte quantificação:

Quadro 10– Níveis de identificação para aposição de status

Categoria de informações	Quantitativo de tipos observados		
	Ausente	Parcial	Integral
Proveniência	0	$1 < x = 4$	5
Contexto	0	$1 < x = 2$	3
Fixidez	0		1

Fonte: Elaborada pela autora (2025).

Com base nesse parâmetro escalonado no Quadro 10, estabeleceu-se que a classificação dos registros de informação seguiria três níveis: ausente, parcial ou integral.

No que se refere à proveniência, o registro foi considerado ausente quando não foram encontrados, no *dataset*, metadados que descrevessem o histórico anterior ao depósito no *dataverse*, tais como informações sobre fontes ou origem dos dados, alterações realizadas desde sua criação, identificação de custodiadores ou trilhas de auditoria prévias. Caso identificados de um a quatro desses tipos de informação, o registro foi classificado como parcial; e, se os cinco tipos estiverem presentes, foi considerado integral.

Para a categoria contexto, o registro foi avaliado como ausente quando não houve informações sobre o ambiente de origem, a motivação da criação dos dados ou suas relações com outros objetos informacionais. Se encontrados um ou dois tipos de informação, foi considerado parcial; e, se os três tipos forem identificados, classificado como integral.

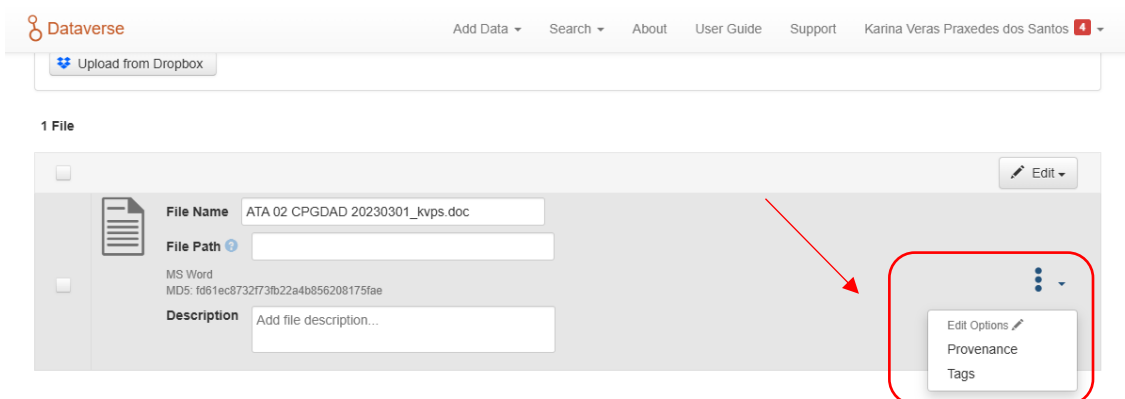
No caso das informações de fixidez, a ausência de registros de verificação de integridade ou de chaves de validação referentes a auditorias anteriores ao depósito no *dataverse* caracteriza o registro como ausente. Não se aplica a classificação parcial, pois a presença de qualquer um desses elementos já é suficiente para considerá-lo integral.

6.1.2 Critérios de análise

Para o estabelecimento dos critérios de análise, foi necessário compreender os termos, recursos, fluxos e funcionalidades disponíveis no Dataverse Project. Assim, foram realizadas leituras do *User Guide* (Guia de usuários) e testes na versão Demo (v.6.6), com objetivo de compreender o comportamento do sistema quanto à estrutura de metadados e às opções de registro de proveniência.

Inicialmente, com foco nas funcionalidades relevantes para a análise proposta, verificou-se no Guia de Usuários (posteriormente confirmado nos testes na versão Demo) que o sistema permite registrar informações de proveniência diretamente no arquivo de dados (nível dado de pesquisa). Para isso, conforme exposto no referido Guia, ao inserir um arquivo e salvá-lo enquanto “*draft*” é possível escolher a opção “*File Options*”, conforme Figura 35:

Figura 35 – Tela Dataverse Project (versão demo) para editar opções do arquivo depositado



Fonte: Dataverse versão Demo.

Ao clicar em “*Provenance*”, nos três pontos laterais ao arquivo, o sistema reforça que proveniência “é um registro da origem do seu arquivo de dados e de quaisquer transformações pelas quais ele passou”, ou seja, trata-se de um registro de onde vieram os dados e como eles chegaram à sua forma atual, descrevendo a origem de um arquivo de dados, suas transformações e quaisquer pessoas ou organizações associadas a esse arquivo, podendo ajudar na reprodutibilidade e para estar em conformidade com os regulamentos legais¹⁰⁰.

De acordo com o Guia de usuários, quando configurado para dar suporte à proveniência, o Dataverse pode contribuir para acompanhar a proveniência dos dados, para tal são proporcionadas duas formas de registrá-la.

A primeira é um “arquivo de proveniência”, considerada a maneira preferencial de enviar informações de proveniência para uma instalação do *dataverse* visto que fornece um registro detalhado e confiável, a partir da seleção de um arquivo JSON. Conforme disposto ao entrar nessa opção do *dataverse*: “Carregue um arquivo JSON de uma ferramenta de captura de proveniência para gerar um gráfico da proveniência dos seus dados”, o que implica a necessidade, então, de uso de um recurso tecnológico para sua extração¹⁰¹. Relevante apontar que esse arquivo de proveniência torna-se permanente, não podendo ser substituído, removido ou editado, garantindo assim seu registro estável e imutável do histórico do arquivo de dados.

¹⁰⁰ Guia do usuário, disponível em: <https://guides.dataverse.org/en/6.6/user/dataset-management.html#data-provenance>. Acesso em: 10 set. 2025

¹⁰¹ Para geração desse arquivo normalmente durante o processo de análise de dados são utilizadas ferramentas de captura de proveniência como, por exemplo, provR, RDataTracker, NoWorkflow, recordr ou CamFlow NoWorkflow. Kohwalter (2021) cita também VisTrails, Kepler, Taverna.

A segunda é “Descrição da proveniência”, de livre escrita, onde o depositante pode descrever as informações relevantes sobre a proveniência dos dados em questão. Esse campo pode ser utilizado para adicionar mais informações de proveniência a um “arquivo de proveniência” ou no lugar desse arquivo, caso não seja possível adicioná-lo. Nesse campo é possível descrever fatores de proveniência, como o sistema operacional utilizado para trabalhar com o arquivo de dados, quais funções ou bibliotecas foram usadas, como os dados foram mesclados no arquivo, qual versão do arquivo foi utilizada, entre outros. Cabe destacar que, segundo consta no próprio Guia, essa descrição não é tão útil ou confiável quanto um “arquivo de proveniência”, mas ainda pode fornecer valor, podendo ser editada, removida ou substituída a qualquer momento.

Figura 36 – Registro de proveniência no Dataverse

Fonte: Dataverse Project versão Demo.

Essas duas modalidades demonstram que o Dataverse reconhece a importância da proveniência; contudo, sua aplicação depende de configurações adicionais e ainda está restrita aos usuários com permissões de edição.

Dessa forma, foi preciso buscar outros caminhos que pudessem levar à identificação de registros relativos às categorias pretendidas. Diante disso, considerou-se que a verificação dos metadados disponibilizados seria o caminho mais viável para alcançar os objetivos da análise. Nesse intuito, foi reconhecido que os *datasets* seguem critérios de preenchimento de metadados previamente estabelecidos na criação do *dataverse* ao qual está vinculado. Faz-se oportuno, então, apresentar como essa definição

se configura no nível de *dataverse*, considerando que todos os *datasets* do recorte empírico possuem essa associação. Para isso, ao criar um *dataverse*, a primeira tela para cadastro apresenta para preenchimento os seguintes campos: Host *dataverse*, Nome do *dataverse*, Identificador, Categoria, E-mail, Afiliação, Descrição e Modo “Livro de visitas”, da seguinte forma:

Figura 37 – Tela de cadastro para criação de um *dataverse*

The screenshot shows the Dataverse registration interface. At the top, there's a navigation bar with links like 'Add Data', 'Search', 'About', 'User Guide', 'Support', and a user profile 'Karina Veras Praxedes dos Santos'. Below the navigation bar, a note states '*Asterisks indicate required fields'. The form fields are arranged in two columns. The left column contains: 'Host Dataverse' (pre-filled), 'Dataverse Name' (required), 'Identifier' (required, pre-filled with a URL), 'Category' (required, dropdown), and 'Email' (required, with a plus icon). The right column contains: 'Affiliation' (required), 'Description' (required, with a note 'This field supports only certain HTML tags.'), and 'Guestbook Mode' (required, dropdown). The 'Guestbook Mode' dropdown is currently set to 'Guestbook Entry At Access Request(inherited from enclosing Dataverse)'.

Fonte: Dataverse Project versão Demo.

Cujas definições são apresentadas ao passar o cursor do mouse por cima das interrogações ao lado de cada campo. Abaixo estão listadas as respectivas definições e destacadas, em fonte azul, todos os campos obrigatórios.

Quadro 11 - Campos e definições da tela de cadastro de um *dataverse*

Campo	Definição
Host dataverse	O Dataverse que contém esses dados
Nome do dataverse	O projeto, departamento, universidade, professor ou periódico para o qual este Dataverse conterá dados
Identificador	Nome curto usado para URL deste Dataverse
Categoria	O tipo que reflete mais de perto este Dataverse
E-mail	O(s) endereço(s) de e-mail do(s) contato(s) do Dataverse
Afiliação	A organização à qual este Dataverse é afiliado
Descrição	Um resumo que descreve a finalidade, a natureza ou escopo deste Dataverse
Modo Livro de visitas	Se os livros de visitas (quando configurados para o conjunto de dados) são exibidos aos usuários quando eles solicitam acesso ao arquivo ou quando baixam arquivos

Fonte: Elaborado pela autora (2025).

O campo categoria apresenta para seleção, em uma barra de rolagem, as opções: Departamento, Diário, Laboratório, Organização ou Instituição, Investigador, Grupo de pesquisa, Projeto de pesquisa, Curso de Docência ou sem categoria.

Nessa mesma tela, são apresentados os campos de metadados a serem usados em modelos de conjuntos de dados e ao adicionar um conjunto de dados a referido *dataverse*. Os seguintes campos são listados para seleção:

Figura 38 – Tela ilustrativa da seleção de metadados para *dataverse*

Metadata Fields

Choose the metadata fields to use in dataset templates and when adding a dataset to this dataverse.

- ☒ Use metadata fields from Demo Dataverse
- ☒ Citation Metadata (Required) [\[+\] View fields](#)
- ☐ Geospatial Metadata [\[+\] View fields](#)
- ☐ Social Science and Humanities Metadata [\[+\] View fields](#)
- ☐ Astronomy and Astrophysics Metadata [\[+\] View fields](#)
- ☐ Life Sciences Metadata [\[+\] View fields](#)
- ☐ Journal Metadata [\[+\] View fields](#)
- ☐ MRA Metadata [\[+\] View fields](#)
- ☐ Graduate School of Design Metadata [\[+\] View fields](#)
- ☐ Alliance for Research on Corporate Sustainability Metadata [\[+\] View fields](#)
- ☐ Political Science Replication Initiative Metadata [\[+\] View fields](#)
- ☐ PSI Metadata [\[+\] View fields](#)
- ☐ CHIA Metadata [\[+\] View fields](#)
- ☐ Digaai Metadata [\[+\] View fields](#)
- ☐ Computational Workflow Metadata [\[+\] View fields](#)
- ☐ Local Contexts [\[+\] View fields](#)
- ☐ 3D Objects Metadata [\[+\] View fields](#)

Fonte: Dataverse Project versão Demo.

São diversas opções de metadados indicados como de domínios, os quais abarcam modelos voltados a formato e especificações informacionais também. No entanto, cabe destacar que somente os **metadados de citação** (*Citation Metadata*) são obrigatórios, os demais são opcionais e podem ser cumulativamente selecionados. Mais adiante, na mesma tela, são dadas opções sobre quais os campos de metadados serão usados como facetas para busca dos conjuntos de dados e *dataverses* no referido *dataverse*.

Figura 39 – Tela ilustrativa da escolha de facetas do Dataverse

Browse/Search Facets Choose the metadata fields to use as facets for browsing datasets and dataverses in this dataverse.

☒ Use browse/search facets from Demo Dataverse

All Metadata Fields
Identifier
Distributor Name
Distribution Date
Time Period Start Date
Time Period End Date
Data Type
Series Name
Geographic Coverage
Country/Region

→

←

↔

↕

Selected
Author Name
Author Affiliation
Subject
Keyword Term
Topic Classification Term
Deposit Date

Fonte: Dataverse Project versão Demo.

Adiante, foi identificado que, ao criar um *dataset*, o sistema abre uma tela para preenchimento das seguintes informações:

Quadro 12 – Campos de preenchimento inicial para cadastro de *dataset*

Campo/subcampo	Categoria de Informação (PDI do OAIS)
Título	Referência
Autor	Proveniência
Nome	Proveniência
Afiliação	Proveniência
Contato	Contexto
Nome	Contexto
Afiliação	Contexto
E-mail	Contexto
Descrição	Contexto
Texto	Contexto
Data	Contexto
Assunto	Contexto
Palavra-chave	Referência
Termo	Referência
URI do termo	Referência
Nome controlado por vocabulário	Referência

Campo/subcampo	Categoria de Informação (PDI do OAIS)
URL do vocabulário controlado	Referência
Publicação relacionada	Contexto
Tipo de relação	Contexto
Citação	Contexto
Tipo de identificador	Contexto
Identificador	Contexto
URL do vocabulário controlado	Contexto
Notas	Proveniência
Depositante	Proveniência
Data do depósito	Proveniência
Subtítulo	Referência
Título Alternativo	Referência
URL alternativo	Contexto
Outro ID	Contexto
Agência	Contexto
Identificador	Contexto
Classificação de tópicos	Contexto
Termo	Contexto
Nome controlado por vocabulário	Contexto
URL do vocabulário controlado	Contexto
Idioma	Referência
Produtor	Proveniência
Nome	Proveniência
Afiliação	Proveniência
Abreviação do nome	Proveniência
URL	Proveniência
URL do logotipo	Proveniência
Data de produção	Proveniência

Campo/subcampo	Categoria de Informação (PDI do OAIS)
Local de produção	Proveniência
Colaborador	Proveniência
Tipo	Proveniência
Nome	Proveniência
Informações de financiamento	Contexto
Agência	Contexto
Identificador	Contexto
Distribuidor	Contexto
Nome	Contexto
Afiliação	Contexto
Abreviação do nome	Contexto
URL do vocabulário controlado	Contexto
Logo da URL	Contexto
Data da distribuição	Contexto
Período de tempo	Proveniência
Data de início	Proveniência
Data de término	Proveniência
Data da coleta	Proveniência
Data de início	Proveniência
Data de término	Proveniência
Tipos de dados	Contexto
Séries	Contexto
Nome	Contexto
Informação	Contexto
Software	Contexto
Nome	Contexto
Versão	Contexto
Material relacionado	Contexto
Dataset relacionado	Contexto
Outras referências	Contexto

Campo/subcampo	Categoria de Informação (PDI do OAIS)
Fonte de dados	Contexto
Origem das fontes históricas	Contexto
Característica das fontes	Contexto
Documentação e acesso a fontes	Contexto

Fonte: Elaborado pela autora (2025).

Onde o título, nome do autor, e-mail de contato, texto de descrição e assunto (em azul na tabela acima) são campos de preenchimento obrigatório¹⁰². Já o depositante é mostrado por *default* a partir do login que foi utilizado para entrar no sistema e a data do depósito é automaticamente preenchida com a data em que o *dataset* está sendo criado. O título pode ser preenchido por replicação. O assunto é mostrado em uma barra de rolagem com 14 assuntos (Medicina, Saúde e Ciências da vida, Ciências sociais, Ciências da Terra e Meio ambiente, Engenharia, Ciências Agrárias, Negócio e Gestão, Computação e Ciência da Informação, Artes e Humanidades, Direito, Química, Ciências Matemáticas, Física, Astronomia e Astrofísica e outros), podendo ser selecionado mais de 1 assunto. No campo publicação relacionada, o tipo de relação pode ser: “é citado por”, “cita”, “é suplementado por”, “é complementado por”, “é referenciado por” e “referências”, já o tipo de identificador da publicação relacionada pode ser escolhido por uma barra de rolagem que apresenta opções como, por exemplo, bibcode, DOI, handle, ISBN, URN, entre outros.

A explicação sobre os campos pode ser acessada ao passar o cursor do *mouse* em cima de pontos de interrogação localizados ao lado de cada um deles. Em suma, para os campos pontuados, são expostas as seguintes definições:

Quadro 13 – Campos e definições da tela de cadastro de um *dataset*

Campo/subcampo	Definição
Título	Título principal do conjunto de dados
Autor	A entidade ou pessoa que criou o conjunto de dados

¹⁰² Cabe apontar que esses metadados obrigatórios são os mínimos que estipulados por *default* pelo sistema. Isso porque, outros metadados podem ser obrigatórios caso tenham sido assim definidos quando na seleção dos metadados do Dataverse ao qual o *dataset* está vinculado.

Campo/subcampo	Definição
Nome	O nome do autor, por ex. o nome da pessoa ou da organização
Afiliação	O nome da entidade afiliada ao autor, por ex. o nome da organização
Contato	A entidade, por ex. uma pessoa ou organização, que os usuários dos conjuntos de dados podem contactar em caso de dúvidas
Nome	O nome do ponto de contato, por ex. o nome da pessoa ou da organização
Afiliação	O nome da entidade afiliada do ponto de contato, por ex. o nome da organização
E-mail	O endereço de e-mail do ponto de contato
Descrição	Um resumo descrevendo a finalidade, a natureza e o escopo do conjunto de dados
Texto	Um resumo descrevendo a finalidade, a natureza e o escopo do conjunto de dados
Data	A data em que a descrição foi adicionada ao conjunto de dados. Se o conjunto de dados contiver mais do que uma descrição, por ex. o produtor de dados forneceu uma descrição e o repositório de dados forneceu outra, esta data é utilizada para distinguir as descrições
Assunto	A área de estudo relevante para o conjunto de dados
Palavra-chave	Um termo chave que descreve um aspecto importante do conjunto de dados e informações sobre qualquer vocabulário controlado usado
Termo	Um termo chave que descreve aspectos importantes do conjunto de dados
URI do termo	Um URI que aponta para a presença na Web do termo de palavra-chave
Nome controlado por vocabulário	O vocabulário controlado usado para o termo da palavra-chave (ex LCSH, MeSH)
URL do vocabulário controlado	A URL onde se pode acessar informações sobre o vocabulário controlado do termo
Publicação relacionada	O artigo ou relatório que usa os dados no conjunto de dados. A lista completa de publicações relacionadas será exibida na guia de metadados
Tipo de relação	A natureza da relação entre este conjunto de dados e a publicação relacionada
Citação	A citação bibliográfica completa para a publicação relacionada
Tipo de identificador	O tipo de identificador que identifica exclusivamente uma publicação relacionada
Identificador	O identificador de uma publicação relacionada
URL do vocabulário controlado	A forma de URL do identificador inserido no campo Identificador, por ex o URL do DOI se um DOI foi inserido nos campos Tipo de ID e Número de ID como um link. Se o que foi inserido no campo Identificador não tiver um formulário URL, o URL da página da publicação será usado, por ex. uma página da Web de um artigo de periódico
Notas	Informações adicionais sobre o conjunto de dados
Depositante	A entidade, como uma pessoa ou organização, que depositou o conjunto de dados no repositório

Campo/subcampo	Definição
Data do depósito	A data em que o conjunto de dados foi depositado no repositório

Fonte: Elaborado pela autora (2025).

Após salvar esse cadastro inicial, ainda em rascunho, é possível que os metadados do *Dataset* sejam editados no campo “Add+Edit Metadata”, conforme figura a seguir, na qual para facilitar a visualização, foi apostado um destaque com elipse amarela:

Figura 40 – Tela de edição de metadados de citação no *dataset*

The screenshot shows a web interface for editing dataset metadata. At the top, there are tabs for 'Files', 'Metadata', 'Terms', and 'Versions'. A yellow arrow points to a button labeled 'Add + Edit Metadata' in the top right corner. Below the tabs, the 'Citation Metadata' section is expanded, showing a list of fields and their values:

Field	Value
Persistent Identifier	doi:10.70122/FK2/PNW4AV
Title	Teste TESE número 4
Author	Veras Praxedes dos Santos, Karina (FIOCRUZ - Fundacao Oswaldo Cruz)
Point of Contact	Use email button above to contact. Veras Praxedes dos Santos, Karina (FIOCRUZ - Fundacao Oswaldo Cruz)
Description	Dataset criado para teses que serão apresentadas na Tese (2025-04-24)
Subject	Demo Only
Keyword	palavra-chave para teste https://URIdoTarmo.com.br (vocabulário controlado) https://URLVocControl.com.br
Related Publication	Is Supplemented By: PRAXEDES,karina
Notes	Anotações para testar a exportação
Depositor	Veras Praxedes dos Santos, Karina
Deposit Date	2025-04-24

Fonte: Elaborada pela autora (2025).

Ao entrar nessa opção, é aberta uma tela onde são apresentados, para além dos campos e subcampos iniciais já registrados (ao menos os obrigatórios), outras possibilidades de metadados para preenchimento, conforme tabela abaixo:

Quadro 14 - Campos adicionais para preenchimento de metadados no *dataset*

Campo	Subcampos
Subtítulo	N/A
Título Alternativo	N/A
URL alternativo	N/A
Outro ID	Agência
	Identificador
Classificação de tópicos	Termo
	Nome controlado por vocabulário
	URL do vocabulário controlado
Idioma	N/A

Campo	Subcampos
Produtor	Nome
	Afiliação
	Abreviação do nome
	URL
	URL do logotipo
Data de produção	N/A
Local de produção	N/A
Colaborador	Tipo
	Nome
Informações de financiamento	Agência
	Identificador
Distribuidor	Nome
	Afiliação
	Abreviação do nome
	URL do vocabulário controlado
	Logo da URL
Data da distribuição	N/A
Período de tempo	Data de início
	Data de término
Data da coleta	Data de início
	Data de término
Tipos de dados	N/A
Séries	Nome
	Informação
Software	Nome
	Versão
Material relacionado	N/A
Dataset relacionado	N/A
Outras referências	N/A
Fonte de dados	N/A
Origem das fontes históricas	N/A
Característica das fontes	N/A
Documentação e acesso a fontes	N/A

Fonte: Elaborado pela autora (2025).

No campo colaborador é possível atrelar o nome do colaborador ao tipo de colaboração, que numa barra de rolagem é apresentada como: curador de dados, gestor de dados, coletor de dados, editor, financiador, instituição de acolhimento, líder de projetos, gerentes de projetos, membro do projeto, pessoa relacionada, investigador,

grupo de pesquisa, detentor dos direitos, patrocinador, supervisor, líder do pacote de trabalho e outro.

Os campos listados acima possuem finalidade de registrar as seguintes informações:

Quadro 15 – Definições dos campos adicionais

Campo/subcampo	Definição
Subtítulo	Um título secundário que amplifica ou declara certas limitações no título principal
Título Alternativo	1) Título comumente usado para se referir ao conjunto de dados ou 2) uma abreviação do título principal
URL alternativo	Outro URL onde se pode visualizar ou acessar os dados no conjunto de dados, por ex. um projeto ou página da Web pessoal
Outro ID	Outro identificador que identifica exclusivamente o conjunto de dados
Agência	O nome da agência que gerou o outro identificador
Identificador	Outro identificador exclusivo para o conjunto de dados, por ex. identificador do produtor ou de outro repositório
Classificação de tópicos	Indica um tópico ou assunto amplo e importante que cobre o conjunto de dados e informações sobre qualquer vocabulário controlado usado
Termo	Um tópico ou termo de assunto
Nome controlado por vocabulário	O vocabulário controlado usado para o termo (ex: LCSH, MeSH)
URL do vocabulário controlado	A URL onde se pode acessar informações sobre o vocabulário controlado do termo
Idioma	O idioma no qual o conjunto de dados foi escrito
Produtor	A entidade, tal pessoa ou organização, que gerencia as finanças ou outros processos administrativos envolvidos na criação do conjunto de dados
Nome	O nome da entidade, por ex, o nome da pessoa ou nome de uma organização - Este campo se tornará obrigatório se for optado por inserir valores em um ou mais desses campos opcionais
Afiliação	O nome da entidade afiliada ao produtor, por ex. o nome de uma organização
Abreviação do nome	O nome abreviado do produtor
URL	A URL do site do produtor
URL do logotipo	A URL do logotipo do produtor
Data de produção	A data em que os dados foram produzidos (não distribuídos, publicados ou arquivados)
Local de produção	O local onde os dados e quaisquer materiais relacionados foram produzidos ou coletados

Campo/subcampo	Definição
Colaborador	A entidade, como uma pessoa ou organização, responsável por coletar, gerenciar ou colaborar de outra forma para o desenvolvimento do conjunto de dados
Tipo	Indica o tipo de contribuição feita para o conjunto de dados
Nome	O nome do colaborador, por ex. o nome da pessoa ou o nome da organização
Informações de financiamento	Informações sobre o apoio financeiro do conjunto de dados
Agência	A agência que forneceu apoio financeiro para o conjunto de dados
Identificador	O identificador de subvenção ou identificador de contrato da agência que prestou apoio financeiro ao conjunto de dados
Distribuidor	A entidade, como uma pessoa ou organização, designada para gerar cópias do conjunto de dados, incluindo quaisquer edições ou revisões
Nome	O nome da entidade, por ex. o nome da pessoa ou o nome de uma organização
Afiliação	O nome da entidade afiliada ao distribuidor, por ex. o nome de uma organização
Abreviação do nome	O nome abreviado do distribuidor
URL do vocabulário controlado	A URL da página da Web do distribuidor
Logo da URL	A URL da imagem do logotipo do distribuidor, usada para mostrar a imagem na página do conjunto de dados
Data da distribuição	A data em que o conjunto de dados foi disponibilizado para distribuição/apresentação
Período de tempo	O período de tempo ao qual os dados se referem. Também conhecido como span. Este é o período de tempo coberto pelos dados, não as datas de codificação, coleta de dados ou para tornar os documentos legíveis por máquinas
Data de início	A data de início do período de tempo ao qual os dados se referem
Data de término	A data de término do período de tempo ao qual os dados se referem
Data da coleta	As datas em que os dados foram gerados ou coletados
Data de início	A data em que a coleta de dados começou
Data de término	A data em que a coleta de dados foi encerrada
Tipos de dados	O tipo de dados incluídos nos arquivos, por ex. dados de página, dados clínicos ou texto legível por máquina
Séries	Informações sobre a série de conjunto de dados à qual o conjunto de dados pertence
Nome	O nome da série de conjunto de dados
Informação	Pode incluir 1) histórico da série e 2) um resumo dos recursos que se aplicam à série

Campo/subcampo	Definição
Software	Informações sobre o software usado para gerar o conjunto de dados
Nome	O nome do software usado para gerar o conjunto de dados
Versão	A versão do software usado para gerar o conjunto de dados, por ex 4.11
Material relacionado	Informações, como um ID persistente ou citação, sobre o material relacionado ao conjunto de dados, como apêndices ou informações de amostragem disponíveis fora do conjunto de dados
Dataset relacionado	Informações, como um ID persistente ou citação, sobre um conjunto de dados relacionado, como pesquisas anteriores sobre o assunto do conjunto de dados
Outras referências	Informações, como um ID persistente ou citação, sobre outro tipo de recurso que fornece informações básicas ou material de apoio ao conjunto de dados
Fonte de dados	Informações, como um ID persistente ou citação, sobre as fontes do conjunto de dados (por ex. um livro, artigo, série ou arquivo de dados legível por máquina)
Origem das fontes históricas	Para fontes históricas, a origem e quaisquer regras seguidas para estabelecê-las como fonte
Característica das fontes	Características ainda não observadas em outro lugar
Documentação e acesso a fontes	1) Métodos ou procedimentos para acessar fontes de dados e 2) quaisquer permissões especiais necessárias para acesso

Fonte: Elaborado pela autora (2025).

A partir dessa verificação, e tomando como referência as categorias de informação do OAIS (2024), elaborou-se o enquadramento dos metadados observados, conforme quadro 16. É relevante reconhecer que as categorias de informações das PDI são compostas e podem ser trabalhadas a partir de diferentes tipos de metadados descritivos, administrativos e estruturais. Além disso, é recomendada a utilização de padrões e elementos de esquemas, registrados e mantidos persistentemente junto aos dados, os quais podem permear distintas categorias ao mesmo tempo.

Portanto, embora basicamente os campos disponibilizados sejam voltados ao registro de metadados descritivos, o seguinte parâmetro de observação foi constituído no intuito de categorizar as informações relativas aos conjuntos de dados e aos próprios dados a partir de possíveis correlações com a compreensão semântica sobre cada uma das categorias de informações da PDI. Dessa forma, basicamente, o sistema apoia o registro de metadados que podem ser categorizados da seguinte forma:

Quadro 16– Categorização dos metadados dos *datasets* à luz das PDI do Modelo OAIS

Campo/subcampo	Categoria de Informações (PDI do OAIS)
Título	Referência
Autor	Proveniência
Nome	Proveniência
Afiliação	Proveniência
Contato	Contexto
Nome	Contexto
Afiliação	Contexto
E-mail	Contexto
Descrição	Contexto
Texto	Contexto
Data	Contexto
Assunto	Contexto
Palavra-chave	Referência
Termo	Referência
URI do termo	Referência
Nome controlado por vocabulário	Referência
URL do vocabulário controlado	Referência
Publicação relacionada	Contexto
Tipo de relação	Contexto
Citação	Contexto
Tipo de identificador	Contexto
Identificador	Contexto
URL do vocabulário controlado	Contexto
Notas	Proveniência
Depositante	Proveniência
Data do depósito	Proveniência
Subtítulo	Referência
Título Alternativo	Referência

Campo/subcampo	Categoria de Informações (PDI do OAIS)
URL alternativo	Referência
Outro ID	Referência
Agência	Referência
Identificador	Referência
Classificação de tópicos	Contexto
Termo	Contexto
Nome controlado por vocabulário	Contexto
URL do vocabulário controlado	Contexto
Idioma	Referência
Produtor	Proveniência
Nome	Proveniência
Afiliação	Proveniência
Abreviação do nome	Proveniência
URL	Proveniência
URL do logotipo	Proveniência
Data de produção	Proveniência
Local de produção	Proveniência
Colaborador	Proveniência
Tipo	Proveniência
Nome	Proveniência
Informações de financiamento	Contexto
Agência	Contexto
Identificador	Contexto
Distribuidor	Contexto
Nome	Contexto
Afiliação	Contexto
Abreviação do nome	Contexto
URL do vocabulário controlado	Contexto
Logo da URL	Contexto

Campo/subcampo	Categoria de Informações (PDI do OAIS)
Data da distribuição	Contexto
Período de tempo	Proveniência
Data de início	Proveniência
Data de término	Proveniência
Data da coleta	Proveniência
Data de início	Proveniência
Data de término	Proveniência
Tipos de dados	Contexto
Séries	Contexto
Nome	Contexto
Informação	Contexto
Software	Contexto
Nome	Contexto
Versão	Contexto
Material relacionado	Contexto
Dataset relacionado	Contexto
Outras referências	Contexto
Fonte de dados	Contexto
Origem das fontes históricas	Contexto
Característica das fontes	Contexto
Documentação e acesso a fontes	Contexto

Fonte: Elaborada pela autora (2025).

Por se referir à análise do todo mediante visão do panorama generalista oferecida pela definição dos campos, é preciso considerar que um mesmo campo poderia ser enquadrado em mais de uma categoria de PDI à luz do delineamento do Modelo OAIS. Além disso, a fixação de uma categoria deve ser traduzida a partir da observação do preenchimento factual dos referidos campos a fim de que consubstanciem uma visão concreta e realista, sendo necessária a análise dos conteúdos informacionais.

Assim, identificaram-se 46 campos relacionados ao contexto (59%), 23 à proveniência (29%) e 9 à referência (12%). Não foram observados campos relativos às categorias de fixidez ou direitos de acesso

Cabe reforçar que outras relações de metadados específicos por domínios, formatos ou especificações também podem ser agrupadas ao rol de informações do *dataset*, dependendo do cadastro do *dataverse* ao qual está vinculado, o que pode vir a incrementar a categorização acima apresentada.

Após conhecer o que poderia ser extraído para consubstanciar a análise quanti-qualitativo a partir da verificação dos metadados de citação, mostrou-se preciso cercar outras possibilidades funcionais do sistema que, de alguma forma, pudessem transparecer indícios informacionais sobre proveniência, contexto e fixidez.

Diante disso, ao verificar que após a publicação do conjunto de dados seus metadados podem ser exportados em vários padrões e formatos de metadados, julgou-se oportuno acrescentar esse caminho para enrobustecer o critério de análise.

Esses “metadados de exportação”, segundo disposto no Guia de usuário, tem por objetivo possibilitar que esses recursos tornem-se mais detectáveis e utilizáveis em outros sistemas, como outros repositórios de dados e, para isso, as seguintes exportações estão disponíveis na guia “metadados” em cada conjunto de dados: Dublin Core¹⁰³, DDI (Data Documentation Initiative Codebook 2.5)¹⁰⁴, DDI HTML Codebook¹⁰⁵ (uma versão HTML mais legível do DDI Codebook 2.5 metadata export), DataCite 4¹⁰⁶, JSON¹⁰⁷ (formato nativo do software Dataverse), OAI_ORE¹⁰⁸, OpenAIRE¹⁰⁹ e Schema.org JSON-LD¹¹⁰.

Figura 41 – Metadados de Exportação no Dataverse¹¹¹

¹⁰³ Mais informações sobre o Dublin Core em: <https://www.dublincore.org/>

¹⁰⁴ Mais informações sobre DDI em: https://ddialliance.org/ddi-codebook_v2.5

¹⁰⁵ Mais informações sobre DDI HTML Codebook em: <https://docs.ddialliance.org/DDI-Lifecycle/guide/Overview/Field%20Level/>

¹⁰⁶ Mais informações sobre DataCite 4 em: <https://schema.datacite.org/meta/kernel-4.4/>

¹⁰⁷ Mais informações sobre JSON em: <https://www.json.org/json-en.html>

¹⁰⁸ Mais informações sobre OAI-ORE em: <https://www.openarchives.org/ore/>

¹⁰⁹ Mais informações sobre OpenAIRE em: <https://www.openaire.eu/>

¹¹⁰ Mais informações sobre JSON-LD (JSON for Linking Data) em: <https://json-ld.org/>

¹¹¹ As informações que poderiam servir para identificação foram tarjadas com uma figura oval cinza.



Fonte: Dataverse Project versão Demo.

Essas opções de exportação ampliam o potencial de interoperabilidade dos metadados entre repositórios e sistemas externos, fortalecendo a rastreabilidade e a transparência das informações de pesquisa. Isso porque, embora, atualmente, ainda não seja possível que os usuários tenham acesso público a informações especificamente registradas como “de proveniência”, buscou-se confirmar se essas informações poderiam estar incorporadas em outros formatos de exportação, visto sua relevância para (re)uso.

É preciso compreender que, para agregar valor, em termos de confiabilidade, aos dados de pesquisa, faz-se oportuno e relevante o uso de ferramentas capazes de registrar procedimentos e processos desde sua criação, de forma que, ao serem depositados para compartilhamento, carreguem consigo todas essas informações em forma de metadados.

Isso pode ser mais acessível econômica, estrutural e administrativamente em projetos com altos investimentos e/ou com coleta/criação/transformação de grandes volumes de dados. Embora ferramentas de captura automatizada de proveniência sejam mais viáveis em projetos com altos investimentos, pequenas equipes e laboratórios independentes também têm produzido coleções de dados relevantes, que não devem ser negligenciadas, a chamada “cauda longa da ciência” (Sayão; Sales, 2018).

Assim, em busca de outras formas para viabilizar a ação desses registros sem o uso de ferramentas tecnológicas sofisticadas e com intuito de torná-los compreensíveis também por humanos, estratégias de registros contextuais estão sendo aplicadas, como a inserção de arquivos *Read.me*¹¹² junto aos dados disponibilizados. Ainda que não haja um padrão de estrutura ou formato para sua elaboração, esses arquivos de texto servem

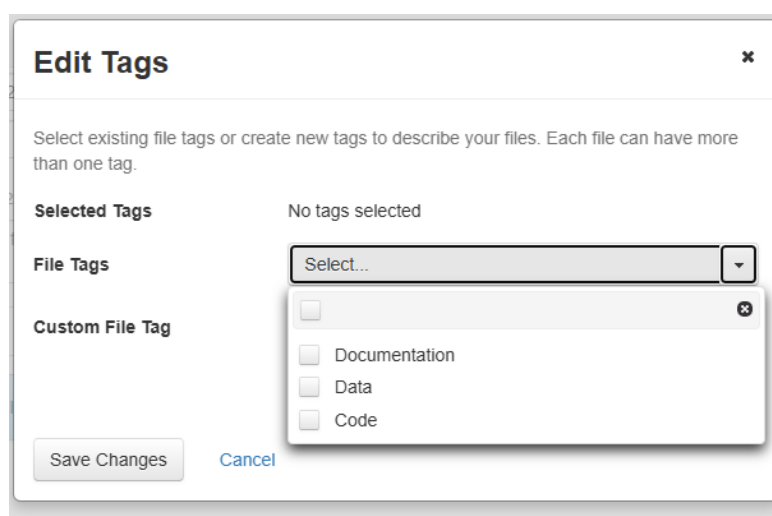
¹¹² No makeareadme.com, junto à explicações sobre o que é um arquivo *Read.me*, além do porquê, por quem, quando deve ser feito e onde deve ser colocado, o site logo de início faz uma alusão interessante à escolha de seu nome “Make a README because no one can read your mind (yet)” (Faça um Readme. Porque ninguém pode ler sua mente (ainda)).

para apresentar e explicar o objeto digital ao qual estão vinculados e contém informações necessárias para entendê-lo¹¹³.

Mesmo que em alguns casos esses arquivos possam se apresentar de uma forma menos estruturada, neles é possível que sejam descritas desde informações sobre a coleta, o processamento dos dados, formas de gestão e manutenção empregadas, até os *softwares* utilizados para tal, métodos aplicados e outras informações contextuais e/ou técnicas julgadas relevantes¹¹⁴.

Outro recurso interessante observado nesse *software*, quando no processo de carregamento dos arquivos (dados de pesquisa), é a possibilidade de inserção de *tags* para descrever e especificar se o arquivo é relativo a alguma categoria como *documentation*, *data* e *code*, podendo ser customizadas outras *tags*, caso seja oportuno.

Figura 42 – Edição de *tags* no Dataverse



Fonte: Dataverse Project versão Demo.

O uso de arquivos *Read.me* e de *tags* descritivas (*data*, *code*, *documentation*) contribui para o registro contextual e para a distinção funcional dos arquivos dentro de cada *dataset*, favorecendo a transparência e a compreensão dos dados por humanos e máquinas. A aposição de especificações distintas contribui, por exemplo, na identificação do arquivo que é o dado de pesquisa em si (*data*), o arquivo que corresponde à sua

¹¹³ Disponível em: makeareadme.com

¹¹⁴ No site do makeareadme.com, é exposto que “Embora um *readme* possa ser muito longo e detalhado, é melhor ser muito longo do que muito curto. Caso esteja longo demais, é possível que sejam consideradas outras formas de documentação ao invés de cortar informações existentes”.

codificação (*code*), o outro que o documenta para fins de contextualização (*documentation*) e assim por diante.

Para aprofundar a análise pretendida, no entanto, foi preciso que, além de verificar os metadados disponíveis, fossem também analisados cada um dos arquivos de dados compartilhados dentro dos conjuntos.

Para tanto, optou-se pelos seguintes passos de análise, na ordem apresentada:

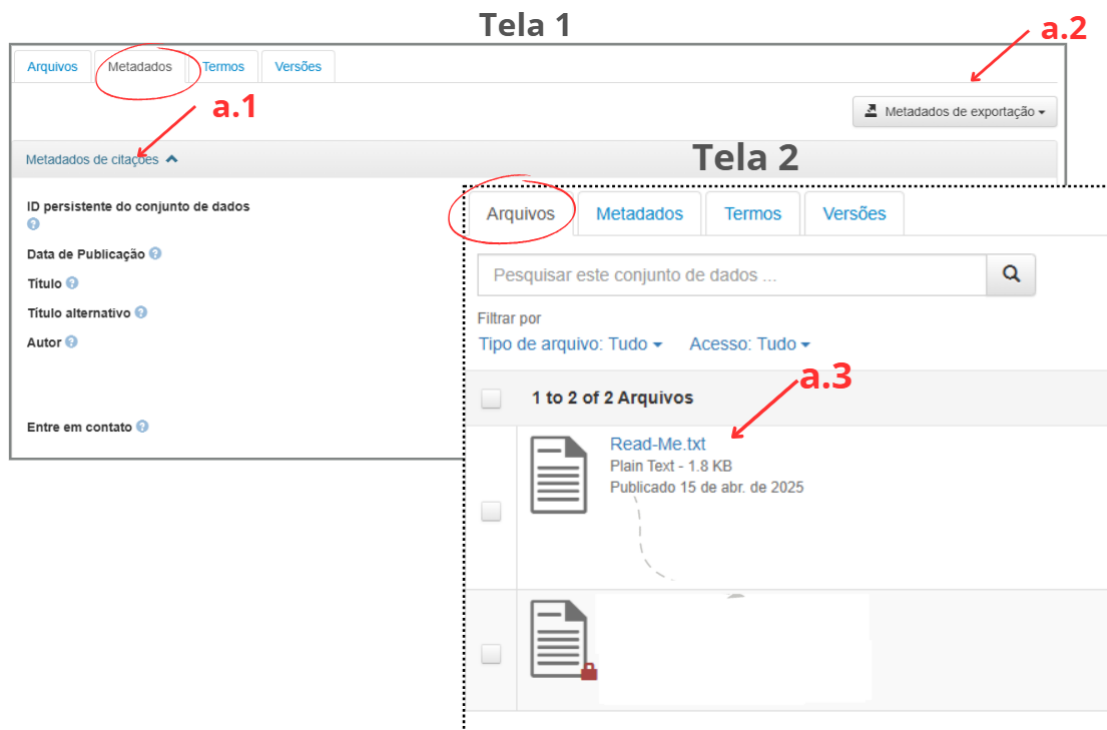
a) **À luz da análise do conjunto de dados (*dataset*)**

a.1) Metadados de citação e de domínios: em busca de informações que possam compor as categorias de proveniência e contexto;

a.2) Metadados de exportação: com objetivo de verificar a existência de informações das categorias de proveniência, contexto e fixidez que não estejam manifestadas de forma transparente e simplificada para o usuário;

a.3) Arquivos *Read.me*: a existência de arquivos que documentem informações que possam consubstanciar registros de proveniência, contexto e/ou fixidez.

Figura 43 – Telas com ilustração das etapas da análise no nível *dataset*



Fonte: Elaborada pela autora (2025).

b) À luz da análise de cada um dos arquivos de dados:

b.1) Configurações do documento: no intuito de verificar informações descritas nas propriedades dos arquivos e, quando não for possível este caminho, identificar os itens informacionais, exclusivamente, pela aba “metadados” do arquivo no sistema.

b.2) Uso de *tags* que possam direcionar a observação de arquivos de dados específicos para documentar o dado ou do conjunto de dados de pesquisa.

Figura 44 – Telas com ilustração das etapas da análise no nível Arquivo de dados



Fonte: Elaborada pela autora (2025).

Esses critérios orientaram o processo de análise aplicado aos *datasets* e arquivos de dados, permitindo identificar a presença e a qualidade das informações relativas à proveniência, contexto e fixidez.

6.1.3 Limitação da pesquisa

Ao analisar os repositórios de instituições públicas brasileiras selecionados conforme os critérios previamente definidos, é importante ressaltar que tais ferramentas tecnológicas representam apenas uma parcela do amplo esforço coletivo, institucional e nacional voltado à transformação das práticas de produção, organização,

compartilhamento e reutilização do conhecimento científico. Por trás dessas iniciativas, encontram-se inúmeros profissionais comprometidos com ações de conscientização, sensibilização e capacitação contínuas, que visam possibilitar o (re)uso dos dados e promover uma ciência mais transparente, democrática e socialmente relevante.

Assim, a análise realizada nesta pesquisa não tem por objetivo avaliar a competência, a expertise ou o empenho das instituições e dos profissionais envolvidos no ecossistema da curadoria digital de dados de pesquisa. Ao contrário, o simples fato de seus repositórios já se encontrarem disponíveis para análise demonstra que avanços significativos foram alcançados. O propósito, portanto, é identificar possibilidades de aprimoramento das práticas já em execução e oferecer subsídios que possam apoiar instituições ainda em fase inicial de implementação, a partir de parâmetros mínimos que orientem o desenvolvimento de suas políticas e fluxos de gestão.

O estudo empírico concentrou-se na análise de repositórios que utilizam o *software* Dataverse Project, considerando um número delimitado de *datasets* selecionados com base em um recorte temporal e temático que permitisse o aprofundamento da análise dentro das restrições de tempo e escopo. Ressalta-se que a pesquisa não tem caráter avaliativo em relação ao sistema, mas sim exploratório, de modo a compreender suas funcionalidades disponíveis sob o ponto de vista de uma arquivista usuária comum, com recursos informacionais limitados, porém adequados aos objetivos estabelecidos.

Os testes foram realizados na versão *demo* do Dataverse, a qual pode apresentar diferenças em relação às versões customizadas utilizadas institucionalmente. Essa etapa teve como finalidade compreender o fluxo e as funções básicas relacionadas ao registro e à estruturação de metadados.

Como exposto anteriormente, o recorte empírico foi composto por 24 *dataverses*, a partir dos quais foram selecionados 24 *datasets* contendo 184 arquivos de dados no total. Com base nos campos disponíveis para preenchimento no sistema, procedeu-se à coleta dos dados necessários para as análises quanti-qualitativa. A verificação dos *datasets* selecionados nos 3 (três) repositórios revelou o uso predominante dos metadados de citação, complementados por alguns metadados de domínio: Geoespaciais e Ecológicos, para descrição dos conjuntos de dados.

Essas constatações se refletem na lista de metadados efetivamente encontrados, conforme Quadro 17.

Quadro 17 – Listagem de metadados encontrados nos *datasets*

Metadados encontrados		
Citação	Geoespacial	Ecológica
ID persistente	Cobertura geográfica	Cobertura geográfica
Data de publicação	Caixa delimitadora geográfica	Cobertura temporal
Título		Cobertura taxonômica
Subtítulo		Classificação Taxonômica
Título Alternativo		Métodos
URL alternativo		
Outro ID		
Contato		
Descrição		
Assunto		
Palavra-chave		
Publicação relacionada		
Classificação de tópicos		
Idioma		
Produtor		
Data de produção		
Conceder Informação		
Notas		
Depositante		
Data do depósito		
Data da coleta		
Período coberto		
Documentação e acesso a fontes		
Tipos de dados		
Software		

Fonte: Elaborado pela autora (2025).

Assim, diante da compreensão dos metadados que foram utilizados para serem a base informacional dos *datasets* do recorte, observou-se que dos 31 campos de metadados utilizados¹¹⁵, a maioria corresponde aos metadados de citação, tendo sido utilizado alguns poucos campos para preenchimento específico do domínio geoespacial e ecológico. Deles, 26 campos (84%) correspondem ao rol padrão (*default*) de metadados de citação

¹¹⁵ Embora o campo “Cobertura geográfica” esteja apresentado 2 vezes, tanto como campo do domínio geoespacial quanto do geológico, por se tratar de um campo com intuito de preenchimento sobre uma mesma informação, optou-se por quantificá-la como 1 campo.

do sistema, enquanto apenas 5 campos (17%) foram acrescentados. Essa distribuição indica que os elementos descritivos fundamentais seguem o modelo predefinido do sistema, ao passo que as customizações por domínio ainda ocorrem em escala reduzida.

Conforme apresentado anteriormente, é importante expor que esses campos de metadados por domínio são definidos da seguinte forma:

Quadro 18 – Campos e definições dos metadados por domínio

Metadados por domínio	Campo	Definição
Ecológica	Cobertura geográfica	Contêiner para informações espaciais sobre um projeto, um recurso ou um entidade dentro de um recurso. Ela possibilita a definição de uma caixa delimitadora para a cobertura global, expressa em coordenadas de latitude e longitude
	Cobertura temporal	Este campo especifica a cobertura temporal e permite que ela seja um único ponto no tempo, vários pontos no tempo ou um intervalo de datas. As datas podem ser expressas em termos de datas calendárias e sistemas de datação geológica
	Cobertura taxonômica	Recipiente para informações taxonômicas. Inclui lista de nome de espécies (ou classificações de nível superior) de um ou mais sistemas de classificação
	Classificação taxonômica	Informações sobre a variedade de taxa abordados no conjunto de dados ou coleção
	Métodos	O campo de métodos documenta os métodos científicos utilizados na coleta deste conjunto de dados. Inclui informações sobre itens como ferramentas, calibração de instrumentos e software
Geoespacial	Cobertura geográfica	Informações sobre cobertura geográfica dos dados. Inclui o escopo geográfico total dos dados
	Caixa delimitadora geográfica	A descrição geométrica fundamental para qualquer conjunto de dados que modela a geografia é a caixa delimitadora geográfica. Ele descreve a caixa mínima, definida pelas longitudes oeste e leste e latitudes norte e sul, que inclui a maior extensão geográfica do conjunto de dados. Este elemento é usado na primeira passagem de uma pesquisa baseada em coordenadas. A inclusão desse elemento no livro de códigos é recomendada, mas é necessária se a caixa de polígonos vinculada estiver incluída.

Fonte: Elaborado pela autora (2025).

Com base nessa compreensão preliminar dos metadados registrados, foi possível dar continuidade à verificação do conteúdo informacional, de forma mais aprofundada e específica, nos *datasets* e arquivos de dados selecionados.

6.2 ANÁLISE DO CAMPO EMPÍRICO: IDENTIFICAÇÃO E CATEGORIZAÇÃO DAS INFORMAÇÕES ENCONTRADAS

Esta seção apresenta a análise do campo empírico, conduzida com o propósito de examinar de forma sistemática o conteúdo informacional disponibilizado nos repositórios selecionados, buscando identificar, categorizar e avaliar as informações que subsidiam a presunção de autenticidade dos dados de pesquisa. Essa análise parte da observação tanto dos *datasets* (conjuntos de dados) quanto dos arquivos que os compõem (dados em si), considerando, sobretudo, os elementos registrados nos metadados de citação, de domínio e de exportação e o uso de recursos complementares, como arquivos *Read.me* e *tags*.

A partir desses componentes, pretende-se verificar em que medida as informações disponibilizadas refletem práticas de registro de metadados que permitam inferir aspectos relacionados à proveniência, ao contexto e à fixidez, dimensões fundamentais para a qualificação da preservação digital e para o (re)uso confiável dos dados de pesquisa.

6.2.1 À luz da análise do conjunto de dados (*dataset*)

Para analisar o campo empírico, nessa subseção são identificadas e categorizadas as informações encontradas à luz da análise do conjunto de dados (*dataset*), a partir da observação dos metadados de citação, de domínios e exportação e o uso do arquivo *Read.me*.

6.2.1.1 Metadados de citação e de domínios

Inicialmente, com utilização do quadro 8 “Quadro-base para quantificação de tipos informacionais”, criada para categorização dos metadados dos *datasets* à luz das PDI do Modelo OAIS, foram consolidados abaixo os campos de metadados de citação e de domínios encontrados nos 24 *datasets* do recorte. Além disso, foram empregados os seguintes destaques, tendo em vista o escopo dessa análise:

a) em azul, os campos identificados como passíveis de registrar informações sobre **proveniência**¹¹⁶;

b) em amarelo, os campos passíveis de registrar informações sobre contexto;

c) em verde, os campos viáveis para registro de informações **referência**; e

d) em rosa, aquele direcionado à informação correspondente a direitos de **acesso**.

Dessa forma, foi possível elencar a categorização abaixo:

Quadro 19– Categorização de informações encontradas nos *datasets*

Campo/subcampo	Categoria de Informação (PDI do OAIS)
ID persistente	Referência
Data de publicação	Proveniência
Título	Referência
Subtítulo	Referência
Título Alternativo	Referência
URL alternativo	Contexto
Outro ID	Contexto
Autor	Proveniência
Contato	Contexto
Descrição	Contexto
Assunto	Contexto
Palavra-chave	Referência
Publicação relacionada	Contexto
Classificação de tópicos	Contexto
Idioma	Referência
Produtor	Proveniência

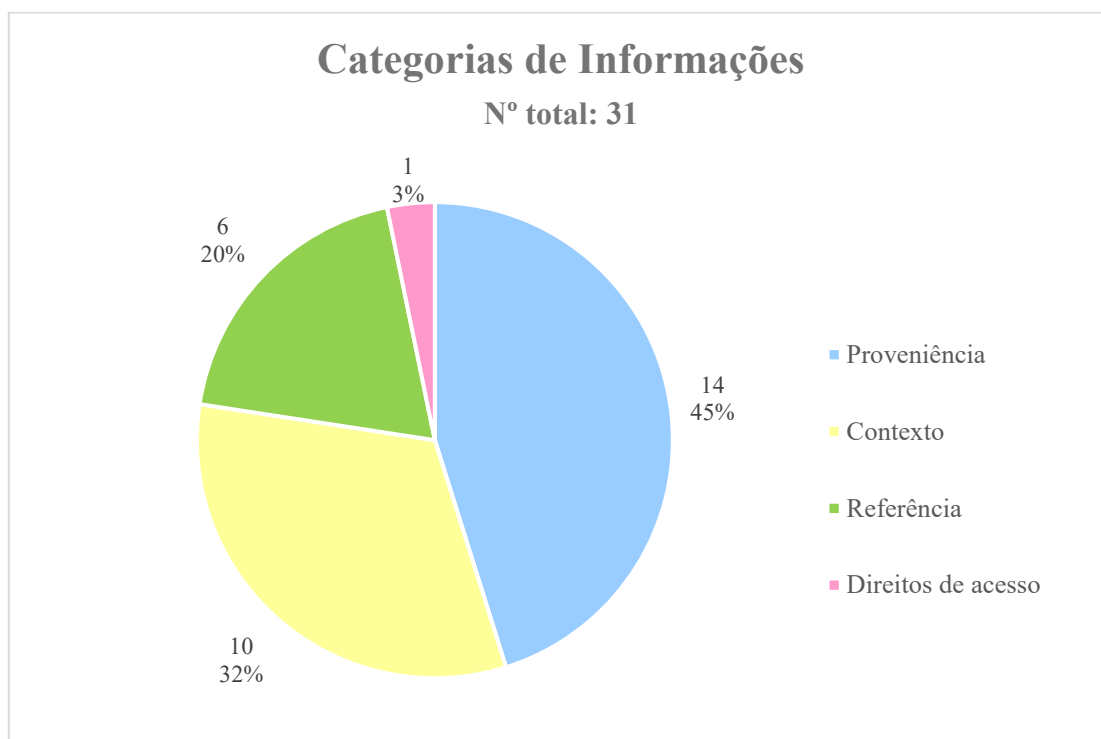
¹¹⁶ Nesse momento, faz-se mister reforçar que a categoria de PDI relativa à informação de proveniência é uma forma especial de informação de contexto, segundo Modelo OAIS (2024). Portanto, para fins deste estudo, foi adotado como critério para a diferenciação entre as duas categorias o seguinte: quando o campo é preenchido com informações contextuais mas que sejam capazes de demonstrar o contexto de produção e manutenção dos conjuntos de dados de pesquisa, foi considerado como correspondente à categoria “proveniência”. Às demais formas de contextos, foi optado por categorizar enquanto contexto.

Campo/subcampo	Categoria de Informação (PDI do OAIS)
Data de produção	Proveniência
Conceder Informação	Direitos de acesso
Notas*	Proveniência
Depositante	Proveniência
Data do depósito	Proveniência
Data da coleta	Proveniência
Período coberto	Proveniência
Documentação e acesso a fontes	Contexto
Tipos de dados	Contexto
<i>Software</i>	Contexto
Caixa delimitadora geográfica	Proveniência
Cobertura geográfica	Proveniência
Cobertura temporal	Proveniência
Cobertura taxonômica	Proveniência
Métodos	Proveniência

Fonte: Elaborada pela autora (2025).

A partir dessa categorização inicial, foi possível aprofundar a análise dos 31 campos verificados. Deles, 14 são passíveis de demonstrar informações sobre proveniência, 10 sobre contextos, 6 sobre referências e 1 sobre direitos de acesso, cuja distribuição apresenta-se conforme gráfico abaixo:

Gráfico 1 – Distribuição de categorias de informações encontradas nos *datasets*



Fonte: Elaborado pela autora (2025).

Embora os 31 campos identificados componham o conjunto completo de metadados disponíveis, somente parte deles é efetivamente capaz de contribuir para a verificação das categorias de informação pertinentes à análise. Por esse motivo, aplicou-se um filtro metodológico, por meio do qual foram desconsiderados alguns campos, conforme listado a seguir:

- a) campos relativos às **categorias de informações de referência** (Título, Subtítulo, Título Alternativo, Palavra-chave e Idioma) e **de direitos de acesso** (Conceder informação);
- b) campos de **preenchimento obrigatório** (Autor, Contato, Descrição e Assunto);
- c) campos **preenchidos automaticamente** (ID persistente e Data de publicação); e
- d) campos que correspondam a **informações do ato de depósito em diante** (Depositante e Data de depósito).

Selecionaram-se, portanto, 17 campos dentre os inicialmente encontrados. Adiante, utilizando a hierarquização definida para fins desta tese, correlacionam-se as categorias, tipos e itens de informação identificados da seguinte forma:

Quadro 20 – Correlação hierárquica dos itens de informação selecionados

Categoria de informação (PDI do OAIS)	Tipo de informação	Itens de informação (campo/subcampo)
Proveniência	Fonte ou origem	Produtor
	Histórico	Data de produção
		Notas
		Data da coleta
		Período coberto
		Cobertura geográfica
		Caixa delimitadora geográfica
		Cobertura temporal
		Cobertura taxonômica
		Métodos
Contexto	Relacionamentos	URL alternativo
		Outro ID
		Publicação relacionada
		Classificação de tópicos
	Relação com ambiente	Documentação e acesso a fontes
		Tipos de dados
		<i>software</i>

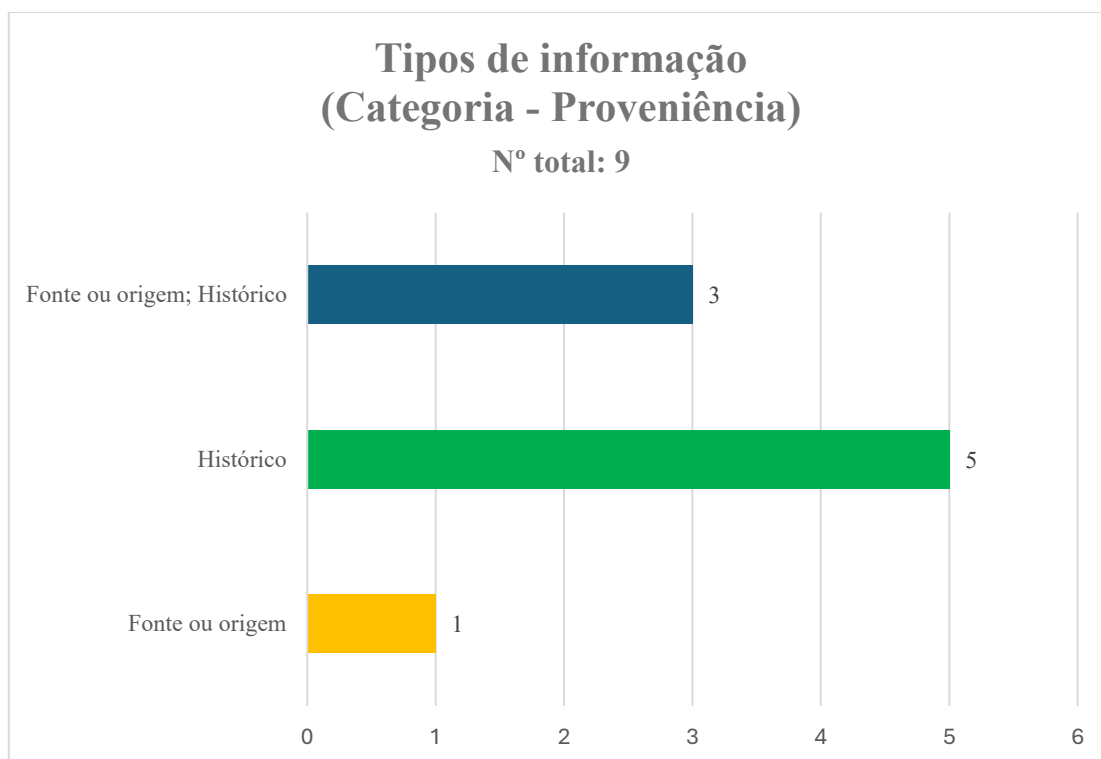
Fonte: Elaborado pela autora (2025).

Com base no modelo de instrumento de coleta para registro de itens de informação (quadro 11) identificados por *dataset* foi possível prosseguir na identificação dos elementos observados no nível informacional, conforme Apêndice A. Portanto, compreendeu-se que:

6.2.1.1.1 Quanto à **Categoria de Informação Proveniência** (que abarca os tipos de informação: Histórico, Fonte ou origem, Alterações desde sua criação, Custodiadores e Trilhas de Auditoria)

Foram encontrados nos *datasets* selecionados, apenas itens correspondentes à **Fonte ou origem** e **Histórico**. A distribuição dos tipos informacionais identificados na categoria Proveniência é apresentada no Gráfico 2.

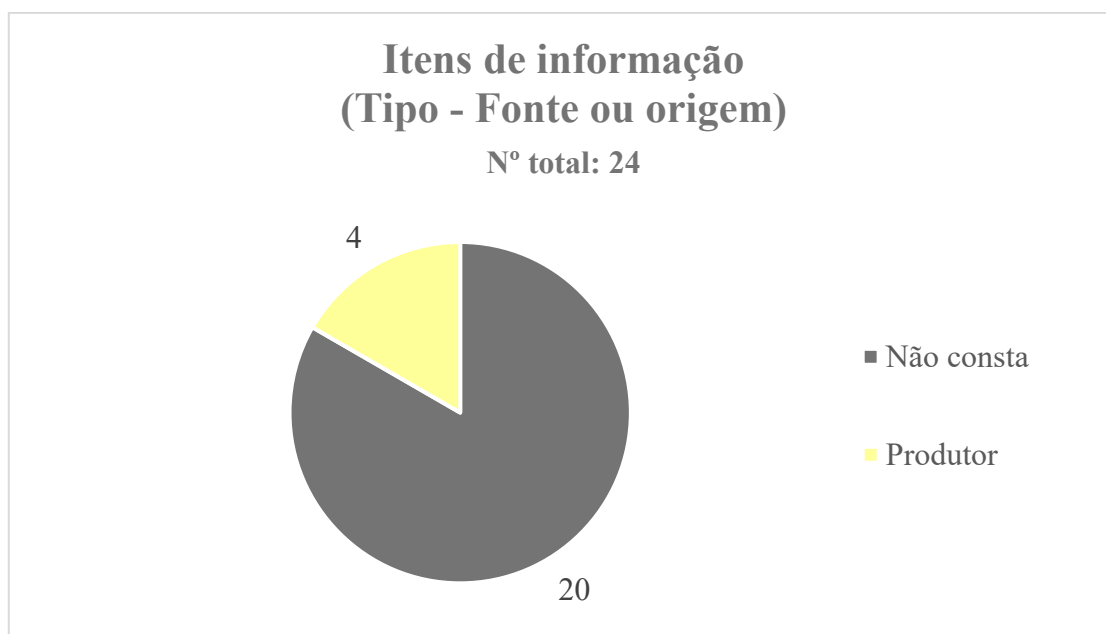
Gráfico 2 – Tipos informacionais da Categoria Proveniência encontrados nos *datasets* selecionados



Fonte: Elaborado pela autora (2025).

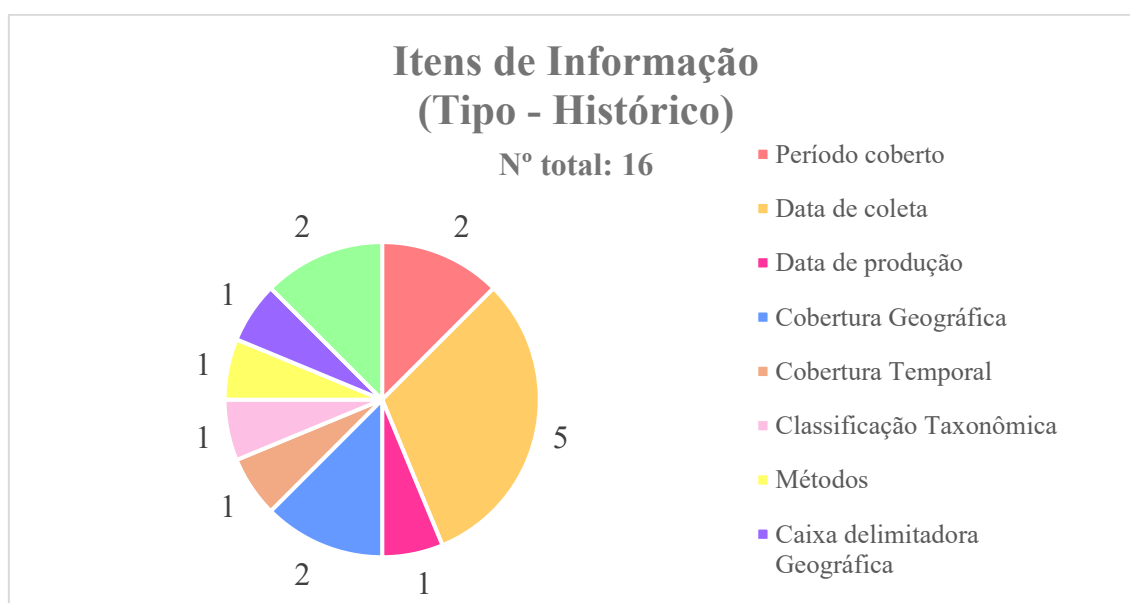
De acordo com o Gráfico 2, 9 *datasets* trazem alguma indicação relativa à Proveniência, sendo que em 1 deles só foram encontrados itens sobre **Fonte ou origem**, 5 indicam apenas itens sobre Histórico, enquanto 3 registram concomitantemente itens sobre os dois Tipos de Informação.

Assim, quanto ao **Tipo de Informação Fonte ou origem**, compreendendo que mediante disposto no quadro 20, ele engloba o item Produtor, é possível afirmar que este item só foi encontrado em 16% dos *datasets* do recorte. A ocorrência dos itens referentes ao tipo “Fonte ou origem” pode ser observada no Gráfico 3.

Gráfico 3 – Itens de informação relativos ao Tipo Fonte ou origem

Fonte: Elaborado pela autora (2025).

Em seguida, ainda dentro dessa categoria, para a observação do **Tipo de Informação Histórico** foram encontrados itens de informação. Os itens informacionais relacionados ao tipo “Histórico” encontram-se ilustrados no gráfico 4.

Gráfico 4 – Itens informacionais relativos ao Tipo Histórico

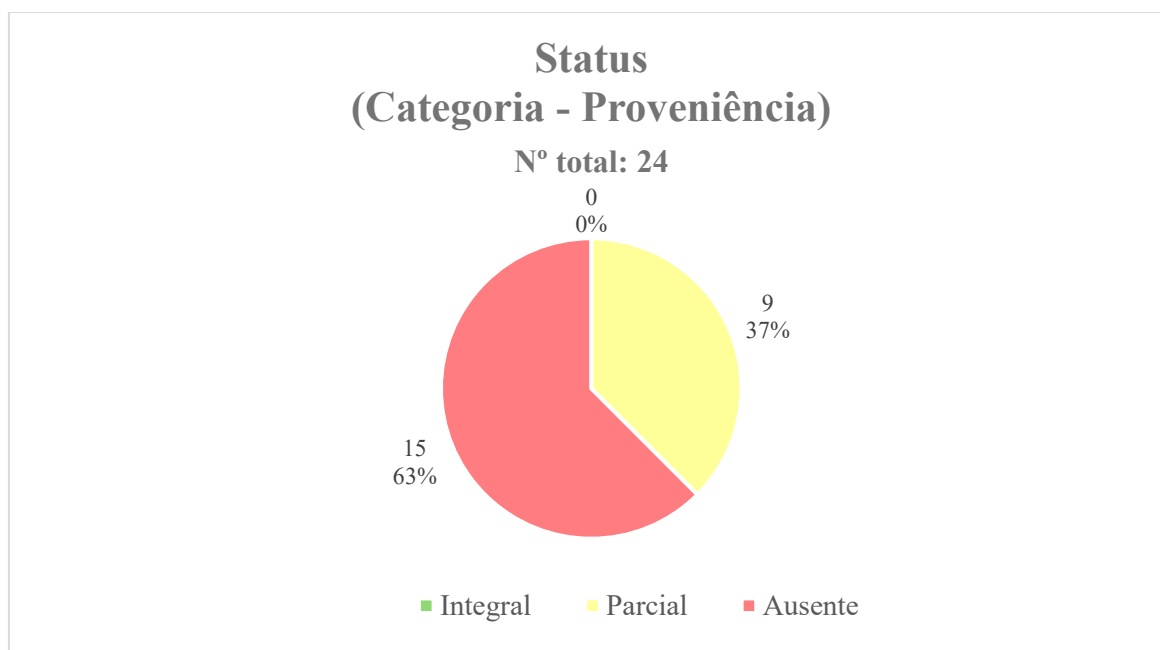
Fonte: Elaborado pela autora (2025).

É possível observar que a maioria dos itens considerados pelo escopo pretendido corresponde ao **item de informação Data de coleta**, encontrado em 5 *datasets*, seguido por Período coberto, Cobertura geográfica e Notas, encontrados individualmente em 2 *datasets*. Os demais itens (Data de produção, Cobertura temporal, Cobertura taxonômica, Métodos e Caixa delimitadora geográfica) foram encontrados apenas 1 *dataset*.

Assim, foi constatado que dos 24 *datasets* da seleção, 1 tem apenas a informação sobre produtor como item informacional do tipo Fonte ou origem, 5 tem no mínimo 1 item informacional do tipo Histórico e apenas 3 possuem itens dos 2 tipos informacionais (Fonte ou origem e Histórico) da categoria, não tendo sido identificados outros tipos informacionais. Cabendo, portanto, sobre o registro de Proveniência, o nível AUSENTE para 15 *datasets* (63%) e PARCIAL para 9 *datasets* (37%) e nenhum tendo atingido o status INTEGRAL.

O panorama geral dos níveis atribuídos à categoria Proveniência está sintetizado no Gráfico 5.

Gráfico 5 – Status de identificação de itens para Categoria Proveniência

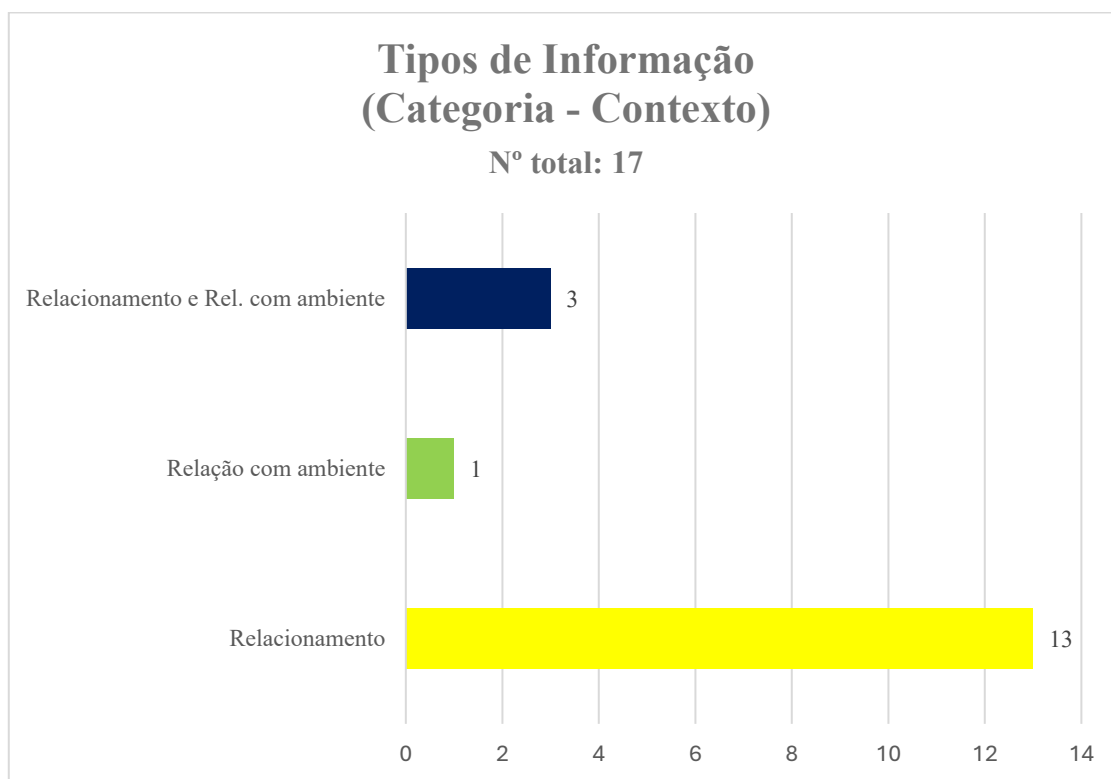


Fonte: Elaborado pela autora (2025).

6.1.1.1.2 Quanto à **Categoria de Informação Contexto** (que engloba os tipos informacionais: Relacionamentos, Relação com ambiente e Motivo de criação)

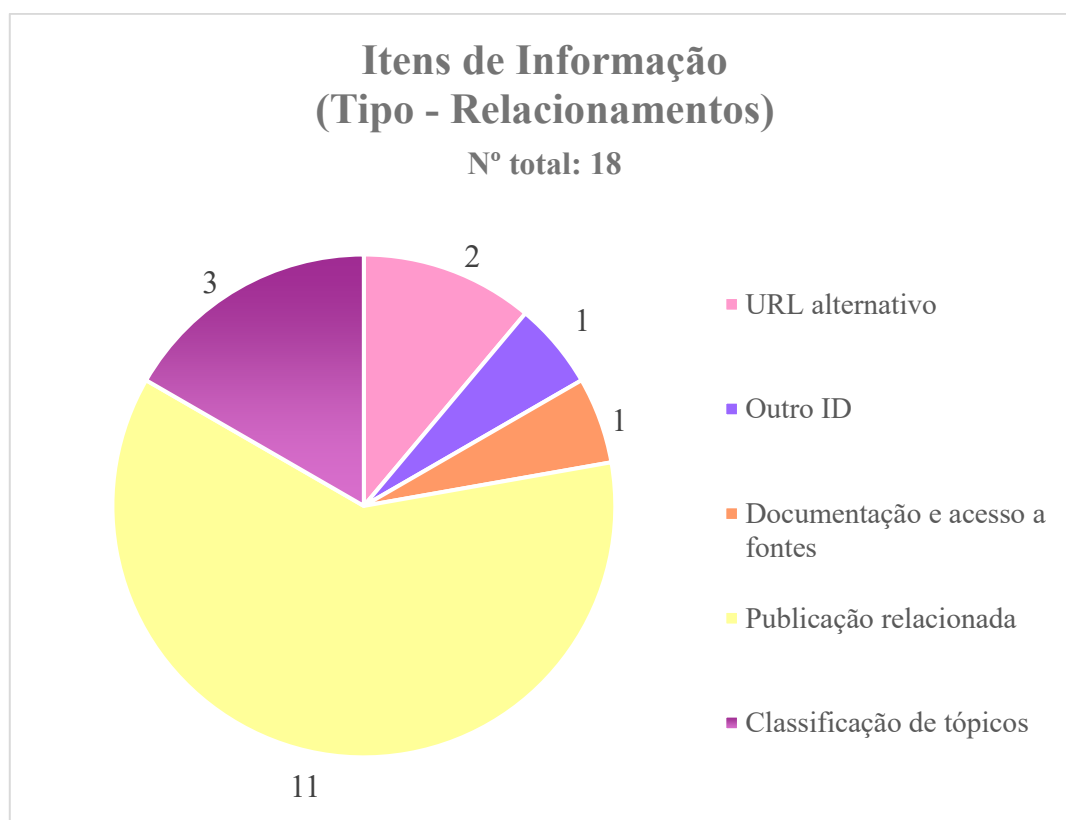
Foram encontrados nos *datasets* selecionados, apenas itens correspondentes à **Relacionamentos e Relação com ambiente**. A distribuição dos tipos informacionais vinculados à categoria Contexto é apresentada no Gráfico 6.

Gráfico 6 – Tipos informacionais da Categoria Contexto encontrados nos *datasets* selecionados



Fonte: Elaborado pela autora (2025).

Com relação aos tipos de informação desta categoria, observa-se que 17 *datasets* (71%) demonstram algum tipo relativo à categoria contexto. Tem-se o quantitativo de 14 *datasets* (58%) apontando apenas um dos tipos de informação de contexto (13 à Relacionamento e 3 à Relação com ambiente). Apenas 1 *dataset* (4%) apresenta 2 tipos referentes a Contexto. A identificação dos itens relacionados ao tipo “Relacionamentos” pode ser visualizada no Gráfico 7.

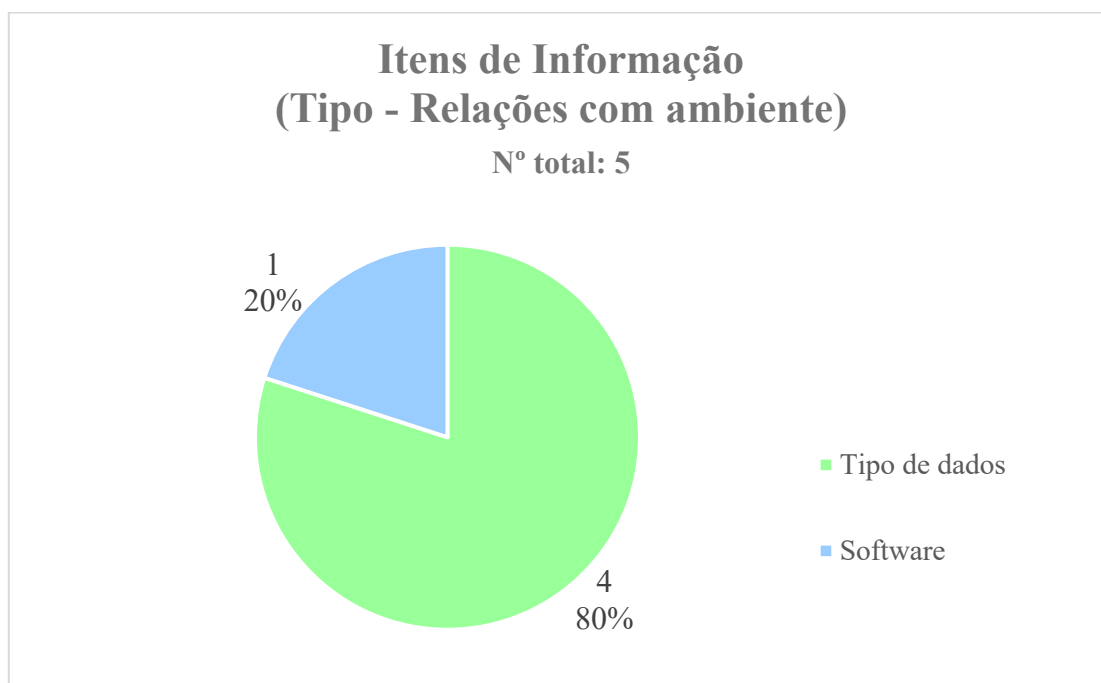
Gráfico 7 – Itens informacionais relativos ao Tipo Relacionamentos

Fonte: Elaborado pela autora (2025).

Dos 24 *datasets* do recorte, 11 tem registros sobre Publicação relacionada, 2 sobre seus URL alternativos, 1 indica outro ID, 1 registra informações no campo Documentação e acesso a fontes e 1 correlaciona Classificação de tópicos.

Já sobre o **tipo de informação Relações com ambiente**, dos 24 *datasets*, 5 registram itens relativos a esse tipo. A ocorrência dos itens associados à relação com o ambiente está detalhada no Gráfico 8.

Gráfico 8 – Itens informacionais relativos ao Tipo Relações com ambiente



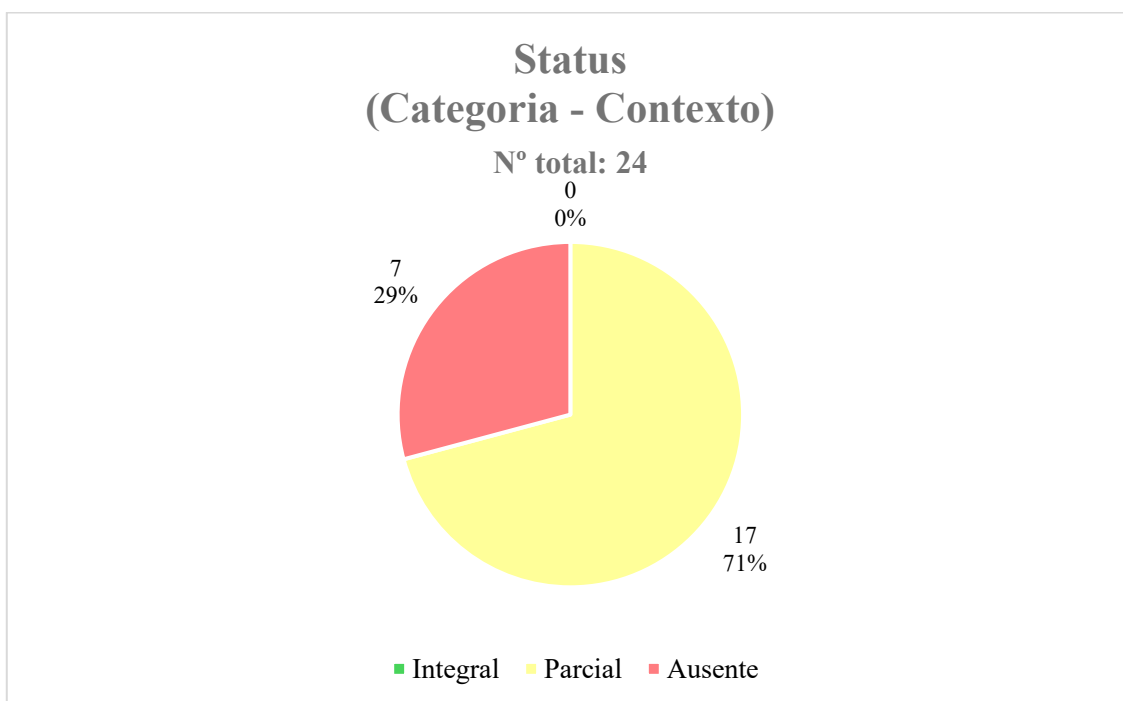
Fonte: Elaborado pela autora (2025).

Conforme Gráfico 8, 4 *datasets* têm registros sobre Tipos de dados e apenas 1 sobre o *software* utilizado.

Assim, foi constatado que todos os 24 *datasets* da seleção, 17 tem no mínimo 1 item de informação de um dos tipos Relacionamento ou Relação com ambiente, sendo que desses, 3 *datasets* registram itens relativos aos 2 tipos concomitantemente. Não foram encontrados itens que indiquem o tipo “motivo de criação”. Portanto, entendeu-se como PARCIAL o nível de registro de Contexto de (71%) dos *datasets* selecionados e AUSENTE para o restante, 7 *datasets* (29%), não tendo sido apostado o status INTEGRAL a nenhum dos *datasets* do recorte.

A síntese dos níveis atribuídos à categoria Contexto encontra-se representada no Gráfico 9.

Gráfico 9 – Status de identificação de itens para Categoria Contexto



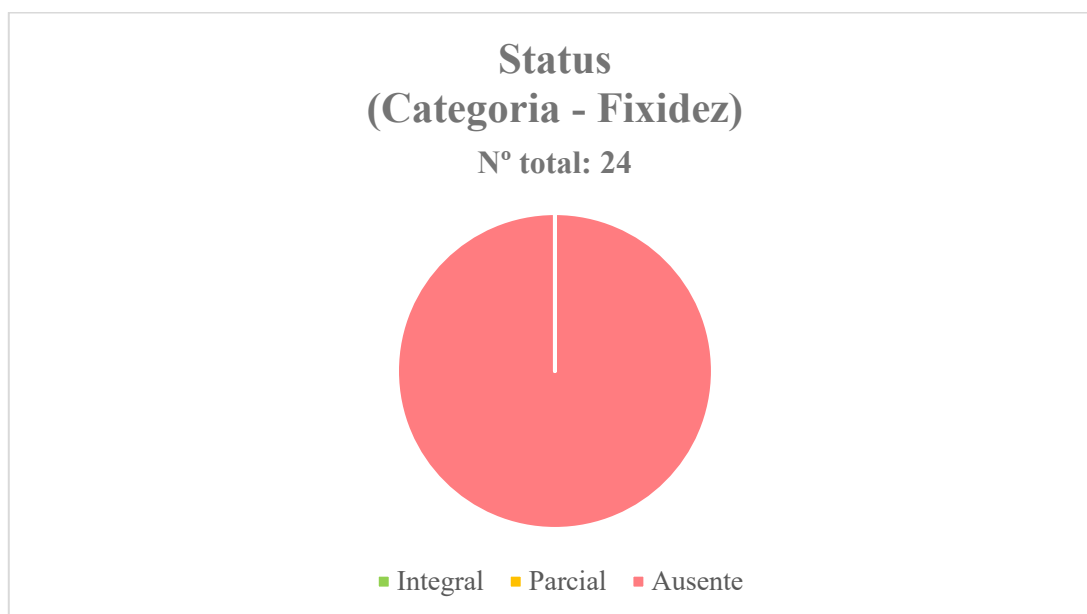
Fonte: Elaborado pela autora (2025).

6.1.1.1.3 Quanto à **Categoria de Informação Fixidez** (que engloba informações sobre verificação de integridade ou chaves de validação)

Nesse momento da análise à luz da verificação de metadados de citação e domínio dos *datasets*, não foi observado nenhum item que pudesse indicar informações relativas à fixidez. Portanto, nessa categoria, 100% dos *datasets* foram classificados como “AUSENTE”.

O Gráfico 10 apresenta o panorama referente à Categoria de Informação Fixidez, demonstrando que, para todos os *datasets* analisados, não foram identificados registros que permitissem aferir esse tipo informacional.

Gráfico 10 – Status de identificação de itens para Categoria Fixidez



Fonte: Elaborado pela autora (2025).

Com base nos dados quantitativos levantados junto ao uso da Tabela de níveis para aposição de status de identificação (Quadro 11), juntamente com a aplicação da abordagem qualitativa, observou-se que dos 24 *datasets*, em nenhum deles houve atendimento INTEGRAL dos tipos informacionais em nenhuma das categorias. Da mesma forma, 7 foram considerados com status AUSENTE para todas as três categorias. De maneira geral, 9 *datasets* foram identificados com registro PARCIAL para proveniência ao mesmo tempo que encontra-se com registro PARCIAL para contexto e AUSENTE para fixidez. Os outros 8 *datasets* foram identificados com registro AUSENTE para proveniência, PARCIAL para contexto e AUSENTE para fixidez.

A consolidação dos registros informacionais encontra-se no Quadro 21.

Quadro 21- Identificação de registros informacionais nos *datasets* selecionados

Número de <i>datasets</i>	Proveniência	Contexto	Fixidez
7	A	A	A
8	A	P	A
9	P	P	A

Fonte: Elaborado pela autora (2025).

Finalmente, quanto aos metadados de citação e domínios, compreendeu-se que 63% não registram metadados que poderiam incrementar os registros relativos à proveniência dos dados de pesquisa, mantendo exposto apenas aqueles de preenchimento

obrigatório para cadastro do *dataset* no repositório. Além disso, 29% também não fazem uso de campos correspondentes ao registro de informações que poderiam aprofundar a contextualização dos dados compartilhados.

6.2.1.2 Metadados de exportação

Ao serem analisados os metadados de exportação de cada um dos *datasets* do recorte, foi possível observar que estão sendo registrados os metadados de citação e de domínio do conjunto de dados (descritivos), junto aos metadados dos arquivos de dados e outras novas informações resultantes da gestão desses objetos a partir do momento de seu depósito no referido repositório.

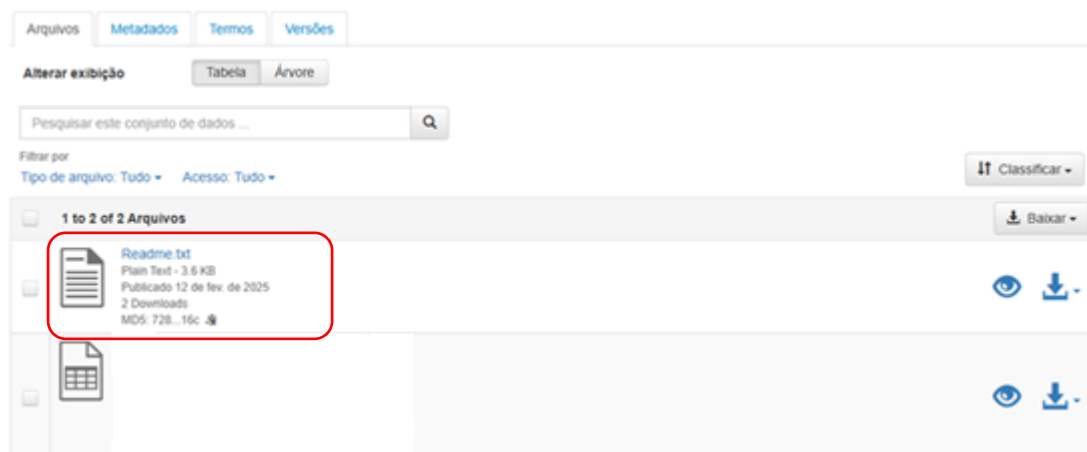
Para verificar registros relativos à proveniência, no primeiro momento por uma busca por localização com o prefixo “prov” e “premis”, não foi obtido nenhum resultado. Após, foi buscado também o prefixo “check”, para verificação de registros de checksum. Quanto a isso, foi identificado que há o registro de checksum por MD5, mas em todos eles esse registro se relaciona exclusivamente à codificação feita automaticamente pelo próprio sistema a partir da admissão do arquivo de dados, ou seja, não há registro quanto à verificação de fixidez até este momento do ciclo.

Para confirmar essa constatação, adiante foram analisados os documentos codificados em XML, HTML e JSON, contendo diversos padrões disponibilizados pelo Dataverse, como Dublin Core, OAI-ORE, CodeBook HTML DDI, OpenAIRE, DataCite, confirmando, pela análise integral, que não foi encontrado registro de informações sobre eventos precedentes ao depósito no referido repositório ou o carregamento de todo seu histórico anterior a este momento.

6.2.1.3 Arquivos *Read.me*

Após a observação dos conteúdos informacionais registrados enquanto metadados, foi necessário explorar outros caminhos para identificar informações sobre proveniência. Assim, buscou-se verificar a existência de arquivos *Read.me* (documentos textuais) disponibilizados enquanto um “dado” que compõe o conjunto de dados do *dataset* da seguinte forma:

Figura 45 - Exemplo de arquivo *Read.me* no *dataset* selecionado do Repositório 3¹¹⁷



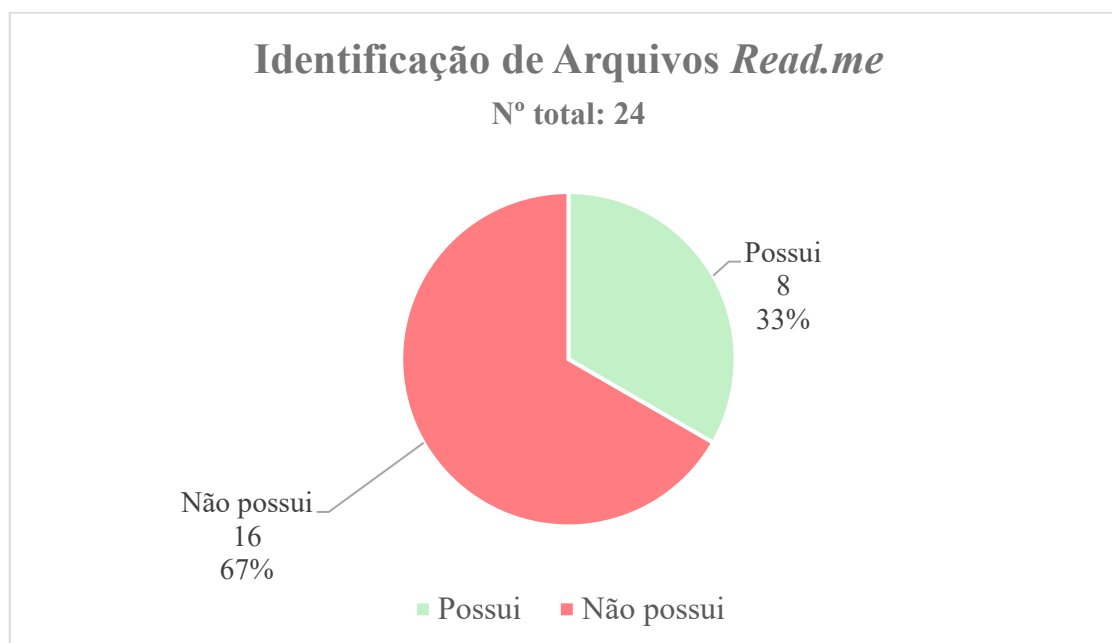
Fonte: Elaborada pela autora (2025).

Dessa forma, verificou-se que dos 24 *datasets*, 8 possuem um documento *Read.me* enquanto um de seus arquivos de dados. No entanto, constatou-se, ao analisar cada um deles, diversas configurações, o que indica não haver um padrão quanto ao tipo de informação, a forma e a estrutura para a elaboração desse arquivo.

A presença de arquivos *Read.me* nos *datasets* selecionados pode ser observada no Gráfico 11, que ilustra a proporção de conjuntos de dados que disponibilizam esse tipo de documento.

Gráfico 11 – Disponibilização de documento *Read.me* nos *datasets* selecionados

¹¹⁷ As informações do outro arquivo disponibilizado que poderiam servir para identificação do *dataset* foram apagadas.



Fonte: Elaborado pela autora (2025).

Dos 8 *datasets* com *Read.me* encontrados, 1 estava com uma *tag* tarjada em vermelho indicando que os metadados estavam incompletos e 1 estava com uma indicação de que a ingestão dos arquivos ainda estava em andamento e, portanto, o *Read.me* nele inserido encontra-se incompleto, apenas com os cabeçalhos ainda sem preenchimento. Portanto, ao analisar cada um dos outros 6 arquivos *Read.me*, foi percebido que, embora possam conter campos para preenchimento sobre autoria, data da coleção, visão geral dos dados, descrição dos métodos para geração e processamento e informações específicas sobre os dados, 3 deles registraram informações de forma enxuta ou bastante resumida¹¹⁸, segundo julgamento desta pesquisadora.

Uma observação que chamou atenção foi que em 1 dos outros 3 restantes, que apresentavam informações mais detalhadas, no campo “*describe any quality assurance procedures performed on the data*” (descreva quaisquer procedimentos de garantia de qualidade realizados nos dados”, a pesquisadora preencheu que “*The information was manually reviewed after exportation*” (A informação foi revisada manualmente após a

¹¹⁸ Alguns dos *Read.me* analisados são preenchidos de maneira mais completa, com informações de autoria, contato e os arquivos do conjunto de dados, junto a campos sobre “visão geral dos dados e arquivos”, “Data de criação de cada arquivo e como os arquivos se relacionam entre si”, “Descrição dos métodos de coleta ou geração de dados”, “Métodos usados para o processamento de dados” e “Informações específicas dos dados”. Outros só apresentam informações como “contato do pesquisador”, “data da coleta”, “visão geral” e “descrição de métodos” ou apresenta o “contexto geral”, mas o campo descrito como “informações sobre métodos, instrumentos e padrões” está incompleto.

exportação). Essa informação indicou que muitos dos processos inerentes ao ciclo de vida dos dados de pesquisa estão sendo feitos de forma manual.

Isso precisa ser pontuado com uma realidade latente a muitas pesquisas científicas, mas também precisa considerar a vulnerabilidade que se impõe ao uso de dados extraídos de fontes terceiras sem o devido registro de metadados precedentes sobre esses dados ou a possibilidade de que a reprodução dessas informações relativas aos dados, como proveniência e relacionamentos, possa conter erros humanos de preenchimento.

6.2.2 À luz da análise de cada um dos arquivos de dados

Para continuar a análise do campo empírico, julgou-se importante identificar, conforme exposto nessa subseção, as informações encontradas à luz da análise de cada um dos arquivos de dados disponíveis no conjunto de dados (*dataset*), observando as configurações de desses arquivos e uso de *tags* neles apostas.

6.2.2.1 Configurações do documento

Os *datasets* do recorte são compostos, no total, por 184 arquivos de dados, dos tipos: documento, textual, imagem, áudio e dados tabulares¹¹⁹ em diversos formatos (.xlsx; xls, .mp3; .pdf; .png, .tab, .csv, .docx, .accdb, .txt, .vcf.gz e .doc. Do total, 48 são de acesso restrito e 136 de acesso público (dentre os quais 79 são públicos mediante assinatura de contrato de licença/uso dos dados e 57 são públicos sem restrição).

Mediante essa análise foi possível identificar que alguns arquivos são abertos na tela ou por link de download no próprio sistema, enquanto outros recomendam a abertura através da “URL de download em um comando Wget ou em um gerenciador de downloads para evitar downloads interrompidos, tempos de espera ou outras falhas ou por link”.

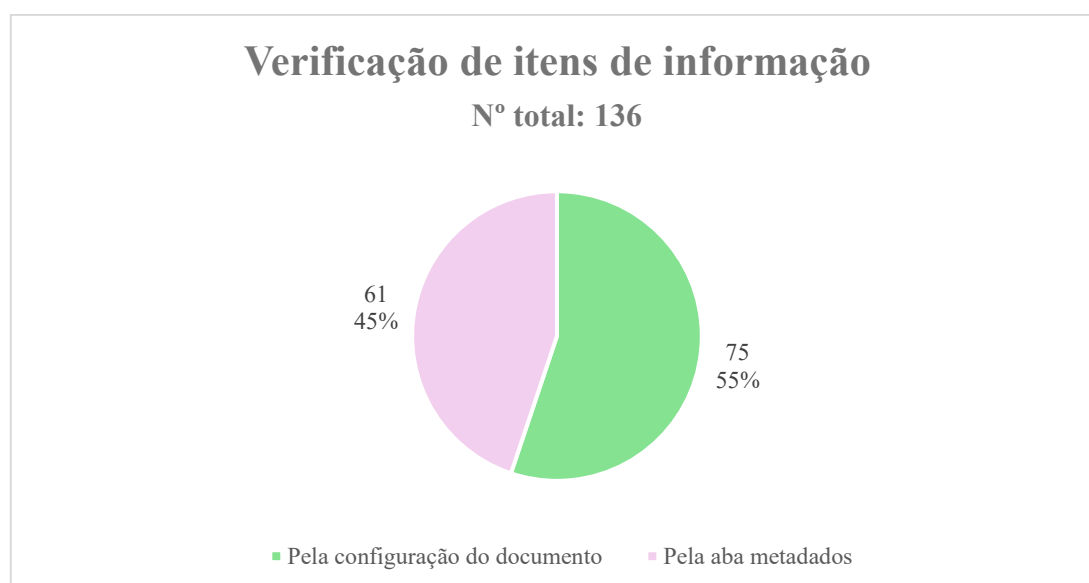
A partir desse primeiro reconhecimento, foi iniciada a verificação de itens de informação relativos à proveniência, contexto e fixidez, conforme categorização hierárquica adotada na metodologia dessa tese, na descrição das propriedades do

¹¹⁹ Essas especificações foram compiladas a partir da disponibilização dos tipos cadastrados nos próprios *datasets*.

documento encontradas na configuração do arquivo. Com isso, na intenção de abrir cada um dos arquivos de acesso público, foi verificada a impossibilidade técnica de efetuar a leitura dos seguintes formatos: .accdb (1 arquivo) e .mp3 (7 arquivos). Junto a eles, para arquivos que não puderam ser baixados, apresentaram erros de download ou sem descrição nas configurações, conforme exposto em detalhamento na tabela de verificação de arquivos de dados dos *datasets*, apêndice B, foram analisadas, exclusivamente, as informações disponíveis pela aba “metadados” dos arquivos.

O Gráfico 12 explicita as fontes utilizadas para identificação dos itens de informação nos arquivos de dados, distinguindo os casos analisáveis pela configuração do documento e aqueles avaliados exclusivamente pela aba de metadados.

Gráfico 12 – Fonte de identificação de itens de informação do dado



Fonte: Elaborado pela autora (2025).

A análise através da configuração dos documentos evidenciou que, quando não estão em branco, os campos preenchidos apontam, em sua maioria: Nome do arquivo; Tamanho do arquivo; Autor e/ou Criador e as Datas (última modificação e criado em). Demonstrando, portanto, itens que poderiam ser considerados como Fonte ou origem e Histórico, na categoria de Proveniência.

Todavia, julgou-se inoportuno aplicar o uso dos instrumentos elaborados para essa tese visto que a análise revelou a possibilidade de incongruências significativas nas informações de autoria (Autor - item de informação do tipo fonte ou origem da categoria proveniência) extraídas das propriedades dos documentos. Isso pode se dar em virtude de

que, embora este campo possa ser preenchido com o nome do autor, em algumas situações, a depender das configurações e ações do usuário, ele pode ocorrer de maneira automática e apresentar imprecisões.

Um exemplo disso é a geração de um documento a partir do uso de aplicativo do sistema operacional instalado no computador de outra pessoa que tem por *default* em suas configurações o registro de autoria em documentos produzidos naquele ambiente. Da mesma forma, a data da última modificação não especifica a natureza da alteração (salvamento ou modificação substancial), introduzindo ambiguidade.

Quanto à análise dos arquivos pela aba “metadados”, foi possível verificar que existe um padrão para o registro de alguns metadados como: ID persistente, URL de download, MD5, Data do depósito, Data de lançamento de metadados, Data de publicação, Tamanho, Tipo e Descrição. Informações estas geradas automaticamente pelo sistema no ato de carregamento do arquivo no *dataset*. Embora a geração de um hash criptográfico em MD5 produza uma “impressão digital” de tamanho fixo do arquivo, nesse contexto ele servirá para verificações de integridade futuras, mas diante do recorte dessa pesquisa, não necessariamente indica se tratar de um MD5 gerado anteriormente ao ato de depósito e submetido a uma verificação, portanto, não está considerado como item de informação relacionado à categoria fixidez.

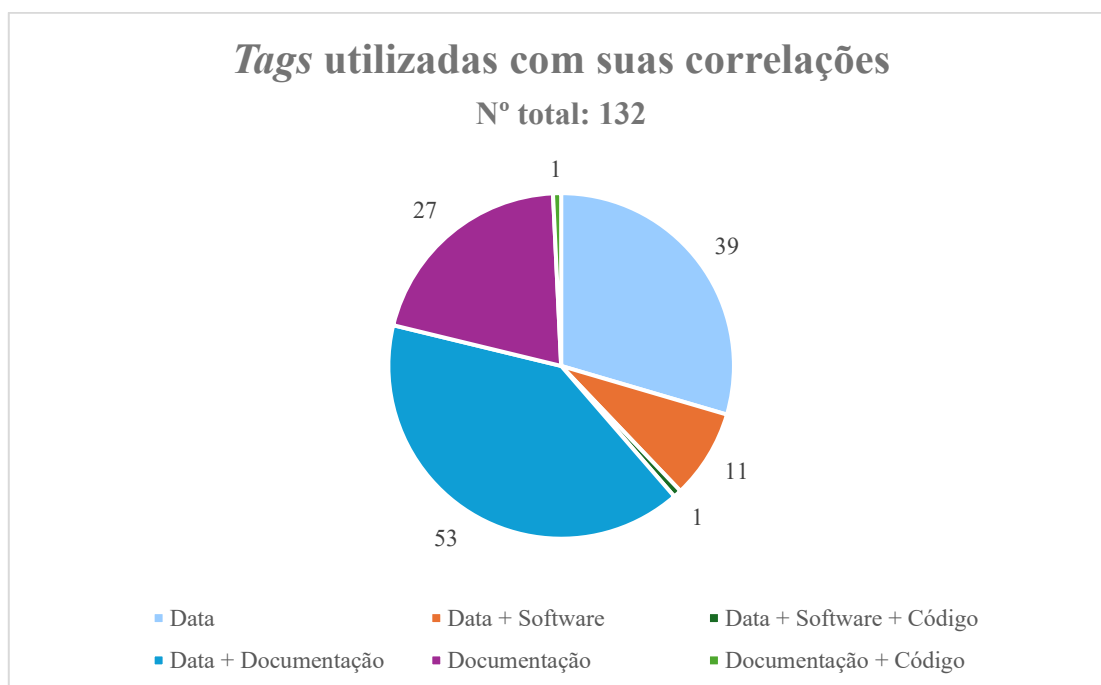
Em suma, a análise dos arquivos de dados destaca a necessidade de cautela ao utilizar as propriedades dos documentos como fonte de metadados confiáveis e reforça a importância dos metadados fornecidos pelo sistema para garantir a integridade e a rastreabilidade dos dados.

6.2.2.2 Uso de *Tags*

Sobre a conferência da utilização de *tag* enquanto um recurso capaz de especificar a natureza do arquivo disponibilizado, foi possível identificar que num total de 184 arquivos de dados, 132 utilizam *tags* (64%), dentre os quais encontram-se “Documentação”, “Data”, “Software”, “Código” e outros termos que correspondam à pesquisa especificamente. Ao desconsiderar o uso de *tags* relativas a termos que correspondam a possíveis palavras-chave que facilitem a busca do arquivo de dados, tem-se por observância que as *tags* “Documentação”, “Data”, “Software”, “Código” estão correlacionadas nos arquivos de dados.

A distribuição do uso de *tags* nos arquivos de dados encontra-se representada no Gráfico 13, indicando como esses marcadores foram empregados para caracterizar a natureza dos arquivos disponibilizados.

Gráfico 13 – Uso de *tags* nos arquivos de dados dos *datasets* selecionados



Fonte: Elaborado pela autora (2025).

Dos 132 arquivos de dados, 53 estão taguados como sendo dados e documentação; 39 são indicados como dados; 27 tem indicação de se tratar de documentação; 11 estão como dados e *softwares*; um arquivo relativo à documentação e código; e outro que indica dado, *software* e código.

Cumprе esclarecer que na tentativa de clicar para abrir os arquivos que abarcam “Documentação” em suas *tags* (81 arquivos de dados), foi verificado que estes são vinculados a 5 *datasets*, dos quais 2 são de acesso restrito, não sendo possível analisar de imediato, e dos 3 com acesso público, os quais foi possível abrir os arquivos para verificar, identificou-se que em 1 *dataset* a escolha para taguear os arquivos como documentação foi adotada para arquivos em formato .pdf produzidos enquanto produtos da pesquisa, capazes de imprimir certa forma de contextualização do projeto de pesquisa ao qual se vincula a partir da disponibilização do que está sendo desenvolvido no decurso do projeto, e 2 optaram por apor a *tag* “documentação” aos arquivos que abordam a

metodologia aplicada aos levantamentos da pesquisa, com registro de etapas, técnicas utilizadas e análises realizadas.

Assim, dos arquivos de dados com *tags* “documentação”, que compõem esses 3 *datasets* com acesso público, entende-se que, de alguma forma, questões relativas a informações sobre a proveniência estão, mesmo que de maneira limitada, sendo registradas a partir da contextualização de coleta/criação dos dados disponibilizados. Observou-se, portanto, que o uso de *tag* não segue um padrão de especificação para seu uso e, em nenhum foi demonstrada informação sobre histórico de eventos e transformações aos quais esses dados possam ter sido submetidos ao longo do seu ciclo de vida até chegarem ao repositório.

A observação integrada dos metadados de citação e domínio, dos metadados de exportação, das propriedades dos arquivos, dos arquivos *Read.me* e do uso de *tags* revelou um panorama caracterizado pela predominância de descrições pontuais, pouco articuladas entre si e insuficientes para materializar, de forma consistente, as dimensões informacionais essenciais à autenticidade: proveniência, contexto e fixidez.

Esses achados indicam a persistência de lacunas que atravessam diferentes etapas do ciclo de vida dos dados e que comprometem a completude dos registros necessários para a confiabilidade no presente e para a preservação qualitativa no futuro. Assim, evidencia-se que a gestão dos dados de pesquisa, tal como observada no campo empírico, necessita ser fortalecida para sustentar práticas de transparência e (re)uso confiável.

Com base nesse diagnóstico, torna-se pertinente avançar para a análise do grau de transparência e maturidade da gestão dos dados de pesquisa. A Seção 6.3 aprofunda essa avaliação, articulando os resultados empíricos às exigências conceituais e normativas discutidas ao longo da tese, de modo a interpretar em que medida as práticas observadas atendem, ou não, às expectativas informacionais que sustentam a confiabilidade e a preservação digital qualitativa no contexto da Ciência Aberta.

6.3 GRAU DE TRANSPARÊNCIA E MATURIDADE: ANALISANDO A GESTÃO DE DADOS DE PESQUISA A PARTIR DAS INFORMAÇÕES ENCONTRADAS NA PESQUISA DE CAMPO

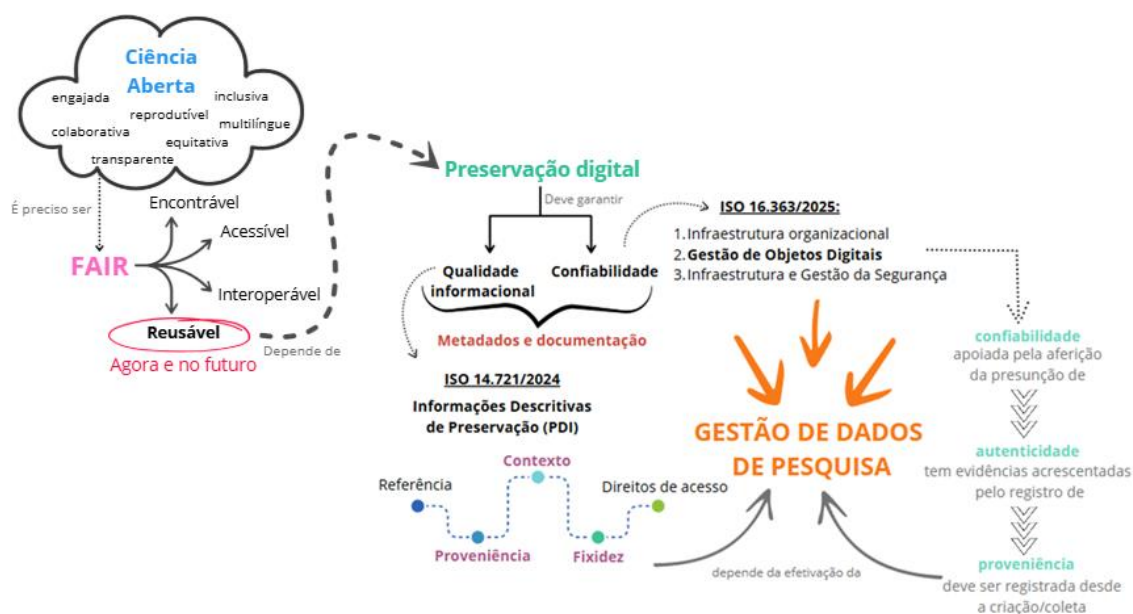
Conforme demonstrado ao longo desta pesquisa, a preservação digital a longo prazo deve assegurar não apenas a gestão dos objetos digitais em repositórios confiáveis (conforme estabelecido na Seção 2 da ISO 16.363:2025), mas também a confiabilidade intrínseca desses próprios objetos. Nesse sentido, compreende-se que a confiabilidade dos dados de pesquisa é uma qualidade sustentada por registros informacionais consistentes, capazes de respaldar a presunção de autenticidade, composta por identidade e integridade.

No âmbito desta tese, essa autenticidade é reforçada por meio de evidências acrescidas pelo registro de informações capazes de demonstrar sua proveniência, que engloba o histórico dos objetos, eventos e agentes envolvidos.

Dessa forma, é imprescindível que essas informações sejam registradas e mantidas de forma persistente desde o início do ciclo de vida dos dados, compondo, com o maior grau de completude possível, as categorias de informação das PDI, especialmente as de contexto, proveniência e fixidez, dos Pacotes de Informação para Arquivamento (*Archival Information Package – AIP*) destinados à preservação em repositórios digitais baseados no Modelo de Referência OAIS (2024).

A compreensão integrada de todos os aspectos teóricos desenvolvidos ao longo desta tese, que articulam as práticas da Ciência Aberta, a curadoria digital de dados de pesquisa, a preservação digital e a Arquivologia, permite visualizar as interdependências entre elas. A seguir, apresenta-se a ilustração dessa relação, esclarecendo como a gestão adequada dos dados de pesquisa, conduzida sob uma perspectiva arquivística, constitui o elo central entre os dados de pesquisa e o contexto da Ciência Aberta

Figura 46 – Ilustração da conexão sistemática entre os temas abordados na pesquisa



Fonte: Elaborado pela autora (2025).

A Figura 46 ilustra a interdependência entre a Ciência Aberta, os princípios FAIR, a preservação digital e a gestão dos dados de pesquisa, demonstrando que a efetividade das práticas de (re)uso depende diretamente da qualidade informacional e da confiabilidade desses dados.

Observa-se que a preservação digital, conforme estabelecido pelas normas ISO 16.363:2025 e ISO 14721:2025, deve garantir tanto a infraestrutura organizacional e tecnológica quanto a gestão dos objetos digitais. Essa confiança, contudo, só é alcançada quando os dados de pesquisa são geridos desde sua criação de modo a assegurar sua proveniência, contexto e fixidez, elementos essenciais à autenticidade e, consequentemente, à sua confiabilidade.

Assim, a Figura 46 reforça que o (re)uso, princípio R de FAIR, somente se concretiza quando os dados são objeto de uma gestão sistemática, capaz de promover sua transparência e manutenção da presunção de sua autenticidade.

Compreende-se que o registro de metadados capazes de apontar a existência dessas informações no repositório de gestão de dados de pesquisa, ambiente considerado “intermediário” no ciclo de vida desses objetos digitais, é relevante para a constituição de um histórico ininterrupto sobre todos os agentes, processos, procedimentos e percursos aos quais esses dados de pesquisa foram submetidos até terem sido nele depositados.

Além disso, compreende-se que qualquer lacuna informacional nesse sentido possa incidir diretamente sobre a confiança do que está sendo disponibilizado e, ainda

mais relevante é compreender que, analisando o “meio do caminho” do ciclo de vida dos dados de pesquisa, se essa mesma lacuna informacional for encontrada, ela será perpetuada nos estágios posteriores podendo, conseqüentemente, comprometer o interesse sobre seu (re)uso.

Portanto, é preciso estabelecer, minimamente, quais informações devem ser registradas e mantidas de forma persistente junto aos dados de pesquisa antes mesmo de sua admissão em um repositório de gestão de dados, como o Dataverse. Considerando que, atualmente, mesmo quando os campos “descrição de proveniência” são preenchidos ou quando é carregado um “arquivo de proveniência”, essas informações não são exibidas aos possíveis (re)usuários, e tampouco integram o conjunto de metadados de exportação em casos de interoperabilidade entre sistemas, torna-se necessário desenvolver estratégias que impeçam a perda de informações relevantes e promovam sua exposição para verificação humana.

É importante esclarecer que todos os metadados observados na análise compõem um rol essencialmente descritivo. A identificação dos agentes (autor, produtor etc.), dos eventos (datas de coleta, criação etc.) e das informações sobre a pesquisa (cobertura geográfica, publicações relacionadas etc.), embora disponibilizadas dessa forma, são apresentadas isoladamente, sem considerar suas correlações. Isso cumpre apenas parcialmente o que se entende por proveniência e não atende integralmente à demanda informacional.

A falta de uma visualização clara das datas reais de coleta ou criação dos dados pode gerar interpretações equivocadas sobre o histórico do conjunto desses dados de pesquisa. Quando essa informação não é explicitada, cria-se a impressão de que esses dados passaram a existir, ou a ser geridos, apenas a partir da data de seu depósito no repositório, gerando lacuna sobre uma informação crucialmente básica no que tange à sua proveniência.

A ausência de registros que demonstrem todo o histórico, com evidências sobre a Cadeia de Preservação, os eventos realizados, os agentes envolvidos anteriormente à admissão no repositório podem indicar que essas informações não serão passíveis de conhecimento nem no presente, nem no futuro.

Diante desse cenário, constata-se que, caso esses conjuntos de dados de pesquisa sejam avaliados como permanentes e precisem ser recolhidos para outro sistema que execute a preservação a longo prazo, conforme o Modelo de Referência OAIS, os

metadados que os acompanharão serão, em grande medida, exclusivamente descritivos, voltados à busca e ao acesso, acrescidos, possivelmente, apenas de metadados registrados a partir do momento de seu depósito. Permanecem, assim, fragilmente documentadas, ou mesmo ausentes, as informações referentes ao período que antecede esse depósito.

Em apenas oito dos 24 conjuntos de dados (*datasets*) analisados foi possível identificar, de forma facilitada, as datas de coleta ou criação, constantes nos campos “Data de criação”, “Data de coleta” ou em “Notas” da aba “Metadados de citação”. Essa constatação reforça a existência de uma lacuna informacional relativa ao histórico de gestão desses conjuntos. Por exemplo, verificou-se que, em um dos casos, os dados foram criados em 2018, mas apenas depositados no repositório em 2022, permanecendo, portanto, cerca de quatro anos em outro ambiente antes de sua inserção.

Apesar de todos esses apontamentos, considera-se inadequado formular qualquer julgamento sobre a gestão dos dados até o momento de sua transferência para o *software* em questão. Ainda que não seja possível avaliar o tratamento dispensado anteriormente, se os dados de pesquisa estão disponíveis para acesso, algum tratamento para que eles pudessem ser compartilhados foi a eles destinado. Se isso foi feito de forma padronizada, adequada e responsável, não é possível reconhecer diante do recorte empírico estabelecido e não foi a intenção pretendida.

Acredita-se, porém, que as chances de analisar a gestão de dados anteriormente aplicada aumentariam a partir da disponibilização de alguns recursos de documentação como, por exemplo, o Plano de Gestão de Dados (para aqueles aos quais cabe sua elaboração), junto ao conjunto de dados. Essa documentação possibilitaria conhecer de maneira mais aprofundada o contexto da pesquisa e, principalmente, o de criação e manutenção de seus dados até então.

Além disso, também seria possível conhecer qual foi o tratamento pretendido e dispensado aos dados, em um período inicial de seu ciclo de vida, os critérios estabelecidos para sua criação, de que maneira foram criados e quais recursos foram utilizados para isso e se o que está disponibilizado é o conjunto integral dos dados da pesquisa ou se para seu compartilhamento foi preciso, ou recomendada, a seleção de uma parte somente e, caso tenha ocorrido, quais foram os critérios e motivos que embasaram essa avaliação.

No entanto, em nenhum dos *datasets* avaliados foi possível identificar a existência de um Plano de Gestão de Dados para as pesquisas cujos dados foram compartilhados.

Em contrapartida, foi possível observar a disponibilização de arquivos *Read.me* em alguns *datasets*. Destaca-se sua relevância, principalmente enquanto o sistema não possibilita que sejam transparecidos os arquivos de proveniência, visto que esse documento pode fornecer informações específicas do domínio e das demais descrições sobre a criação e manutenção dos dados, apoiando sua interpretação e entendimento.

Também é possível que seja inserida nesse arquivo, uma declaração da integridade ou das limitações do conjunto de dados, para que a Comunidade Designada possa avaliar a qualidade substantiva dos dados publicados no repositório.

Todavia, é preciso o estabelecimento de um rol de informações mínimas a serem registradas nesse arquivo em estrutura e formato padrões. Padrões esses inexistentes nos repositórios analisados. Em alguns *Read.me* havia apenas o cabeçalho, já em outros, foi possível verificar incompletude de informações.

Para solucionar essa incongruência, é necessário que os pesquisadores tenham acesso, por meio da conscientização de sua necessidade e importância, a orientações que precisam ser previamente consensuadas nas instâncias responsáveis pela normatização da gestão de dados na instituição.

Desta forma, compreendeu-se que, embora seja possível aplicar estritamente apenas os elementos mínimos descritivos (título, autor, descrição, assunto, data de depósito), para permitir a descoberta dos conjuntos de dados, a transparência sobre os processos que pretenda apoiar a confiabilidade dos dados de pesquisa demanda necessariamente a adoção de metadados significativamente mais ricos que incorporem informações desde sua criação e sejam persistentemente mantidos aos dados aos quais se vinculam.

Para além da gestão dos dados de pesquisa, é imprescindível que seja realizada a gestão inicial e contínua desses dados e seus metadados. Ou seja, reforça-se que para fins de confiabilidade a gestão de dados deve englobar, inexoravelmente, a gestão de seus metadados.

Revela-se, portanto, o achado dessa análise: **a centralidade da gestão de metadados para a qualificação da confiabilidade de dados de pesquisa em ambiente de repositório para sua gestão.** Compreende-se que tanto a preservação quanto a gestão de dados de pesquisa são indispensáveis à Ciência Aberta, mas a qualidade desses dados se apoia nas informações registradas nos metadados, de forma clara, transparente, persistente e cumulativa, ao longo do seu ciclo de vida. Pode-se acrescentar, também, a

importância de toda documentação capaz de demonstrar sua proveniência digital. Isso ampara a confiabilidade e seu potencial de (re)uso.

Assim, a gestão de dados deve incorporar, de forma indissociável, a gestão de seus metadados, articulando-se aos fundamentos arquivísticos que orientam a produção e a gestão de documentos.

Na subseção seguinte, aprofunda-se essa discussão, relacionando o potencial de confiabilidade dos dados às exigências de uma preservação digital qualitativa.

6.4 A ANÁLISE DO POTENCIAL DE CONFIABILIDADE PARA PRESERVAÇÃO DIGITAL QUALITATIVA DOS DADOS DE PESQUISA

Ao considerar a necessidade de preservar os dados de pesquisa e todas as informações a eles associadas, de modo a garantir sua qualidade informacional e viabilizar seu (re)uso, compreende-se que o dado, isoladamente, muitas vezes não constitui evidência suficiente das etapas e procedimentos da pesquisa, muito menos contém elementos que assegurem sua interpretação e confiabilidade. Para tanto, torna-se imprescindível o registro de sua proveniência, a manutenção de seu contexto e o controle das informações que trilhem sua fixidez e reforcem sua autenticidade.

Desse modo, considera-se adequada a adoção de uma perspectiva abrangente e integrada de preservação digital, que contemple de forma sistêmica todas as ações realizadas ao longo do ciclo de vida dos dados, desde sua produção, gestão e compartilhamento até a preservação a longo prazo propriamente dita. Essa abordagem amplia o escopo da preservação, incorporando-a como dimensão estratégica que deve estar presente em todas as fases da pesquisa.

Como proposição de embasamento metodológico para a preservação digital dos dados de pesquisa, destaca-se a pertinência de recorrer ao arcabouço teórico e prático da Arquivologia, sobretudo aos saberes e procedimentos consolidados na gestão de documentos.

Entende-se, pela natureza dos dados de pesquisa, que o campo arquivístico oferece contribuições essenciais para a gestão desses objetos digitais. Entre elas, destaca-se a reflexão sobre a autenticidade, tema amplamente estudado na Arquivologia, sobretudo após o advento das tecnologias digitais. O Projeto InterPARES, uma das iniciativas mais

relevantes nesse contexto, define autenticidade como “[...] qualidade de um registro ser o que pretende ser e que está livre de adulteração ou corrupção” (InterPARES, 2012, tradução nossa).

No Brasil, o Conselho Nacional de Arquivos (CONARQ), por intermédio da sua Câmara Técnica de Documentos Eletrônicos (CTDE), ao atualizar o “Glossário Documentos Arquivísticos Digitais” define a autenticidade a partir da concepção proposta do Projeto InterPares acrescentando a afirmativa de que ela “é composta de identidade e integridade” as quais “são constatadas à luz do contexto (jurídico-administrativo, de proveniência, de procedimentos, documental e tecnológico) no qual o documento arquivístico foi produzido e usado ao longo do tempo” (CTDE, 2020, p. 12). A identidade caracteriza o documento como único e a integridade garante que ele esteja completo e que não tenham sofrido nenhum tipo de corrupção ou alteração não autorizada nem documentada (CTDE, 2020, p. 34-35).

Sob essa perspectiva, ao se estabelecer uma analogia entre documentos arquivísticos e dados de pesquisa, infere-se que a presunção de autenticidade desses últimos depende da análise integral de seu ciclo de vida. Qualquer adulteração indevida ao longo desse processo compromete sua confiabilidade e valor probatório. Por isso, torna-se necessário que o ciclo de vida tanto dos dados quanto de seus metadados seja planejado e controlado, com critérios de proteção contra modificações indevidas e registro de todas as alterações, sejam elas deliberadas ou acidentais.

Cabe esclarecer que a precisão no registro de metadados não é mero preciosismo, mas sim condição essencial para assegurar a confiabilidade dos dados de pesquisa. A ausência de metadados sobre o histórico do ciclo de vida desses dados, ainda que não comprometa sua acessibilidade, afeta diretamente a presunção de sua autenticidade. Lacunas informacionais sobre a proveniência, o contexto e a fixidez, decorrentes de registros incompletos ou incipientes, impactam negativamente seu potencial de (re)uso.

Para que os dados sejam confiáveis ao ponto de serem reutilizáveis, é fundamental saber quem os gerou, como, quando e com que finalidade.

Eu (e as minhas máquinas) precisamos saber quem gerou os dados, como, em que momento e com que finalidade, para podermos avaliar se quero confiar e reutilizar esses dados no meu processo de descoberta ou de tomada de decisão. Portanto, na ciência moderna a proveniência é fundamental (Mons, 2018, p. 35, tradução nossa).

Essa afirmativa reforça que a proveniência é um aspecto fundamental na ciência moderna. Em complemento, Peset e Gonzalez-Moreno (2016) trazem à reflexão o entendimento de que a atenção aos dados de pesquisa não pode ser destinada apenas à etapa final de seu ciclo, ou seja, a publicação, a ponto de que, se assim for, “a abertura será menos fluida e generalizada”, reforçando que a abertura dos dados, ao ponto de cumprirem com seu aspecto de prestação de contas só é possível se eles forem “gerenciados a partir de sua produção, padronizando a forma como serão coletados, produzidos e tratados durante os projetos” (Peset; Gonzalez-Moreno, 2016, p. 243).

Além da importância da gestão desde a produção, Fernandes (2012) explica que é extremamente importante a preservação dos processos de análise científica como forma de garantir a proveniência dos dados de pesquisa, o que permite o rastreamento da história e propriedade do experimento, e a autenticidade, enquanto um dos principais fatores na validação dos resultados obtidos durante a pesquisa.

Todavia, ainda faltam métodos comprovados tanto para garantir a continuidade a longo prazo das informações digitais quanto para assegurar que elas se manterão acessíveis e, se sim, que serão autênticas e confiáveis (Fernandes, 2012). Nesse sentido, o autor reforça, ainda, que essas informações só sobreviverão e permanecerão aptas ao acesso e uso se houver uma preocupação em sua gestão e avaliação desde o início do seu ciclo de vida, já que são necessárias intervenções ativas durante todo seu período de vida útil.

Em convergência com essa perspectiva, Akmon *et al.* (2011) reforçam que

A experiência com a seleção, a preocupação com a preservação do contexto para manter o significado e o reconhecimento de que as práticas de preservação de materiais digitais devem começar no início do ciclo de vida dos documentos seriam de particular valor para a curadoria de dados científicos (Akmon *et al.*, 2011, p. 123, tradução nossa).

Nesse sentido, a preservação digital abarca todas essas ações de manutenção do objeto digital, seu contexto e informações que demonstrem que ele não foi comprometido por nada do que foi feito durante esses processos de gestão, uso, preservação, compartilhamento e acesso.

Para isso, além dos requisitos técnicos e tecnológicos envolvidos diretamente na ação de preservar, destaca-se que é preciso um investimento no tratamento ao longo do ciclo de vida dos dados de pesquisa que viabilize efetivamente a possibilidade de reusá-los. Isso porque um conjunto de dados sem documentação sobre os métodos de pesquisa

que os geraram, a análise e a interpretação das descobertas e o seu contexto, pode não passar de uma sequência de números, ou seja, os dados se tornam mais úteis aos outros quanto melhor for documentado e curado (Borgman, 2018).

Diante dessa compreensão sobre documentação e contexto, o planejamento deve ser reconhecido como etapa essencial para a manutenção a longo prazo dos dados de pesquisa. Além disso, é importante ressaltar que ainda na fase precedente ao início da pesquisa, é quando deve ser desenvolvido o Plano de Gestão de Dados (PGD), ao menos o primeiro, já que o ideal é que seja atualizado e acompanhe as trajetórias escolhidas no decorrer da pesquisa.

Ao abranger informações sobre Coleta de dados, Metadados e Documentação, Ética e conformidade legal, Armazenamento e *Backup*, Seleção e Preservação, Compartilhamento de dados e Responsabilidades e Recursos (Digital Curation Centre, 2013), o PGD fornece contexto sobre as intenções e escolhas dos criadores, e, portanto, oferecem informações úteis tanto para os curadores quanto para possíveis reutilizadores dos dados na avaliação da transparência e maturidade das práticas de gestão adotadas para seu tratamento até a efetivação de sua abertura.

A partir desse entendimento, a análise empírica e a verificação do grau de transparência e maturidade observados permitem identificar que a confiabilidade potencial dos dados de pesquisa disponibilizados nos repositórios analisados é limitada, especialmente quando considerada como elemento fundamental para o amparo da preservação digital qualitativa.

À luz das lacunas informacionais evidenciadas, justifica-se, portanto, a proposição de uma abordagem teórico-metodológica que estabeleça a articulação entre a gestão arquivística de documentos e a gestão de dados de pesquisa, consubstanciando um Modelo de fluxo para a gestão inicial dos dados de pesquisa, fundamentado na Arquivologia e voltado ao registro de evidências da proveniência desses dados, aumentando sua presunção de autenticidade e, conseqüentemente, fortalecendo sua confiabilidade.

7 A ARTICULAÇÃO DOS FUNDAMENTOS ARQUIVÍSTICOS NO CONTEXTO ATUAL DE GESTÃO DE DADOS DE PESQUISA

A análise do campo empírico foi conduzida a partir da observação dos conjuntos de dados de pesquisa (*datasets*), dos próprios arquivos de dados, com o propósito de identificar e categorizar as informações registradas nos metadados e na documentação complementar disponibilizada nos repositórios selecionados. Dessa forma, o objetivo central desta seção é discutir os resultados alcançados à luz dos fundamentos teóricos delineados ao longo da tese, especialmente aqueles que tratam da proveniência, da autenticidade e confiabilidade dos dados de pesquisa enquanto evidências do processo científico.

Partindo das categorias de informação: proveniência, contexto e fixidez, foi avaliado o grau de maturidade e transparência dos dados de pesquisa compartilhados nos repositórios de gestão de dados analisados, considerando que a confiabilidade desses dados é um atributo diretamente relacionado à qualidade, consistência e completude da documentação que os acompanha. Assim, a análise ultrapassou o exame dos metadados de descrição, mas procurou compreender de que maneira as informações registradas, mas também as ausentes, interferem na aferição da presunção de autenticidade e na possibilidade de preservação digital qualitativa a longo prazo.

Os resultados evidenciaram que, embora os repositórios analisados demonstrem aderência parcial a normas e boas práticas internacionais, persistem lacunas informacionais significativas quanto ao registro sistemático do ciclo de vida dos dados de pesquisa. Tais lacunas comprometem a rastreabilidade e dificultam o reconhecimento da trajetória desses dados desde sua origem, o que impacta diretamente na sua confiabilidade e no potencial de (re)uso em novos contextos científicos e, também, na sociedade.

Diante desse diagnóstico, esta seção propõe reinterpretar os resultados à luz da Arquivologia, evidenciando como seus fundamentos podem orientar a formulação de um modelo de fluxo para gestão inicial de dados de pesquisa capaz de fortalecer a presunção de autenticidade e amparar a preservação digital qualitativa. Parte-se do entendimento de que a Arquivologia, ao oferecer um arcabouço teórico-metodológico consolidado sobre proveniência, contexto e fixidez, fornece subsídios essenciais para o registro e a manutenção das informações que sustentam a confiabilidade dos dados de pesquisa e, consequentemente, sua preservação e (re)uso.

Assim, tem-se o objetivo de articular fundamentos arquivísticos tradicionais e contemporâneos às demandas da gestão de dados de pesquisa, sobretudo em seus estágios iniciais, considerando as perspectivas de uma preservação digital qualitativa no contexto da Ciência Aberta. Ao explorar conceitos arquivísticos essenciais, pretende-se demonstrar como podem ser aplicados e adaptados para garantir a gestão cumulativa e contínua dos metadados de forma transparente. Essa articulação é fundamental para identificar lacunas e desafios nas práticas institucionais atuais, além de possibilitar a proposição de um modelo de fluxo, apresentado adiante na subseção 7.3, que visa fortalecer a preservação digital qualitativa e o (re)uso confiável dos dados de pesquisa ao longo do tempo.

Esta seção, portanto, consolida o objeto geral desta tese ao demonstrar como a gestão arquivística de documentos se configura como um caminho metodológico viável para apoiar a execução da gestão de dados de pesquisa desde seus estágios iniciais e, a partir dessa interlocução, apresentar o modelo de fluxo para a gestão inicial de dados de pesquisa.

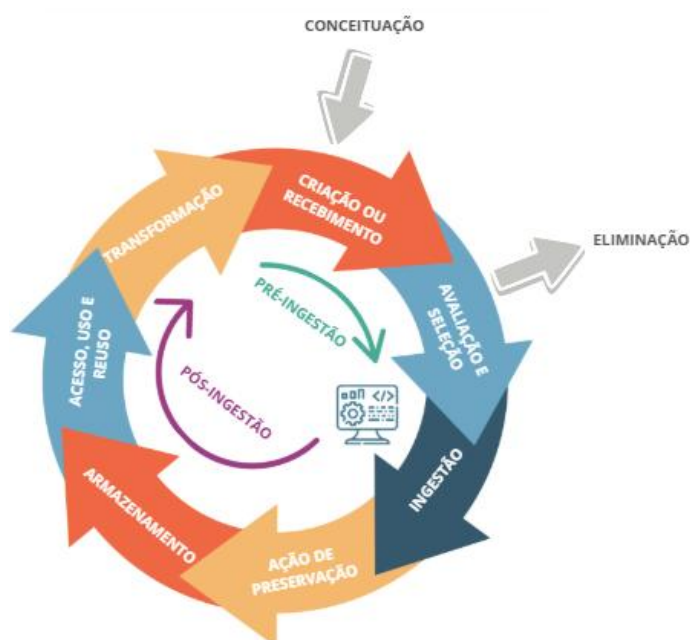
7.1 LACUNAS EMPÍRICAS E O PAPEL DA ARQUIVOLOGIA

A análise dos repositórios selecionados como campo empírico desta pesquisa revelou a ausência de registros consistentes de informações sobre a proveniência e a gestão dos dados de pesquisa, desde sua criação até o momento do depósito. A falta de metadados que registrem a proveniência digital e de documentos adicionais compromete a confiabilidade desses dados no ambiente digital. Ainda que os repositórios possam se embasar em normas internacionais para segurança e preservação, o uso de *softwares* que garantam a integridade dos dados de pesquisa em um momento intermediário do seu ciclo de vida não garante que tais dados tenham chegado confiáveis a esses repositórios.

Essa lacuna informacional relativa à proveniência pode afetar negativamente a qualidade e o valor dos dados disponibilizados para compartilhamento e (re)uso. Borgman (2017, p. 102) alerta que “sem metadados adequados, os dados científicos perdem seu contexto, tornando-se difíceis de interpretar e reutilizar”, reforçando a necessidade de uma gestão padronizada dos metadados que registrem informações essenciais sobre contexto, origem e tratamento dos dados desde os estágios iniciais de seu ciclo de vida.

A Arquivologia, conforme apresentado na seção primária 5 desta tese, oferece fundamentos teóricos e práticos para estruturar a gestão dos dados de pesquisa. Ao utilizar como referência o Modelo de ciclo de vida da curadoria digital do DCC (Higgins, 2008), apresentado na subseção 3.1.2, compreendeu-se que as funções e técnicas arquivísticas executadas na gestão de documentos (identificação, classificação, avaliação e descrição) e que conferem proveniência e organicidade devem ser realizadas anteriormente à “ingestão” nos repositórios de gestão de dados. O que seria, portanto, uma etapa denominada como “pré-ingestão”.

Figura 47 - A gestão inicial dos dados de pesquisa (etapa pré-ingestão)



Fonte: Adaptada pela autora com base em Higgins (2008).

Como pode ser visto na Figura 47, a “pré-ingestão” incorpora as ações de conceituação, criação ou recebimento, avaliação e seleção e eliminação, inferindo-se a atividade de transferência dos dados de pesquisa para depósito em repositórios antes da ingestão, de fato. Ao serem ingeridos, ou seja, no que está sendo compreendido como “pós-ingestão”, eles são submetidos a ações de preservação, armazenamento, acesso, uso e (re)uso e transformação, quando, então, podem continuar o fluxo em outros contextos de pesquisa. Um ciclo de vida que pode se perpetuar.

Preferiu-se, contudo, adotar a denominação “gestão inicial” para essa fase, em alusão ao conceito arquivístico de gestão de documentos, pois ela reforça a convergência

entre as práticas da gestão arquivística e a gestão de dados de pesquisa. Portanto, as ações que devem ser realizadas antes do depósito (ingestão) nos repositórios de gestão estão englobadas na **“gestão inicial dos dados de pesquisa”**

Assim, conforme demonstrado ao longo desta tese, a gestão inicial dos dados de pesquisa deve ser considerada etapa crucial para o registro da proveniência, a aferição da autenticidade e, conseqüentemente, a consolidação da confiança nesses dados. Dessa maneira, essa gestão inicial demanda uma abordagem arquivística capaz de manter de forma contínua, cumulativa e transparente as evidências sobre a proveniência digital desde a criação/coleta dos dados de pesquisa depositados nos repositórios.

A expertise arquivística, especialmente no que tange à gestão de documentos digitais, é essencial para o estabelecimento de metadados completos, precisos e contextualizados, permitindo o registro contínuo e cumulativo de informações que sustentam uma preservação digital qualitativa e fortalecem o potencial de (re)uso dos dados de pesquisa em consonância com os princípios FAIR, segundo os quais “os dados devem ser descritos com metadados ricos, que permitam sua descoberta, acesso e reutilização” (Wilkinson *et al.*, 2016, p. 5) e a prática da Ciência Aberta.

Como discutido na subseção 5.2 desta tese, a manutenção da organicidade é um fator estratégico para sustentar a presunção de autenticidade dos dados de pesquisa e, portanto, sua confiabilidade. Por ser manifestada a partir do contexto de produção desses dados, da sua relação com os demais conjuntos de dados e com a própria pesquisa, observou-se que ela pode ser verificada em metadados contidos na “informação de contexto”, na “informação de proveniência” e na “informação de fixidez”, definidas no OAIS (2024).

Assim, foi possível estabelecer correspondências entre as lacunas informacionais identificadas na análise empírica e as contribuições arquivísticas capazes de mitigá-las, conforme sintetizado no quadro a seguir.

Quadro 22 - Correspondência entre as lacunas informacionais encontradas e etapas do fluxo de gestão inicial dos dados de pesquisa

Categoria de informação (Subseção 6.4 - Figura 34)	Lacuna informacional identificada na análise empírica	Contribuição arquivística
Proveniência	Ausência de evidências processuais com vinculação dos dados de pesquisa, eventos, agentes e outras informações associadas à criação/coleta, gestão, transformação e preservação do Objeto Digital.	O registro da fonte ou origem, histórico, alterações desde a criação, custodiadores e trilhas de auditoria orienta a vinculação explícita entre dado, produtor, contexto de criação e eventos pelos quais foi submetido, garante a rastreabilidade e a responsabilidades envolvidas, fortalecendo a autenticidade e sua confiabilidade.
Contexto	Falta de documentação sobre o contexto metodológico da coleta, produção e manutenção dos dados de pesquisa (instrumentos, parâmetros, datas, locais, <i>softwares</i> etc.).	Metadados e documentação que descrevam os relacionamentos, as relações com o ambiente, e os motivos de criação/coleta, permitem a reinterpretação, a avaliação e o tratamento técnico adequados aos dados de pesquisa.
Fixidez	Inexistência de controle de integridade e versionamento dos arquivos digitais ao longo do processo de pesquisa até seu depósito no repositório de gestão de dados de pesquisa.	A verificação de integridade ou chaves de validação, por meio do controle de versões, <i>checksums</i> e logs de alteração, consubstanciam a integridade e, assim, aumentam a presunção de autenticidade dos dados de pesquisa.
Proveniência/Contexto	Deficiências na descrição técnica e temática dos dados (metadados insuficientes para renderização e (re)uso).	A Arquivística Integrada propõe o registro contínuo e cumulativo de metadados e documentação que evidenciam os aspectos técnicos referentes aos dados em si (formatos, por ex.), eventos (por ex. criação, verificação de vírus, normalização) e suas relações com agentes (por ex. <i>softwares</i> utilizados para transformação, pessoa física responsável, entidade custodiadora) qualificando a curadoria digital e a preservação futura.
	Falta de critérios consistentes para avaliar o valor dos dados e definir sua destinação (preservação ou descarte) e de registros que evidenciem essa avaliação.	A avaliação arquivística, amparada nos aspectos relativos à proveniência e contexto, ampara critérios para destinação dos dados de pesquisa a partir da verificação de valor científico, cultural e/ou histórico, orientando decisões fundamentadas sobre retenção (e depósito) ou eliminação.

Fonte: Elaborado pela autora (2025).

É relevante expor que essas contribuições arquivísticas são fruto de um campo historicamente dedicado à garantia da autenticidade, integridade e rastreabilidade dos documentos. Portanto, a Arquivologia oferece subsídios teóricos e práticos para a gestão inicial dos dados de pesquisa, que faz parte da curadoria digital desses dados.

Como exposto, a curadoria digital de dados de pesquisa passou, ao longo de seu desdobramento técnico e conceitual, a abranger todas as atividades inerentes ao ciclo de vida desses dados, mesmo que seja possível perceber que esse ciclo, na maioria das vezes, se inicia somente quando os dados passam a ser disponibilizados em repositórios de gestão.

Por ser um estágio sob o qual os dados ainda se encontram sob ingerência do pesquisador, ou da equipe de pesquisa ao qual está vinculado, a tentativa de estabelecer um padrão para os procedimentos técnicos da gestão desses dados pode não produzir os resultados esperados. E, de fato, o será, se as necessidades informacionais e as especificidades de cada área e da própria pesquisa em si não forem consideradas como elementos-chave na definição das melhores práticas.

Ressalta-se que todos os tratamentos destinados à manutenção da integridade e autenticidade dos dados de pesquisa possuem relevância singular em cada etapa de seu ciclo de vida, exigindo que os agentes envolvidos estejam capacitados e comprometidos com suas responsabilidades.

Qualquer processo de curadoria de dados deve levar em conta a qualidade e contexto na captura de dados no ciclo de vida. Do ponto de vista da curadoria de dados, em contextos muito distribuídos, os próprios utilizadores têm um papel potencialmente valioso como curadores, garantindo tanto a qualidade como a validade dos dados e, sancionando a utilização que lhes é dada, incluindo a reutilização para outros fins, originalmente não previsto" (Ure *et al.*, 2011, p. 132, tradução nossa).

Considerando essa interdependência entre os agentes envolvidos, eventuais vulnerabilidades em qualquer etapa podem comprometer a confiança e o potencial de (re)uso dos dados de pesquisa. Nesse contexto, tais atividades demandam uma governança estruturada capaz de apoiá-las.

Em uma tentativa de elucidar as ações práticas que precisam ser realizadas para viabilizar um possível (re)uso, expõe-se abaixo um diagrama que foi elaborado em formato de *iceberg*, uma imagem análoga àquilo que está aparente, ou que tem maior

visibilidade, em contrapartida a tudo o que se encontra submerso, mas que é a verdadeira base de apoio e sustentação do que se vê na superfície.

Figura 48 - Iceberg do (re)uso sob a perspectiva das ações de gestão de dados de pesquisa



Fonte: Elaborada pela autora (2025).

Na Figura 48, a parte visível do *iceberg*, ou seja, aquela que se encontra acima da linha d'água, corresponde às ações mais frequentemente reconhecidas e valorizadas pela comunidade científica e pelas instâncias relativas à sua avaliação: a publicação dos resultados da pesquisa e, mais recentemente, a publicação de seus dados. Tendo como objetivo o (re)uso por outros pesquisadores ou em estudos subsequentes. Essa camada, no entanto, omite uma vasta gama de atividades subjacentes que permanecem, muitas vezes, invisibilizadas.

Na parte submersa do *iceberg* estão representadas as etapas fundamentais do ciclo de vida dos dados de pesquisa, as quais sustentam todo o processo de produção científica baseada em dados. Essas etapas incluem, em ordem cronológica aproximada: o planejamento da pesquisa, a criação e/ou coleta dos dados, a identificação, a classificação, o uso, a avaliação e seleção dos dados relevantes, o compartilhamento e, junto a todas elas, a descrição e a preservação contínua. Cada uma delas requer decisões metodológicas e técnicas que impactam diretamente na qualidade, na interoperabilidade e na longevidade dos dados de pesquisa produzidos e coletados.

Essa representação reforça que a gestão de dados de pesquisa é uma atividade estruturante e contínua, cuja efetividade depende da integração entre planejamento, documentação e preservação, desde a origem até o (re)uso. Portanto, uma gestão adequada e sistemática dos dados de pesquisa, desde o planejamento até a preservação permanente, constitui condição necessária para a efetivação da Ciência Aberta.

Para tanto, o uso de mecanismos que facilitem e dinamizem a realização dessas atividades essenciais e basilares deve ser um objetivo dos profissionais que atuam na gestão de dados de pesquisa. Do ponto de vista operacional, a implantação de um *software* que apoie o início do ciclo de vida desses dados possibilita a automatização de uma série de atividades, principalmente das etapas preliminares, cuja execução, em grande parte das vezes, é atribuída aos próprios pesquisadores no contexto das práticas de pesquisa.

Essa implementação desde os estágios iniciais permite que os dados sejam criados e/ou coletados a partir de padrões previamente planejados, com documentação e metadados suficientemente capazes de registrá-los, assim como seus códigos, fluxos de trabalho, metodologias e insumos utilizados na pesquisa, atentando-se aos acordos de governança e uso. Essa automatização permite que o pesquisador possa dedicar-se com mais assertividade às atividades de colaboração e às análises necessárias ao desenvolvimento de sua pesquisa.

Optou-se pelo uso do termo preservação contínua para reforçar que, embora exista um ambiente específico para a preservação digital de objetos digitais permanentes, as ações preservacionistas devem ser executadas desde o início da pesquisa, antes mesmo da criação ou coleta de seus dados, conforme vem sendo exposto nesta tese (Figura 19 “Camadas de preservação ao longo do ciclo de vida dos dados de pesquisa”).

É preciso considerar que o apoio pragmático de todo o ciclo de vida dos dados de pesquisa advém da possibilidade de manutenção da descoberta, acesso e uso dos dados criados e coletados na pesquisa e, portanto, ações de preservação desses dados junto à documentação que contextualiza e demonstra proveniência e a informações que vão sendo incorporadas enquanto metadados.

Cumprе reforçar que, nesta tese, como foi demonstrada preservação não é sinônimo de armazenamento. A distinção entre esses dois termos se evidencia pelo fato de que a preservação visa garantir que os dados de pesquisa possam manter sua disponibilidade, identidade, persistência, renderização, interpretabilidade e autenticidade pelo tempo que for preciso. Já o armazenamento se limita a guardá-los.

Visto a experiência adquirida na integração entre os ambientes de gestão de documentos e preservação, promulgada principalmente pela Arquivística Integrada e pela compreensão da relevância de ações que viabilizem a observância da presunção de autenticidade desde a criação dos documentos e no tratamento contínuo que deve ser franqueado durante todo o seu ciclo, é possível afirmar que essas práticas consolidam um entendimento sistêmico do controle arquivístico, essencial para aferir a confiabilidade dos objetos digitais ao longo do tempo.

Considerando que os dados de pesquisa constituem objetos informacionais que, de forma transversal, permeiam diferentes campos e demandam intervenções interdisciplinares, entende-se que a gestão de documentos, aprimorada ao longo de décadas pela Arquivologia, oferece contribuições metodológicas e conceituais valiosas para a efetivação da gestão de dados de pesquisa.

É importante ratificar que a Teoria das três idades continua sendo a base do ciclo vital dos documentos arquivísticos públicos no Brasil. Isso não transpõe qualquer prejuízo para a adoção de novas dinâmicas à medida que, ao contrário, reforça os pilares dos princípios e teorias arquivísticas na consolidação do campo do saber científico, independentemente do suporte ao qual as informações estejam registradas.

Notoriamente o que vem sendo orientado pela CCSDS encontra bases nos conhecimentos arquivísticos atuais, principalmente na Arquivística Integrada, quando reitera a necessidade do estabelecimento de uma Cadeia de Preservação que dê suporte a todo o ciclo dos dados de pesquisa, por exemplo.

Reforça-se, portanto, a necessidade de que sejam estabelecidas metodologias factíveis à execução da gestão de dados de pesquisa e que, atualmente, no contexto brasileiro, o que se tem de mais consolidado nesse acompanhamento de todo o ciclo de vida é que se ancora no conceito da gestão de documentos, que até o momento é realizado com base na Teoria das Três Idades e que pode sim, ser utilizado para embasar a gestão de dados de pesquisa.

Assim, propõe-se que seja considerada a seguinte estrutura no fluxo do ciclo de vida dos dados de pesquisa, diante do reconhecimento de 3 (três) idades distintas: corrente, intermediário e permanente, em ambientes de gestão e preservação, conforme figura abaixo.

Figura 49 – Ciclo de vida dos dados de pesquisa e seus ambientes



Fonte: Elaborada pela autora (2025).

A perspectiva apresentada pela Figura 49 é relevante para reconhecer eventos e agentes envolvidos no tratamento técnico dos dados de pesquisa. Contudo, mesmo sendo apresentados com divisões, trata-se de um fluxo fluído e integrado.

Entende-se que o ambiente “Gestão inicial” compreende a gestão dos dados enquanto a pesquisa está em andamento, razão pela qual são aqui denominados de dados correntes, por adaptação ao conceito arquivístico de documentos correntes. Nesse ambiente, podem ser utilizados *softwares* disciplinares, assim como outros sistemas de negócio específicos voltados à criação, coleta e/ou transformação dos dados. Nele estão previstos: o planejamento da pesquisa, a criação e/ou coleta dos dados de pesquisa, sua identificação, classificação, uso e avaliação prévia, que culmina na transferência daqueles que precisam ser ingeridos em repositórios de gestão, por necessidade de compartilhamento, aberto ou restrito, e manutenção segura e estendida de sua guarda. Para isso, passam a ser objeto de um depósito assistido pelo curador.

Em seguida, os dados de pesquisa são admitidos em um Repositório de Gestão de Dados, ambiente denominado como “Gestão assistida”. Nele passam a ser geridos por equipe responsável pela sua curadoria, por intermédio do uso de um *software*, o qual é recomendável que cumpra os requisitos dispostos no *CoreTrustSeal*. Junto ao ambiente anterior, esse momento faz parte do que se compreende por gestão de dados de pesquisa.

Posteriormente, após nova avaliação, a avaliação final, caso sejam considerados para guarda permanente, os dados de pesquisa são recolhidos para admissão em

Repositório Digital de Preservação, idealmente aderente aos requisitos da ISO 16.363:2025. Ambiente denominado então de “Preservação”.

A integração entre os ambientes decorre do entendimento de que, ao se conhecer o objetivo final de todo o processo, torna-se possível estabelecer procedimentos, ferramentas e padrões adequados aos diferentes profissionais que, embora não sejam preservadores, integram o ecossistema da preservação. A efetividade da gestão, nesse contexto, constitui condição *sine qua non* para a preservação digital qualitativa.

Considerando que o produtor atua diretamente no ambiente “Gestão inicial”, é fundamental que receba orientações sobre a gestão dos dados de pesquisa, bem como sobre as ações e ferramentas que possam apoiar sua operacionalização.

Diante das lacunas informacionais analisadas no campo empírico e com base nos fundamentos arquivísticos aqui delineados, apresenta-se, a seguir, uma proposta de modelo de fluxo para gestão inicial de dados de pesquisa, que operacionaliza essas contribuições em etapas sistematizadas.

7.2 MODELO DE FLUXO PARA GESTÃO INICIAL DE DADOS DE PESQUISA

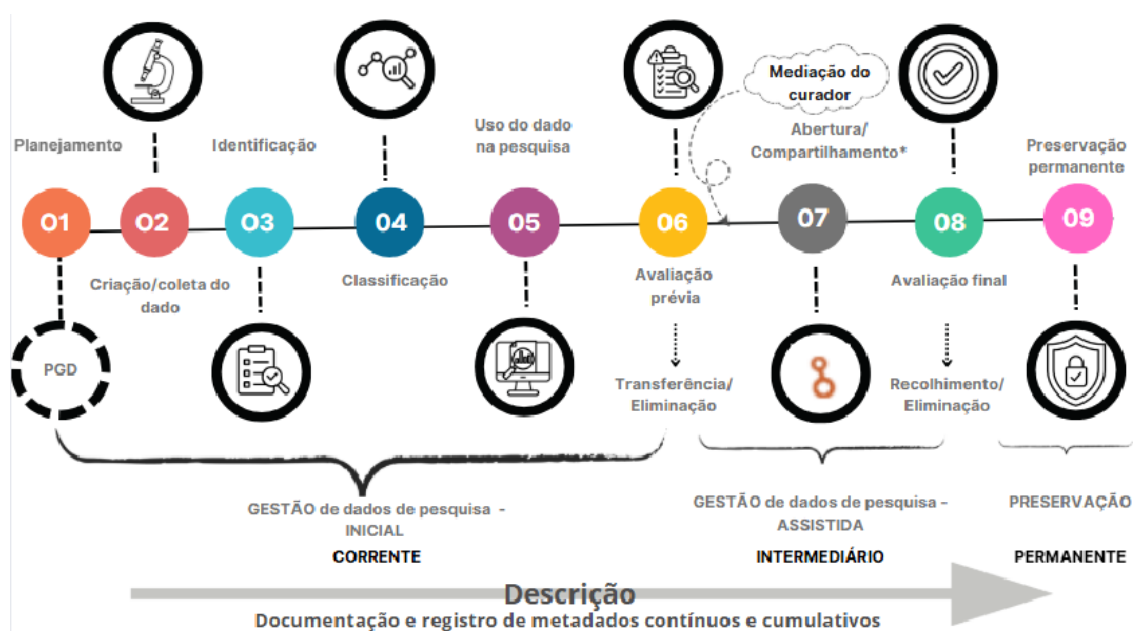
Com base na interlocução entre a gestão arquivística de documentos e a gestão de dados de pesquisa, apresenta-se o Modelo de fluxo para a gestão inicial de dados de pesquisa. A proposta visa amparar a preservação digital qualitativa e potencializar práticas científicas alinhadas à Ciência Aberta.

Nesse sentido, o modelo complementa *frameworks* consolidados de gestão e curadoria digital, oferecendo uma estrutura clara para o estágio inicial do ciclo de vida: do planejamento à transferência para repositório de gestão de dados, com definição das etapas de planejamento, criação/coleta, identificação, classificação, uso e avaliação prévia, assegurando a descrição contínua e cumulativa desses dados em documentação e registro de metadados.

Além do caráter conceitual, o modelo adota uma abordagem pragmática e operacional, em consonância com as demandas contemporâneas e as premissas da Ciência Aberta, enfatizando transparência, qualidade e sustentabilidade. Para tanto, incorpora requisitos para decisões criteriosas de retenção ou eliminação, otimização de recursos e registro documentado das decisões.

Com base nas premissas, conceitos e métodos arquivísticos aplicados a dados de pesquisa e seus metadados, apresenta-se a instanciação do ciclo de vida desses dados e sua correlação com a Teoria das Três Idades, visando sustentar a confiabilidade e o potencial de (re)uso a partir da presunção de autenticidade ancorada na documentação de proveniência.

Figura 50 – A gestão dos dados de pesquisa e sua integração com a preservação digital à luz da perspectiva arquivística



Fonte: Elaborada pela autora (2025).

Nesse sentido, a Figura 50 sintetiza essa articulação, evidenciando a integração entre os ambientes de gestão inicial (corrente), gestão assistida (intermediário) e preservação (permanente). Destaca-se que a documentação e o registro de metadados contínuos e cumulativos relativos à proveniência digital, que engloba informações sobre contexto e fixidez. Esse eixo transversal de metadados e documentação sustenta a organicidade, a preservação qualitativa e fortalece o (re)uso desses dados.

A leitura de modelos e *frameworks* da curadoria digital, apresentados na subseção 3.1 “Curadoria digital: modelos e frameworks”, que especificou o Modelo do ciclo de vida dos dados de pesquisa da JISC, o Modelo do ciclo de vida da curadoria digital do DCC e os Princípios FAIR, integrada ao que foi desenvolvido nas seções primárias 4 “Preservação digital qualitativa: o registro de metadados para o fortalecimento do (re)uso dos dados de pesquisa” e 5 “Gestão de documentos: contributos da Arquivologia para a

qualificação da preservação digital dos dados de pesquisa”, permitiu correlacionar necessidades e soluções.

A partir dessa correspondência conceitual entre a gestão arquivística de documentos e a gestão dos dados de pesquisa, o quadro abaixo sintetiza as etapas distribuídas entre os ambientes de “gestão inicial” e “gestão assistida”, conforme compreensão de apanhado teórico desta tese.

Quadro 23 – Correlações entre as etapas da gestão de dados e as funções inerentes à gestão de documentos

ETAPAS DA GESTÃO DE DADOS DE PESQUISA – DO PLANEJAMENTO À DESTINAÇÃO							
GESTÃO INICIAL					GESTÃO ASSISTIDA		
PRODUÇÃO				UTILIZAÇÃO		DESTINAÇÃO	
Planejamento da pesquisa	Criação/coleta dos dados	Identificação dos dados	Classificação dos dados	Uso dos dados	Avaliação prévia dos dados	Compartilhamento dos dados	Avaliação final dos dados
• Identificação de fontes e recursos necessários • Definição de padrões de formatos e nomenclatura • Verificação de políticas/procedimentos a serem seguidos • Definição de <i>softwares</i> a serem utilizados • Definição de responsabilidades entre os agentes envolvidos • Elaboração de Plano de Gestão de Dados (PGD) • Estabelecimento de estratégias para gestão inicial e assistida dos dados de pesquisa • Definição de metadados mínimos a serem registrados ao longo da pesquisa	• Criação de dados em formatos recomendados • Coleta de dados necessários • Transformação de dados coletados, quando necessário • Organização lógica dos arquivos de dados • Atenção às boas práticas de gestão de dados • Registro contínuo de metadados • Documentação	• Identificação de quais são os dados da pesquisa • Diferenciação dos dados DE pesquisa e dados PARA pesquisa • Execução das ações estratégicas para a gestão inicial dos dados de pesquisa • Registro contínuo de metadados • Documentação	• Classificação por categorias • Definição de prazos mínimos para retenção • Análise e estabelecimento do fluxo ideal dos dados de pesquisa • Registro contínuo de metadados • Documentação	• Observação de acordos de governança e uso dos conjuntos de dados • Colaboração entre equipes • Análises e processamento • Resultados • Registro contínuo de metadados • Documentação dos processos	• Observação dos prazos de retenção dos dados • Seleção de dados a serem eliminados • Seleção de dados para compartilhamento • Registro contínuo de metadados do processo • Documentação	• Transferência iniciada pelo gestor inicial e assistida pelo curador ao repositório de gestão de dados de pesquisa • Registro de metadados do conjunto de dados • Disponibilização dos dados • Gestão de dados de pesquisa no repositório • Registro contínuo de metadados • Documentação	• Observação dos prazos de retenção dos dados • Seleção de dados a serem eliminados • Seleção de dados para preservação permanente • Início do processo de recolhimento dos dados para Repositório Digital de Preservação (RDP) • Registro de metadados do processo • Gestão de dados de pesquisa no RDP • Documentação

Fonte: Elaborado pela autora (2025).

Essa síntese dialoga diretamente com as lacunas identificadas na análise empírica, evidenciando que cada etapa descrita no Quadro 23 demanda registro contínuo e cumulativo de metadados e documentação associada, condição para a manutenção da proveniência digital.

Na gestão inicial dos dados de pesquisa foram identificados como ações essenciais o planejamento, a criação/coleta, identificação, classificação, uso e avaliação prévia. Como é visível, a avaliação é encontrada em dois momentos do ciclo de vida da gestão, cabendo à gestão inicial o que está sendo denominado como “avaliação prévia”, visto que posteriormente, uma outra avaliação deverá ser realizada para a seleção dos dados de pesquisa que irão ser recolhidos ao Repositório Digital de Preservação.

Ao observar as propostas de ciclo de vida dos dados de pesquisa, reconhece-se a existência de semelhanças nas segmentações estabelecidas para criar uma noção de fluxo, as quais, assim como na gestão arquivística de documentos, devem ser compreendidas como encadeamentos de ações sequenciais, que delimitam procedimentos específicos e responsabilidades para os agentes envolvidos em cada etapa.

Cabe ressaltar que não se trata de rupturas, mas de interfaces operacionais e sistêmicas entre ambientes. Assim, se em um primeiro momento os dados se encontram sob a responsabilidade da equipe de pesquisa que os produziu e utilizou no processo investigativo, em seguida a manutenção necessária à sua gestão e uso passa a ser de responsabilidade da equipe gestora do repositório em que foram depositados para compartilhamento.

A partir dessa estrutura integrada, torna-se essencial aprofundar a análise sobre o ambiente “gestão inicial”, entendido como a fase em que os dados de pesquisa ainda se encontram sob a responsabilidade direta dos pesquisadores ou das equipes de pesquisa, antes de sua ingestão no repositório de gestão de dados. É nesse momento que se definem os parâmetros que iniciam e orientam o registro da proveniência, do contexto e da fixidez ao longo do ciclo de vida dos dados de pesquisa, sendo, portanto, um momento determinante para a conformação da preservação digital qualitativa.

A seguir, apresenta-se o Modelo de fluxo para gestão inicial de dados de pesquisa, que concatena todo o conhecimento adquirido no desenvolvimento desta pesquisa articulado sistemicamente com os fundamentos arquivísticos e orientados pela Ciência Aberta.

Figura 51 – Modelo de fluxo para gestão inicial de dados de pesquisa



Fonte: Elaborada pela autora (2025).

No Modelo de fluxo para gestão inicial de dados de pesquisa, apresentado na Figura 51 dialoga com o Quadro 23 e tem a seguinte constituição:

O Modelo é iniciado com o planejamento da pesquisa. Nessa etapa são realizadas:

- a identificação das fontes e dos recursos necessários;
- a definição de padrões de formatos e nomenclatura;
- a análise de possíveis dependências tecnológicas;
- a ordenação das etapas e ações da pesquisa;
- o delineamento de estratégias para criação e/ou coleta de dados e para o registro de metadados¹²⁰;
- a verificação das políticas e procedimentos institucionais aplicáveis;
- a definição de *softwares* a serem empregados;
- as responsabilidades dos agentes envolvidos;

¹²⁰ O registro dos metadados descritivos e técnicos contribui para documentar ações e, se bem descritos e completos, também contribui para evidenciar características e o valor de evidência do dado de pesquisa.

- a elaboração do Plano de Gestão de Dados (PGD) inicial, já que, conforme demonstrado na subseção 3.2.2 desta tese, deve ser atualizado conforme demandas que se apresentam ao longo do processo científico;
- o estabelecimento do fluxo e das estratégias para a gestão inicial e assistida dos dados de pesquisa.

A etapa seguinte corresponde à criação e/ou coleta dos dados de pesquisa, onde deve ser assegurado:

- a produção de dados de pesquisa em formatos recomendados;
- a transformação, quando necessária, de modo controlado para garantir a integridade e a interoperabilidade dos dados;
- a organização lógica dos arquivos;
- a adoção de padrões institucionais de nomenclatura e desenho de planilhas;
- a observância das boas práticas de gestão de dados conforme as normas da instituição responsável pela pesquisa;
- a estruturação do plano de registro e coleta de metadados;
- o registro sistemático das informações relacionadas aos dados, *softwares*, fluxos de trabalho, métodos, insumos, códigos e demais recursos empregados; e
- a documentação de todas as atividades realizadas, reunindo os PGD, os documentos de contextualização da pesquisa, os registros institucionais de suporte à gestão e preservação dos dados de pesquisa e outros documentos considerados relevantes para fundamentar as decisões tomadas até esta etapa.

No que diz respeito especificamente à produção de dados de pesquisa em formatos recomendados, é importante que sejam seguidas as normativas definidas pela instituição à qual a pesquisa se vincula, priorizando-se o uso de formatos abertos e editáveis. Compreende-se, no entanto, que dados de pesquisa são objetos digitais complexos e, muitas vezes, produzidos e/ou coletados, exclusivamente, por *softwares* ou aplicativos proprietários que geram formatos fechados. Nesse caso, devem ser analisados caso a caso.

Para os dados de pesquisa originais, do gênero audiovisual, iconográfico, sonoro ou textual, a título de referência, o Quadro 24 apresenta os formatos de preservação recomendados e os formatos aceitáveis, em casos específicos, para a fase de gestão inicial

dos dados de pesquisa, conforme orientação contida no Plano de Gestão de Documentos Arquivísticos Digitais da Fiocruz (2024).

Quadro 24 – Formatos de arquivo recomendados e aceitos para a fase de gestão inicial dos dados de pesquisa

Gênero documental	Formatos de preservação recomendados	Formatos aceitáveis
Audiovisual	* Contêiner: MKV (Matroska Multimedia Container) / Codec: FFV1 * Contêiner: AVI (Audio Video Interleave) / Codec: v210 * Contêiner: MOV (QuickTime) / Codec: v210 * Contêiner: MXF / Codec: JPEG2000	* Contêiner: MP4 (MPEG-4) / Codec: h.264; h.265; * ProRes; DV25; DV50
Iconográfico	* TIFF (TIFF version 6 uncompressed)	* JPEG * PNG (Portable Network Graphics)
Sonoro	* AIFF (Audio Interchange File Format) * ALAC (Apple Lossless Audio Codec) * BWF (Broadcast Wave File) * FLAC (Free Lossless Audio Codec) * WAV (Waveform Audio File Format)	* AAC (Advanced Audio Coding) * MP3 (MPEG-1/2 * Audio Layer 3)
Textual	* Portable document format for preservation (PDF/A) * Texto simples (TXT)	* Microsoft Word (DOC) (DOCX) * OpenDocument text (ODT) * Portable Document Format (PDF)

Fonte: Fiocruz (2024).

Portanto, é interessante que, na medida do possível, os dados de pesquisa já sejam depositados nos repositórios de gestão nesses formatos, o que reduz custos de migração e riscos para (re)uso.

Dando sequência ao fluxo, a terceira etapa consiste na identificação dos dados, momento em que há:

- a determinação de quais são os dados efetivamente produzidos/coletados pela pesquisa (dados de pesquisa);
- o reconhecimento dos dados utilizados na pesquisa (dados para pesquisa); e
- a identificação dos dados utilizados apenas como referência.

Essa distinção é essencial para definir as estratégias de gestão, assegurar o tratamento técnico adequado dos dados de pesquisa e avaliar a necessidade de recursos para a preservação digital qualitativa dos dados de pesquisa e, eventualmente, dos dados para pesquisa, caso sejam entendidos como necessários à contextualização do conjunto de dados de pesquisa.

Essa etapa precede a aplicação da classificação e da avaliação, sendo essencial para o tratamento desses dados, permitindo determinar padrões e definir estratégias que busquem garantir o acesso contínuo, confiabilidade e a presunção de autenticidade dos dados de pesquisa, além de tornar possível conhecer o volume e embasar uma previsão da capacidade de armazenamento necessária tanto na gestão inicial, quanto na gestão assistida e na preservação permanente.

Na continuidade do processo, em seguida, ocorre a classificação, que deve considerar critérios conforme descritos na subseção 5.2.1.1. Nessa fase, tem-se:

- a categorização dos dados (Sayão; Sales, 2019) quanto à(ao):
 - origem;
 - grau de processamento;
 - abordagem da pesquisa;
 - natureza;
 - nível de sensibilidade;
 - materialidade; e
 - perenidade.
- a definição dos prazos mínimos de retenção;
- a análise e estabelecimento do fluxo ideal dos dados de pesquisa.

Todo esse processo e decisões tomadas devem ser registrados de maneira persistente como metadados inerente à gestão dos dados de pesquisa. Geralmente essa informação consta como metadado Assunto (identificador *Subject*).

A próxima etapa corresponde ao uso dos dados de pesquisa, na qual se enquadram:

- a observação dos acordos de governança, dos direitos autorais e de propriedade intelectual;
- as possibilidades e condições de colaboração entre equipes;
- as análises;
- os processamentos;
- as transformações; e
- a obtenção de resultados.

Todos os procedimentos devem ser documentados e acompanhados por registros contínuos e cumulativos de metadados. Além disso, nessa etapa, recomenda-se a elaboração do arquivo *Read.me*, com as informações específicas do domínio junto às descrições que auxiliem a interpretação, renderização e contextualização dos dados de pesquisa.

Quando pertinente, conforme exposto na subseção 6.3 desta tese, pode ser necessária uma declaração da integridade ou das limitações do conjunto de dados como contribuição para que a análise da qualidade dos dados disponibilizados.

A última etapa da gestão inicial dos dados de pesquisa é a avaliação prévia que se pauta nos prazos de retenção definidos durante a classificação. Nessa fase são realizadas:

- a atribuição de valores primário e secundário;
- a análise dos seguintes critérios, conforme Sá e Pereira (2024):
 - Requisitos legais, éticos ou outras razões regulatórias;
 - Requisitos do instituto, agência de fomento ou editor;
 - Importância dos dados;
 - Unidade dos dados;
 - Valor dos dados;
 - Significativo interesse público;
 - Questões econômicas e financeiras;
 - Singularidade;
 - Potencial para (re)uso; e
 - Critérios de seleção secundários.
- a seleção dos dados a serem eliminados;

- a eliminação dos dados selecionados para tal, mediante procedimentos padronizados e devidamente documentados;
- a seleção dos dados passíveis de compartilhamento; e
- a transferência dos dados selecionados para serem mantidos em repositórios de gestão.

É importante que a temporalidade prevista para os dados de pesquisa compartilhados no repositório de gestão esteja exposta como uma informação atrelada persistentemente ao *dataset*. Essa ação imprime transparência aos pesquisadores que venham a fazer uso desses dados, demonstra que os dados são objetos de gestão contínua e comprometida, além de fundamentar com registros informacionais a avaliação final que será realizada posteriormente.

Ao proceder uma eliminação de dados de pesquisa, de maneira criteriosa e registrada, reforça-se o compromisso com a sustentabilidade não apenas sob o ponto de vista dos recursos da gestão institucional como também sob as perspectivas ambiental e da própria Ciência.

Quanto aos dados de pesquisa selecionados para compartilhamento, estes devem ser preparados enquanto pacote de submissão, o que se entende por SIP (Pacote de Informações para Submissão), aos moldes do que se preconiza no Modelo de Referência OAIS (2024), conforme descrito na subseção 4.1 desta tese. Nesse pacote, os dados de pesquisa são empacotados junto à toda documentação necessária para contextualização e aos metadados registrados até esse momento de seu ciclo.

Essa transferência é iniciada no “ambiente gestão inicial”, conforme figura 49 “Ciclo de vida dos dados de pesquisa e seus ambientes”, onde o produtor e sua equipe de pesquisa são os responsáveis pelas ações de gestão. Os dados de pesquisa, nessa etapa estão armazenados em recursos tecnológicos como computador, sistema, aplicativos sob sua ingerência. Contudo, é importante reforçar a importância do acompanhamento de profissionais especializados nessa gestão para que ela possa ser, efetivamente e adequadamente, realizada, conforme disposto na subseção 5.2.2 desta tese.

Isso porque a preparação do pacote SIP demanda que seja criado um dossiê contendo duas pastas: “*metadata*” e “*objects*”, em consonância com a figura 16 “Composição das Informações de Descrição de Preservação (PDI)”, de Machado (2023).

Nesse dossiê, a pasta “*metadata*” deve conter os metadados relativos aos dados de pesquisa selecionados. A transferência desses metadados a essa pasta deve ser feita mediante uso do aplicativo TeraCopy¹²¹. Já na pasta “*objects*” são inseridos, via TeraCopy também, os dados em si. Após esse passo, deve ser feito o empacotamento em BagIt para a geração do *checksum*, enquanto um registro de “informação de fixidez”.

Caso haja documentação relevante para a interpretação e contextualização dos dados a serem transferidos, a constituição desse SIP deve conter mais uma pasta em seu dossiê, a chamada “*Submission Documentation*”. Nesse cenário, os dados de pesquisa ficam associados aos metadados e à documentação necessários, de maneira cumulativa e contínua, evidenciando suas informações de contexto, proveniência e fixidez.

Após esse processamento, o referido SIP deve ser submetido à conferência do curador. Ao ser validado, esse pacote é submetido à ingestão no repositório de gestão de dados escolhido pela instituição, com a observância das eventuais restrições.

No âmbito do suporte tecnológico, cabe apontar que o uso de um sistema que apoie a execução dessa gestão inicial, como os cadernos eletrônicos de laboratório (ELN), por exemplo, colabora na otimização dos recursos da pesquisa, automatiza algumas funções previstas no fluxo e propicia o registro de todas as informações adicionais correlacionados aos dados, aos seus metadados, aos eventos aos quais foram submetidos e seus respectivos agentes. Portanto, a utilização de uma ferramenta tecnológica para esse momento, junto ao apoio de um profissional dedicado à função da gestão inicial aumenta as chances de que essas ações sejam, efetivamente, realizadas e, então, deve ser considerada pelas instituições de pesquisa científica.

Contudo, mesmo com a adoção de um *software*, as etapas inerentes à essa gestão inicial precisam ser, necessariamente, planejadas, padronizadas e normatizadas institucionalmente para, então, serem realizadas de forma automatizada. Isso incrementa a composição de uma Cadeia de preservação íntegra e confiável para a manutenção de dados de pesquisa autênticos, conforme exposto na subseção 2.3.2 desta tese. Além disso, o uso de mecanismos de registros viabiliza o fortalecimento dessa Cadeia, já que integra

¹²¹ “O TeraCopy é um utilitário de cópia desenvolvido pela Code Sector, uma empresa de software fundada por Oleg Kupershtakhin. Ele substitui as funções padrão de copiar e colar do Windows Explorer e oferece vários recursos avançados para otimizar a transferência de arquivos entre dispositivos” (Disponível em: <https://www.controle.net/faq/teracopy-utilitario-de-backup-mais-eficiente-que-windows-explorer>).

um sistema de controles desde o início do ciclo de vida dos dados de pesquisa, constituindo um ambiente seguro e rastreável.

Salienta-se a importância de que o *software* escolhido como repositório de gestão de dados de pesquisa tenha a capacidade técnica e funcional para absorver todos os metadados e documentação do SIP, constantes nas pastas “*metadata*” e “*submission documentation*” incorporando-os aos demais metadados produzidos após a submissão no repositório de gestão, incluindo-os nos metadados de exportação do *dataset*. Essa composição deve ser realizada em um documento METS, por exemplo, como descrito na subseção 4.2.1 desta tese.

Mais uma vez, essa iniciativa reforça a construção da Cadeia de Preservação, já que todos os metadados até então registrados passam a ficar persistentemente acoplados aos seus respectivos dados de pesquisa.

Caso o *software* do repositório de gestão não viabilize essa ação, é preciso que esses metadados sejam inseridos no sistema enquanto arquivos de dados. Nesse caso, há necessidade de diferenciá-lo dos demais arquivos disponibilizados. Para isso é possível utilizar a *tag* “documentação” para os metadados, enquanto os dados em si podem ser taguados como “dado”.

A partir desse ponto do fluxo, o dado de pesquisa passa a compor o rol de atividades do que está sendo chamado de “Gestão assistida” e “Preservação”.

Assim, em virtude de toda complexidade envolvida na gestão inicial dos dados de pesquisa, é relevante apontar a necessidade de que a instituição de pesquisa desenvolva uma política que defina responsabilidades, escopo, requisitos, princípios, metodologias e sistemas a serem utilizados, formulando e documentando o desenvolvimento sistemático dos procedimentos de gestão, estabelecendo uma linha estratégica de mecanismos de monitoramento e revisão periódica.

Além disso, é preciso evidenciar que o Modelo de fluxo proposto pode fundamentar o desenvolvimento de políticas em instituições que fazem, ou não, uso de sistemas para a gestão inicial dos dados de pesquisa, visto que aborda ações que devem ser realizadas independentemente do meio pelo qual serão efetivadas (automatizada ou manualmente).

Diante do exposto, é possível identificar que ao implantar o Modelo de fluxo para gestão inicial de dados de pesquisa, as lacunas informacionais identificadas na análise empírica podem ser solucionadas ao longo de suas etapas. O Quadro 26 apresenta essa

correspondência, evidenciando em que momento do fluxo cada categoria de informação pode ser registrada integralmente, utilizando por parâmetro o “Quadro-base para quantificação de tipos informacionais” (Quadro 9) para que a “identificação de registros informacionais dos *datasets* selecionados” (Quadro 21), possa ser melhorado.

Quadro 255 - Correspondência entre as lacunas informacionais encontradas e etapas do fluxo de gestão inicial dos dados de pesquisa

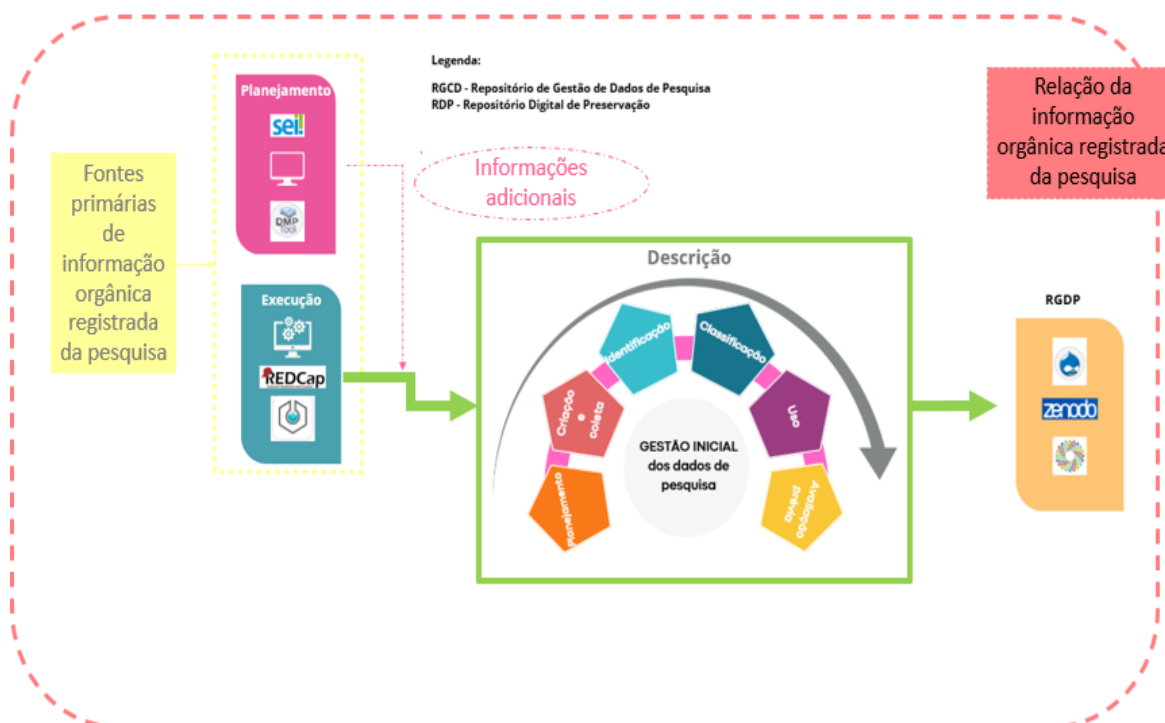
Categoria de informação (Subseção 6.1.1 – Figura 34)	Lacuna informacional identificada na análise empírica	Etapas do Modelo de fluxo para gestão inicial de dados de pesquisa (Figura 51)
Proveniência	Ausência de evidência processuais com vinculação dos dados de pesquisa, eventos, agentes e outras informações associadas à criação, gestão e preservação do Objeto Digital.	1. Plano de Gestão de Dados (PGD)
		3. Identificação
Contexto	Falta de documentação sobre o contexto metodológico da coleta, produção e manutenção dos dados de pesquisa (instrumentos, parâmetros, datas, locais, <i>softwares</i> etc.).	2. Criação/Coleta
		3. Identificação
		4. Classificação
Fixidez	Inexistência de controle de integridade e versionamento dos arquivos digitais ao longo do processo de pesquisa até seu depósito no repositório de gestão de dados de pesquisa.	2. Criação/Coleta
		5. Uso
		6. Avaliação Prévia
Proveniência/Contexto	Deficiências na descrição técnica e temática dos dados (metadados insuficientes para renderização e (re)uso).	1. Plano de Gestão de Dados (PGD)
		3. Identificação
		4. Classificação
	Falta de critérios consistentes para avaliar o valor dos dados e definir sua destinação (preservação ou descarte) e de registros que evidenciem essa avaliação.	6. Avaliação Prévia

Fonte: Elaborado pela autora (2025).

Essa sistematização explícita no Quadro 26 demonstra a coerência entre os resultados da análise empírica e as contribuições arquivísticas delineadas no Modelo de fluxo proposto nesta tese.

Com isso, evidencia-se que o Modelo de fluxo, quando aplicado ao ciclo de vida dos dados de pesquisa, possibilita a manutenção da relação orgânica de todas as fontes primárias de informações registradas na pesquisa.

Figura 52 – O Modelo de fluxo para gestão inicial de dados de pesquisa no registro da informação orgânica da pesquisa

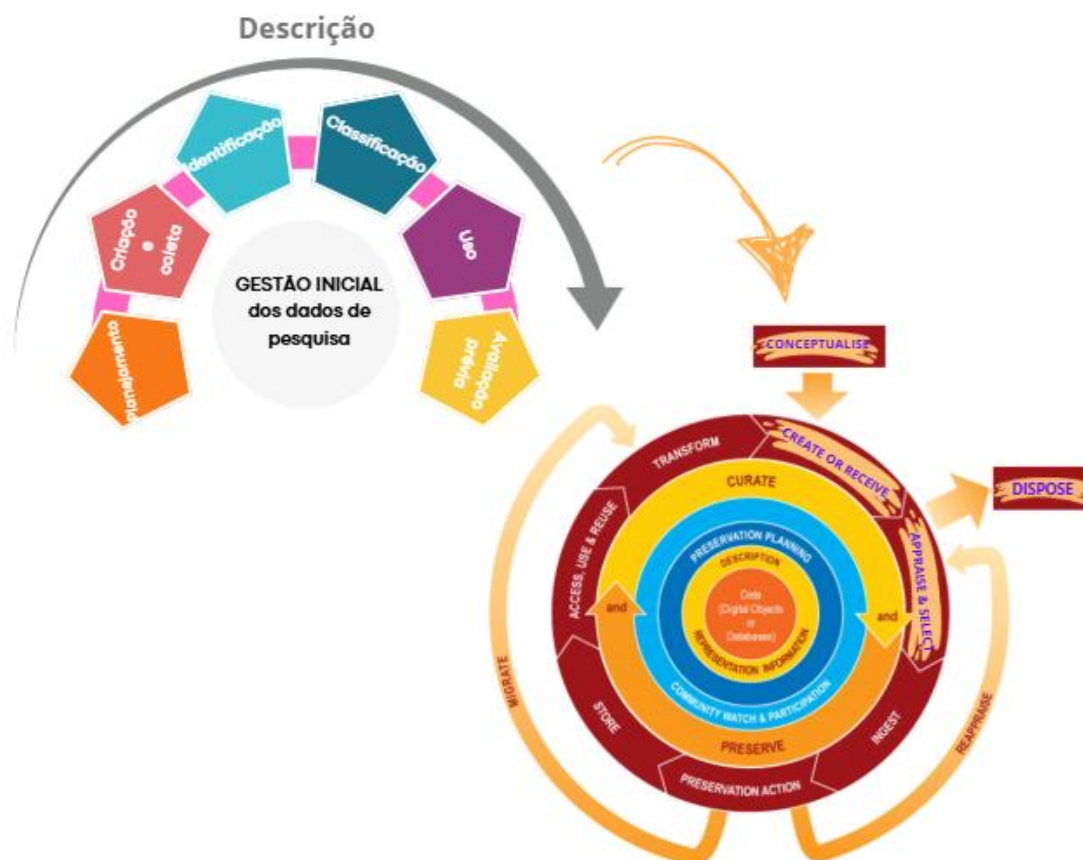


Fonte: Elaborada pela autora (2025).

A Figura 52 evidencia que o modelo preserva a organicidade da informação ao longo do ciclo, ao manter encadeados dados, metadados e documentação.

Adicionalmente, a partir da leitura arquivística dos resultados empíricos, elaborou-se uma representação integradora que demonstra como as ações de gestão inicial dos dados de pesquisa se articulam ao Modelo do ciclo de vida da Curadoria Digital, proposto pelo DCC (2008), exposto na subseção 3.1.2 desta tese. Essa representação, apresentada abaixo, permite visualizar a correspondência entre as práticas arquivísticas e as etapas do ciclo de vida da curadoria, evidenciando o papel da gestão inicial como instância estruturante do processo de preservação digital qualitativa dos dados de pesquisa.

Figura 53 - Articulação entre a gestão inicial dos dados de pesquisa e o modelo de Curadoria Digital do DCC



Fonte: Elaborada pela autora (2025).

Do lado esquerdo da Figura 53 são apresentadas as ações essenciais baseadas nos fundamentos arquivísticos: planejamento da pesquisa (com elaboração do Plano de Gestão de Dados - PGD), criação e coleta, identificação, classificação, uso e avaliação prévia, compreendidas em um conjunto de práticas de descrição, que devem incorporar o registro de metadados e a documentação de forma contínua e cumulativa.

Enquanto o DCC descreve o percurso de tratamento técnico e preservação digital (do *conceptualise* ao *transform*), a proposta arquivística em questão destaca o momento anterior que ampara a execução das etapas *conceptualise*, *create or receive*, *appraisal & select* e *dispose*, compondo uma documentação contínua e cumulativa que visa assegurar a autenticidade e confiabilidade desses dados.

Assim, a Figura 53 evidencia que a gestão inicial dos dados de pesquisa não substitui o modelo do DCC, **complementa-o e qualifica-o**, oferecendo critérios técnicos e conceituais para a preservação digital qualitativa sob uma mesma lógica de ciclo de vida ao revelar a interlocução da Arquivologia com a Curadoria Digital.

Em termos epistêmicos, o resultado traz uma abordagem aplicada, capaz de articular gestão de documentos arquivísticos, curadoria digital e preservação digital sob uma mesma lógica de ciclo de vida.

Conclui-se, portanto, que a adoção dos fundamentos arquivísticos desde as fases iniciais da gestão de dados de pesquisa constitui um requisito estratégico para a Ciência Aberta, ao assegurar elementos transparentes de confiabilidade e sustentar a qualidade necessária para o (re)uso desses dados. A integração entre gestão arquivística, curadoria digital e preservação digital configura a principal contribuição desta tese, demonstrando que a Arquivologia pode atuar como referencial estruturante para práticas de gestão inicial de dados de pesquisa, qualificando processos, fortalecendo a proveniência digital e garantindo as condições para a preservação digital qualitativa ao longo de todo o ciclo de vida.

8 CONCLUSÃO E PERSPECTIVAS FUTURAS

Ao considerar o contexto científico como um campo de contradições e disputas (Bourdieu, 1983) no qual a capitalização do conhecimento no capitalismo cognitivo (Lazzarato, 2010) incentiva a apropriação e o controle de informação até mesmo públicas e socialmente produzidas (Albagli, 2015), a Ciência Aberta é um movimento contracorrente ao *status quo* atual.

A possibilidade de abrir dados e processos de pesquisa em um cenário onde os dados são considerados o “novo petróleo” tem impactos econômicos, sobretudo na direção de uma ciência mais equitativa, sustentável e democrática, especialmente quando se trata de dados produzidos a partir de financiamento com recursos públicos, que devem ser publicamente acessíveis, de acordo com a perspectiva da escola de pensamento democrático da Ciência Aberta, trazida por Fecher e Friesike (2013).

Também há impactos sociais já que a ciência deve servir à humanidade (“Declaração sobre Ciência e o Uso do Conhecimento Científico”, 1999). Diante do crescimento de *fake news* e do negacionismo (Perini, 2019), a transparência dos dados e seus processos são relevantes para elevar a confiança da sociedade na ciência.

Diante de todo esse cenário, na tentativa de adequar as práticas científicas às premissas da Ciência Aberta, dados de pesquisa são cada vez mais disponibilizados em repositórios voltados à gestão e ao compartilhamento, o que suscita reflexões sobre sua qualidade para o (re)uso e a justificativa dos esforços de preservação a longo prazo. Partiu-se, portanto, do **pressuposto** de que a gestão de dados de pesquisa, ao abranger ações estratégicas para o registro e a manutenção de informações que sustentam a confiabilidade desses objetos digitais, constitui condição *sine qua non* para a efetivação de uma preservação digital qualitativa.

Esse panorama evidenciou tanto as dificuldades quanto a relevância das ações voltadas à manutenção do acesso e da qualidade dos conteúdos informacionais ao longo do tempo. Essas ações exigem atenção e procedimentos não apenas sobre os próprios dados de pesquisa, mas também sobre todas as informações essenciais que amparam seu (re)uso, possibilitando a aferição de sua proveniência, autenticidade e confiabilidade.

O (re)uso de dados de pesquisa pressupõe sua preservação a curto, médio e longo prazos, não apenas em termos de integridade e qualidade informacional, mas sobretudo de confiabilidade. Para que possam ser efetivamente reutilizados em novos contextos

científicos, é essencial que esses dados sejam acompanhados de registros que evidenciem sua trajetória, desde a criação até o armazenamento final. Tais registros atuam como garantias de autenticidade, permitindo que os dados sejam validados como provas legítimas da pesquisa e de seus resultados. Portanto, a documentação contínua e cumulativa ao longo do ciclo de vida dos dados de pesquisa é indispensável para assegurar e viabilizar novos usos.

Reconheceu-se que as estratégias preservacionistas devem ser implementadas não apenas nas fases avançadas do ciclo de vida dos dados, mas desde o planejamento da pesquisa e durante todo seu decorrer. Dessa forma, tornou-se relevante considerar aspectos que ultrapassam a simples manutenção de acesso ao objeto digital, incluindo o fornecimento de informações sobre o contexto de produção e o histórico de manutenção dos dados, que fundamentam a confiança necessária ao seu (re)uso. Essa confiança, entretanto, não deve ser amparada, exclusivamente, na disponibilização em repositórios de gestão de dados confiáveis nem na credibilidade institucional de quem fornece seu acesso.

A partir da revisão de literatura, percebeu-se que embora não seja um tema pouco explorado, a abordagem dada à preservação digital de dados de pesquisa tem se voltado a ações que precisam ser planejadas e executadas para a viabilização efetiva de manter dados acessíveis e confiáveis, porém numa visão consubstanciada a partir de sua disponibilização em repositórios que os compartilhem.

O foco, por vezes limitado, à manutenção da integridade e confiabilidade dos dados de pesquisa nessas ferramentas deixou às sombras o fato de que mantê-los confiáveis no repositório em questão, e em diante, não pressupõe que os dados tenham sido mantidos suficientemente confiáveis até o momento de seu depósito. Informações relativas a todo o tratamento que foi a eles dispensados são essenciais para que essa confiança possa ser cristalizada e mantida quando eles passam a compor um conjunto de dados em um *software* para tal.

Julgou-se relevante destacar que, em se tratando de evidências, conforme demonstrado ao longo desta tese, os dados oriundos de pesquisas científicas devem ser disponibilizados em repositórios de gestão de dados como tais, isto é, como registros que testemunham o processo científico. Essa perspectiva difere substancialmente daquela aplicada, no contexto nacional, aos produtos da pesquisa depositados em repositórios institucionais.

Factualmente, tal questão perpassa e impõe-se à gestão de dados de pesquisa (Research Data Management – RDM), especialmente em seus estágios iniciais, ou seja, no período compreendido entre a criação desses dados e seu depósito em repositórios de gestão. Todavia, a metodologia de execução dessas ações tem sido pouco explorada na literatura especializada ou evidenciada na prática.

Observa-se um incentivo crescente para que os pesquisadores depositem seus dados de pesquisa em repositórios de gestão, contudo é preciso reconhecer que esses ambientes não constituem a origem desses dados. Embora sejam ferramentas relevantes e possam integrar os estágios iniciais ou acompanhar o percurso da pesquisa, o processo de gestão começa muito antes de sua submissão a esses sistemas. Paralelamente, nota-se um estímulo cada vez maior, por parte dos principais *stakeholders*, como agências de fomento, editores, sociedades científicas, instituições e os próprios pesquisadores, à elaboração de Planos de Gestão de Dados (PGD).

Diante das observações possibilitadas por esta tese, identificou-se a adoção de um fluxo fundamentado na curadoria digital de dados de pesquisa, com base em modelos como o DCC, que abrange tanto a gestão de dados de pesquisa (RDM) quanto sua preservação digital. Esse fluxo envolve o uso de instrumentos de planejamento e acompanhamento, como os Planos de Gestão de Dados (PGD), bem como o emprego de ferramentas tecnológicas, entre elas, repositórios para gestão e compartilhamento de dados, como o Dataverse, e a intermediação de curadores, que atuam na recuperação, no acesso e no compartilhamento desses recursos digitais conforme políticas de acesso aberto.

Idealmente, os conjuntos de dados com potencial de (re)uso futuro e valor permanente devem ser recolhidos, em momento oportuno, para repositórios digitais de preservação desenvolvidos com base no Modelo de Referência OAIS (2024).

Nessa conjuntura, observa-se a ausência de transparência sobre como a gestão dos dados de pesquisa é realizada durante as etapas de criação, coleta, colaboração e transformação, enquanto permanecem sob a responsabilidade direta do pesquisador ou de sua equipe. Considerando que a preservação de dados (re)usáveis requer o registro sistemático e a manutenção contínua de todas as informações necessárias à garantia de sua confiabilidade, torna-se imperativo que a documentação das ações executadas ao longo do ciclo de vida desses dados seja incorporada de forma indissociável a eles.

A partir da análise empírica realizada, **confirma-se a hipótese** desta tese de que **registros abertos, documentados e integrados desde as fases iniciais do ciclo de vida dos dados de pesquisa são não apenas necessários, mas também determinantes para a promoção de sua preservação digital qualitativa, a fim de fortalecer sua confiabilidade e seu potencial de (re)uso.**

Isso implica reconhecer que qualquer lacuna informacional, em qualquer etapa do ciclo de vida dos dados de pesquisa, pode comprometer a qualidade do que é compartilhado e, conseqüentemente, seu potencial de (re)uso. Dessa forma, quando a qualidade para (re)uso é comprometida, o esforço de preservação pode não se justificar, pois atributos insuficientes para demonstrar confiabilidade desqualificam a necessidade de sua manutenção.

Esse entendimento evidencia a centralidade das estratégias de gestão adotadas nas fases iniciais do ciclo de vida (anteriores à própria intervenção preservacionista) e revela a importância da gestão dos metadados como elemento estruturante desse processo.

Nessa perspectiva, os registros sobre a proveniência, o contexto e a fixidez dos conjuntos de dados de pesquisa assumem papel essencial na análise de sua confiabilidade, especialmente no cenário digital atual. Embora a gestão desses dados abarque ações ampliadas ao longo de todo o ciclo de vida, observa-se com frequência a predominância de estímulos que não contemplam sua totalidade, abrindo brechas que comprometem a confiança integral dos dados de pesquisa admitidos em repositórios destinados à sua gestão e compartilhamento. A manutenção dos dados acompanhados de metadados que registrem informações adicionais desde as fases iniciais desse ciclo contribui significativamente para mitigar essas lacunas.

Ainda que os repositórios de gestão de dados estejam aderentes a normas internacionais de segurança e preservação, e utilizem *softwares* que garantam a confiabilidade dos objetos digitais em seu ambiente, isso não assegura que os dados tenham chegado a esses repositórios em condições qualitativamente confiáveis e acessíveis. Se os dados de pesquisa constituem evidências da própria pesquisa, não é adequado que sejam tratados ou disponibilizados em repositórios de gestão e compartilhamento como meros produtos.

Diante disso, o aprofundamento do conhecimento sobre Ciência Aberta, Curadoria Digital, Preservação digital e Gestão de documentos (**objetivo específico 1**),

realizado a partir da revisão de literatura sobre os referidos temas, conforme exposto nas seções 2 “Ciência Aberta em construção: fundamentos epistemológicos, desafios e a qualidade dos dados para o (re)uso”, 3 “O escopo da curadoria digital para o (re)uso confiável dos dados de pesquisa” e 4 “A preservação digital qualitativa: o registro de metadados para o fortalecimento do (re)uso dos dados de pesquisa” e 5 “A gestão de documentos: contributo da Arquivologia para a qualificação da preservação digital dos dados de pesquisa” embasou o estabelecimento do referencial teórico que subsidiou a compreensão necessária para a proposição do modelo de fluxo, produto essencial desta tese.

Em seguida, o percurso da pesquisa de campo (seção 6), com a delimitação do recorte empírico, especificando as categorias analíticas e os níveis de registro informacional, os critérios de análise e a limitação da pesquisa (subseção 6.1), juntamente com a análise quanti-qualitativa (subseção 6.2), permitiu a verificação do grau de maturidade da gestão aplicada antes do depósito dos dados de pesquisa nos repositórios do recorte a partir da análise da transparência de seus metadados foi conduzida por meio da delimitação do recorte (subseção 6.3), o que levou ao atingimento do **objetivo específico 2**.

Com isso, o **objetivo específico 3** foi realizado a partir de evidências sobre a importância dos estágios iniciais da gestão de dados de pesquisa para viabilizar a efetivação da preservação digital qualitativa à luz do potencial de confiabilidade dos dados de pesquisa (subseção 6.4).

Finalmente, diante de toda pesquisa o **objetivo específico 4** também foi atingido quando foi apresentada a proposição do “Modelo de fluxo para gestão inicial de dados de pesquisa” (subseção 7.2). Esse Modelo, ao articular fundamentos arquivísticos às exigências da Ciência Aberta, da Curadoria digital e da Preservação digital, delineia etapas que operacionalizam registros de proveniência, contexto e fixidez em um fluxo aplicável e replicável e, portanto, tem potencial de aplicação transversal e institucional, podendo subsidiar políticas de gestão e preservação de dados de pesquisa em instituições científicas e acadêmicas.

Ao atingir todos os objetivos específicos **concluiu-se o objetivo geral desta tese**: propor a gestão arquivística de documentos como fundamentação teórico-metodológica para o aprimoramento da gestão de dados de pesquisa, desde seus estágios iniciais, com vistas a viabilizar a preservação digital qualitativa.

Esse percurso evidencia a Arquivologia como um referencial teórico capaz de dialogar com os fundamentos epistemológicos da Ciência Aberta, ao traduzir seus fundamentos clássicos, em prol da proveniência, organicidade e autenticidade, em instrumentos operacionais para a curadoria e preservação digital. Assim, a pesquisa demonstra que o campo arquivístico não apenas oferece fundamentos conceituais, mas também uma estrutura metodológica para solucionar desafios informacionais contemporâneos.

Cumpra esclarecer que, inicialmente, a proposta dessa tese era propor um fluxo de preservação digital dos dados de pesquisa produzidos em cadernos eletrônicos de laboratório. Todavia, assim como a maioria dos percursos científicos, obstáculos foram se apresentando, levando esta pesquisadora a recalcular a rota, mas sem perder de vista aquilo que acredita, defende desde o mestrado (2013-2015) e que tem o privilégio de praticar em sua rotina profissional: a gestão de documentos (e agora de dados de pesquisa) como condição *sine qua non* para a preservação (agora refinado com o entendimento da necessidade dos aspectos qualitativos para sua efetiva e contínua realização).

Ao revisitar os fundamentos da área e reinterpretá-los diante do cenário digital, esta tese contribui para o reposicionamento da Arquivologia como disciplina científica cuja atuação ultrapassa o tratamento documental tradicional e alcança o domínio da gestão de dados de pesquisa. Essa ampliação reafirma a relevância da Arquivologia como mediadora entre práticas informacionais, tecnológicas e científicas.

Reforça-se que a preservação digital qualitativa, tal como delineada nesta tese, ultrapassa a dimensão técnica da manutenção de *bits* para abarcar a permanência de contexto e caráter de evidência dos dados de pesquisa, sustentada em registros de proveniência, contexto e fixidez que lhes conferem autenticidade e fortalecem sua confiabilidade.

Assim, o percurso iniciado pela identificação de lacunas informacionais nos dados de pesquisa disponibilizados em repositórios de gestão de dados de três instituições brasileiras culmina em uma proposta metodológica fundamentada que busca saná-las por meio da gestão de documentos, reafirmando a importância dos registros de metadados e documentação contínua e cumulativa, que também devem ser geridos junto a esses dados, a partir da agregação de elementos que fortalecem a sua confiabilidade, enquanto pilar da preservação digital qualitativa.

É importante esclarecer que compartilhar dados de pesquisa exige subjetividade e engajamento, dessa forma, questões estratégicas devem ser consideradas. A gestão desses dados sem o estabelecimento de uma governança que alicerce cria obstáculos para sua adoção em várias instâncias. Ademais, o contexto para aplicação do que foi desenvolvido nesta tese também deve ser observado e, como todo modelo tem limitações de aplicação, é preciso que o Modelo ora proposto seja analisado e adaptado à cada realidade.

Encerrando este percurso, após demonstrar o diálogo entre a análise empírica e os fundamentos teóricos, confirma-se que a gestão de documentos, quando aplicada desde as fases iniciais do ciclo de vida dos dados de pesquisa, constitui o alicerce da preservação digital qualitativa. Esse resultado sintetiza a integração entre empiria e teoria, demonstrando que os fundamentos arquivísticos mantêm sua força explicativa e operacional mesmo diante dos novos ambientes digitais. A consolidação dos resultados obtidos abre novas possibilidades de aprofundamento, especialmente no que tange ao desenvolvimento de práticas e instrumentos que ampliem a efetividade da gestão inicial dos dados de pesquisa.

As **perspectivas futuras** desdobram-se em três dimensões complementares: teórica, metodológica e tecnológica.

a) Teórica: mais estudos sobre as articulações entre a Arquivologia e Ciência Aberta, com a proposição de linhas de pesquisa para avaliação de maturidade da gestão de dados de pesquisa e/ou modelagem semântica de proveniência, e o apontamento da necessidade de políticas institucionais sustentadas pelo modelo;

b) Metodológica: validações interdisciplinares do modelo de fluxo proposto, além de aplicação, testes, registro e publicação de resultados em diferentes áreas disciplinares para aprimorá-lo diante de realidades específicas que possam embasar capacitações para a execução dessa gestão pelas equipes responsáveis; e

c) Tecnológica: com os dois aspectos anteriores em andamento, é possível contribuir no desenvolvimento de ferramentas tecnológicas que apoiem a gestão inicial dos dados de pesquisa e/ou aprimorar as ferramentas já existentes.

Essas possibilidades de aprofundamento reforçam o papel integrador da Arquivologia no ecossistema da Ciência Aberta, apontando caminhos para o aperfeiçoamento de metodologias, políticas institucionais e tecnologias que garantam a confiabilidade, interoperabilidade e sustentabilidade dos dados de pesquisa a longo prazo. Além disso, as implicações desta pesquisa extrapolam o campo acadêmico para alcançar

também o campo social, ao contribuir para a consolidação de práticas científicas mais transparentes, confiáveis e sustentáveis. Ao propor um modelo de fluxo para gestão inicial de dados de pesquisa, embasado em fundamentos arquivísticos, esta tese oferece caminhos concretos para que a preservação digital seja qualitativa e incorporada como uma responsabilidade científica compartilhada.

Conclui-se, assim, que a Arquivologia se mostra em permanente movimento. Trata-se de um campo que aprende com as dinâmicas contemporâneas da ciência e, concomitantemente, contribui para ensiná-la sobre a importância da autenticidade, da confiabilidade e da preservação qualitativa. Esta tese reafirma o potencial transformador da Arquivologia ao integrar-se à Ciência Aberta e demonstrar que seus fundamentos clássicos podem ser traduzidos em práticas que unem rigor metodológico, valor informacional e compromisso com a preservação do conhecimento científico.

Ao reafirmar a centralidade da proveniência, da autenticidade e da organicidade, evidencia-se que a Arquivologia permanece não apenas atual, mas necessária para enfrentar os desafios informacionais contemporâneos. A proposição de um modelo de fluxo fundamentado e aplicável confirma que a preservação digital qualitativa é indissociável da gestão inicial dos dados de pesquisa e configura uma responsabilidade compartilhada entre pesquisadores, instituições e sociedade.

REFERÊNCIAS

- AKMON, D. *et al.* The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs. **Archival Science**, Dordrecht, v. 11, p. 329-348, 2011. Disponível em: <https://doi.org/10.1007/s10502-011-9151-4>. Acesso em: 20 out. 2025.
- AMERICAN LIBRARY ASSOCIATION (ALA). Definições de | de Preservação Digital Associação para Coleções de Bibliotecas & Serviços Técnicos (ALCTS). Disponível em: <https://www-ala-org>. Acesso em: 15 set. 2025.
- ALBAGLI, S. Ciência aberta em questão. In: ALBAGLI, S. (org.). **Ciência aberta, questões abertas**. Brasília: IBICT, 2015. p. 9-25.
- ALBAGLI, S. Inovação no capitalismo cognitivo. In: SIQUEIRA, M.; COCCO, G. (orgs.). **Por uma política menor: arte, comum e multidão**. Rio de Janeiro: Fundação Casa de Rui Barbosa, 2014. p. 211-224.
- ALBAGLI, S.; CLINIO, A.; RAYCHTOCK, S. Ciência aberta: correntes interpretativas e tipos de ação. **Liinc em Revista**, v. 10, n. 2, p. 434-450, 2014. Disponível em: <https://revista.ibict.br/liinc/article/view/3593>. Acesso em: 20 out. 2025.
- ALBAGLI, S.; MACIEL, M. L. Informação, poder e política: a partir do Sul, para além do Sul. In: MACIEL, M. M.; ALBAGLI, S. (orgs.). **Informação, conhecimento e poder: mudança tecnológica e inovação social**. Rio de Janeiro: Garamond, 2011. p. 9-39.
- ALBAGLI, S.; ROCHA, L. Ciência urbana em tempos de emergência: iniciativas brasileiras diante da pandemia de COVID-19. **ARBOR Ciencia, Pensamiento y Cultura**, v.197-799, enero-marzo 2021, a589. Disponível em: <https://arbor.revistas.csic.es/index.php/arbor/article/view/2404/3606>. Acesso em: 20 out. 2025.
- ALMEIDA, C. da S. “**Make science great again**”? o impacto da covid-19 na percepção pública da ciência. 2020. Disponível em: <https://arca.fiocruz.br/items/cc4b3525-fdfa-46e4-80fd-580759852576>. Acesso em: 20 out. 2025.
- ALTMAN, K. *et al.* The Role of Understanding, Trust, and Access in Public Engagement with Environmental Activities and Decision Making: A Qualitative Study with Water Quality Practitioners. **Environmental Management**, v. 71, n. 6, p. 1162–1175, 2023. Disponível em: <https://doi.org/10.1007/s00267-023-01803-2>. Acesso em: 20 out. 2025.
- ANDRADE, W. O. de *et al.* **O conceito de informação na Arquivologia contemporânea: da tradução conceitual à delimitação do objeto de estudo na produção científica brasileira**. 2019.
- ARAKAKI, F. A. **Metadados administrativos e a proveniência dos dados: modelo baseado na família PROV**. 2019. Disponível em: [Metadados administrativos e a proveniência dos dados: modelo baseado na família PROV](#). Acesso em: 20 out. 2025.
- ARAKAKI, F. A.; SANTOS, P. L. V. A. C.. Proveniência e contexto digital: contribuições da ciência da informação. **Palavra Chave** (La Plata), La Plata, v. 10, n. 2, 1 abr. 2021. Disponível em: <https://doi.org/https://doi.org/10.24215/18539912e124>. Acesso em: 20 out. 2025.
- ARAÚJO, D. G. *et al.* Contribuições para a gestão de dados científicos: análise comparativa entre modelos de ciclo de vida dos dados. **Liinc em Revista**, v. 15, n. 2, 2019. Disponível em: <https://revista.ibict.br/liinc/article/view/4686>. Acesso em: 10 out. 2025.
- ARQUIVO NACIONAL (Brasil). **Guia de gestão de documentos para os órgãos e entidades do Poder Executivo Federal**. Rio de Janeiro: Arquivo Nacional, 2025. (Publicações Técnicas; 68). Disponível em: <https://www.gov.br/arquivonacional/pt-br/servicos/publicacoes/Guiadegestaodedocumentos.pdf>. Acesso em: 20 out. 2025.

ARQUIVO NACIONAL (Brasil). **Código de classificação e tabela de temporalidade e destinação de documentos relativos às atividades-meio do Poder Executivo Federal [recurso eletrônico]**. Rio de Janeiro: Arquivo Nacional, 2024. Disponível em: <https://bdan.an.gov.br/items/6d9644bd-5911-440b-a7ab-1a42e0f422a8/full> . Acesso em: 4 set. 2025.

ARQUIVO NACIONAL (Brasil). **Dicionário brasileiro de terminologia arquivística**. Rio de Janeiro: O Arquivo, 2005. (Publicações Técnicas-AN, n. 51). Disponível em: https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/dicionrio_de_terminologia_arquivistica.pdf. Acesso em: 20 out. 2025.

ATHERTON, J. From life cycle to continuum: some thoughts on the records management–archives relationship. **Archivaria**, n. 21, p. 43-52, 1985. Disponível em: <https://archivaria.ca/index.php/archivaria/article/view/11233>. Acesso em: 4 set. 2025.

AZEVEDO, R. A. de; INDOLFO, A. C. Classificação de documentos arquivísticos: revisão sistemática de literatura (RSL). **Revista Bibliomar**, p. 1–28, 21 jun. 2024. Disponível em: <https://periodicoseletronicos.ufma.br/index.php/bibliomar/article/view/22685>. Acesso em: 22 fev. 2025.

BALL, A. Making the case for sharing with indicators of research data impact. In: **Library Connect Webinar: How to Assist Researchers in Sharing their Research Data**. 2015. Disponível em: <https://researchportal.bath.ac.uk/en/publications/making-the-case-for-sharing-with-indicators-of-research-data-imp/>. Acesso em: 20 out. 2025.

BALL, A. **Tools for research data management**. 2012. Disponível em: <https://researchportal.bath.ac.uk/en/publications/tools-for-research-data-management/>. Acesso em: 20 out. 2025.

BALL, A. **Review of the state of the art of the digital curation of research data**. 2010. Disponível em: https://www.researchgate.net/publication/242710555_Review_of_the_State_of_the_Art_of_the_Digital_Curation_of_Research_Data. Acesso em: 20 out. 2025.

BARROS, T. H. B. de; MORAES, J. B. E. de. **Da classificação biológica à classificação digital: perspectivas de renovação em classificação arquivística**. 2013. Disponível em: <https://repositorio.unesp.br/entities/publication/b39f4c38-f7ad-4267-abe8-389ef7b61311>. Acesso em: 20 out. 2025.

BATINI, C. *et al.* Methodologies for data quality assessment and improvement. **ACM Computing Surveys**, v. 41, n. 3, p. 1-52, 2009.

BAUCOM, E. A brief history of digital preservation. In: **Digital preservation in libraries: preparing for a sustainable future**. Washington: American Library Association, p. 3-19, 2019.

BAUTIER, R. H. Le Conseil International des Archives *et al.* Commission (Provisoire) du Guide des Sources de l'Histoire d'Amérique Latine: session des 7-9 décembre 1959. **Revista de Historia de América**, p. 207-213, 1960.

BELLOTTO, H. L. **Arquivos permanentes: tratamento documental**. 4. ed. São Paulo: T. A. Queiroz, 2008.

BERLIN DECLARATION OPEN ACCESS TO KNOWLEDGE IN THE SCIENCES AND HUMANITIES. Berlin: Open Access, 2003. Disponível em: https://openaccess.mpg.de/67605/berlin_declaration_engl.pdf. Acesso em: 10 out. 2025.

BESSER, H. Longevidade digital. **Acervo: Revista do Arquivo Nacional**, v. 23, n. 2, jul./dez. 2010. Disponível em: <http://arquivistica.fci.unb.br/au/longevidade-digital/>. Acesso em: 20 out. 2025.

BETHESDA STATEMENT ON OPEN ACCESS PUBLISHING. **Bethesda Statement on**

Open Access Publishing. Bethesda, 2003. Disponível em: <https://legacy.earlham.edu/~peters/fos/bethesda.htm>. Acesso em: 20 out. 2025.

BETTIVIA, R.; CHENG, Y.; GRYK, M. R. **Documenting the future: navigating provenance metadata standards.** Cham: Springer, 2022.

BIZER, C. *et al.* Linked data on the web (LDOW2008). In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 17, 2008, Beijing, China. **Anais...** Beijing, China: ACM, 2008. p. 1265-1266. Disponível em: <https://doi.org/10.1145/1367497.1367760>. Acesso em: 20 out. 2025.

BOECKHOUT, M.; ZIELHUIS, G. A.; BREDENOORD, A. L. The FAIR guiding principles for data stewardship: fair enough? **European Journal of Human Genetics**, v. 26, n. 7, p. 931-936, 2018. Disponível em: <https://www.nature.com/articles/s41431-018-0160-0>. Acesso em: 10 out. 2025.

BORGMAN, C. L. **Big data, little data, no data: Scholarship in the networked world.** MIT press, 2017.

BORGMAN, C. L. **Big science, little science, and open science: sustainability, stewardship, and knowledge infrastructures.** Los Angeles: UCLA Center for Knowledge Infrastructures, 2018. Disponível em: <https://escholarship.org/uc/item/99g200q8>. Acesso em: 10 out. 2025.

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

BRASIL. Arquivo Nacional. Sistema de Gestão de Documentos e Arquivos (SIGA). [S. l.]: Arquivo Nacional, [s. d.]. Disponível em: <https://siga.arquivonacional.gov.br>. Acesso em: 20 out. 2025.

BRASIL. Lei nº 8.159, de 8 de janeiro de 1991. Dispõe sobre a política nacional de arquivos públicos e privados. Diário Oficial da União: Brasília, DF, 9 jan. 1991. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8159.htm. Acesso em: 20 jan. 2025.

BRITO, V. M. de. **A preservação da memória científica da Fiocruz: a visão de quem faz ciência.** 2002. Disponível em: https://minerva.ufjf.br/F/?func=direct&doc_number=000587268&local_base=UFR01. Acesso em: 20 out. 2025.

BROWN, A. **Practical digital preservation: a how-to guide for organizations of any size.** Facet Publishing, 2013. Disponível em: <https://archive.org/details/practicaldigital0000brow>. Acesso em: 20 out. 2025.

BUCKLAND, M. K. Information as thing. **Journal of the American Society for Information Science**, v. 42, n. 5, p. 351-360, 1991. Disponível em: [https://doi.org/10.1002/\(SICI\)1097-4571\(199106\)42:5<351::AID-ASI5>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3). Acesso em: 20 out. 2025.

BULLOCK, A. Preservation of digital information: issues and current status. **Network Notes**, National Library of Canada, Ottawa, n. 60, abr. 1999. Disponível em: <http://epe.lac-bac.gc.ca/100/202/301/netnotes/netnotes-h/notes60.htm>. Acesso em: 10 out. 2025.

CALAZANS, A. T. S. Qualidade da informação: conceitos e aplicações. **Transinformação**, v. 20, p. 29-45, 2008.

CAMARGO, A. M. A.; BELLOTTO, H. L. (coord.). **Dicionário de terminologia arquivística.** São Paulo: Associação dos Arquivistas Brasileiros – Núcleo Regional de São Paulo; Secretaria de Estado da Cultura, 1996.

CAREGNATO, S. E. *et al.* Práticas e percepções dos pesquisadores brasileiros sobre serviços de acesso aberto a dados de pesquisa. **Liinc em Revista**, v. 15, n. 2, p. 121-141, 2019. Disponível em: <https://revista.ibict.br/liinc/article/view/4771>. Acesso em: 4 set. 2025.

CAREGNATO, S. E. *et al.* Reúso de dados: princípios FAIR e o ecossistema de pesquisa. In: SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 187 - 200. DOI: 10.22477/9786589167242. Disponível em: [Reúso de dados: princípios FAIR e o ecossistema de pesquisa | Request PDF](#). Acesso em: 20 out. 2025.

CARPENTER, T. **Welcome to NISO!** [S. l.: s. n.], 2017. Disponível em: <https://www.niso.org/>. Acesso em: 15 set. 2025.

CCSDS. The Consultative Committee for Space Data Systems. **Audit and Certification of Trustworthy Digital Repositories**. Recommended Practice CCSDS 652.0-M-2. Washington: Magenta Book, 2024. Disponível em: https://ccsds.org/wp-content/uploads/gravity_forms/5-448e85c647331d9cbaf66c096458bdd5/2025/01/652x0m2.pdf. Acesso em: 11 set. 2025.

CCSDS. The Consultative Committee for Space Data Systems. **Information Preparation to Enable Long-Term Use**. Recommended Practice CCSDS 653.0-M-1. Washington: Magenta Book, 2024. Disponível em: ccsds.org/Pubs/653x0m1.pdf. Acesso em: 11 set. 2025.

CCSDS. The Consultative Committee for Space Data Systems. **Reference Model for an Open Archival Information System (OAIS)**. Recommended Practice CCSDS 650.0-M-3. Washington: Magenta Book, 2024. Disponível em: https://ccsds.org/wp-content/uploads/gravity_forms/5-448e85c647331d9cbaf66c096458bdd5/2025/01/650x0m3.pdf. Acesso em: 11 set. 2025.

CHAO, T. Mapping methods metadata for research data. **International Journal of Digital Curation**, v. 10, n. 1, p. 82-94, 2015. Disponível em: https://www.researchgate.net/publication/276397508_Mapping_Methods_Metadata_for_Research_Data. Acesso em: 20 out. 2025.

CODATA RDM. CODATA RDM Terminology. Disponível em: <https://codata.org/initiatives/data-science-and-stewardship/rdm-terminology-wg/>. Acesso em: 4 set. 2025.

CONARQ. Conselho Nacional de Arquivos (Brasil). **Diretrizes para a presunção de autenticidade de documentos arquivísticos digitais**. Rio de Janeiro: Arquivo Nacional, 2012.

CONARQ. Conselho Nacional de Arquivos (Brasil). **Diretrizes para implementação de repositórios arquivísticos digitais confiáveis (RDC-Arq)**. 2. versão. Rio de Janeiro: Arquivo Nacional, 2023. (Resolução Conarq nº 51, de 25 ago. 2023). Disponível em: <https://www.gov.br/conarq/pt-br/assuntos/noticias/diretrizes-para-a-implementacao-de-repositorios-arquivisticos-digitais-confiaveis-versao-2>. Acesso em: 4 set. 2025.

CONARQ. Conselho Nacional de Arquivos (Brasil). **e-ARQ Brasil: Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos**. 2. versão. Rio de Janeiro: Arquivo Nacional, 2022. Disponível em: <https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/EARQV205MAI2022.pdf>. Acesso em: 10 out. 2025.

CONARQ. Conselho Nacional de Arquivos (Brasil). **Glossário de Documentos Arquivísticos Digitais**. Câmara Técnica de Documentos Eletrônicos (CTDE). Versão 6.0. Rio de Janeiro, 2020. Disponível em: https://www.gov.br/conarq/pt-br/assuntos/camaras-tecnicas-setoriais-inativas/camara-tecnica-de-documentos-eletronicos-ctde/glosctde_2020_08_07.pdf. Acesso em: 02 out. 2025.

CONARQ. Conselho Nacional de Arquivos (Brasil). **ISAD(G): Norma Geral Internacional de Descrição Arquivística**. 2. ed. Rio de Janeiro: Arquivo Nacional, 2000.

CONWAY, P. **Preservação no universo digital**. 2. ed. Rio de Janeiro: Arquivo Nacional, 2001.

(Conservação Preventiva em Bibliotecas e Arquivos; 52).

COOK, T. Arquivos pessoais e arquivos institucionais: para um entendimento arquivístico comum da formação da memória em um mundo pós-moderno. **Estudos Históricos**, v. 11, n. 21, p. 129-150, 1998.

COOK, T. Fashionable nonsense or professional rebirth: postmodernism and the practice of archives. **Archivaria**, n. 51, p. 14-35, 2001.

COOK, T. What is past is prologue: a history of archival ideas since 1898, and the future paradigm shift. **Archivaria**, n. 43, 1997.

CORRÊA, J. G.; DE SOUSA, J. A. P. Perspectivas arquivísticas na gestão de dados de pesquisa: uma análise a partir da Arquivística Integrada. **Revista Ibero-Americana de Ciência da Informação**, v. 15, n. 2, p. 436-451, 2022.

CORUJO, L.; REVEZ, J. Modelos de requisitos para sistemas de organização do conhecimento. In: **Organização do conhecimento no horizonte 2030: desenvolvimento sustentável e saúde**. Atas do V Congresso ISKO Espanha-Portugal. Lisboa: Colibri, 2021. p. 243-261.

COSTA FILHO, C. M. A.; DE SOUSA, R. T. B. Compreendendo o records continuum: contextualização, objetivos e reflexões. **Informação Arquivística**, v. 6, n. 2, p. 34-60, 2017.

COUTURE, C. Arquivística, os arquivistas e os arquivos no Canadá. **Acervo**, v. 28, n. 2, p. 147-163, 2015. Disponível em: <https://revista.an.gov.br/index.php/revistaacervo/article/view/613>. Acesso em: 25 fev. 2025.

COUTURE, C.; DUCHARME, D.; ROUSSEAU, J. L'Archivistique a-t-elle trouvé son identité? **Argus**, v. 17, n. 2, p. 51-60, 1988.

COX, A. M. *et al.* Developments in research data management in academic libraries: towards an understanding of research data service maturity. **Journal of the Association for Information Science and Technology**, v. 68, n. 9, p. 2182-2200, 2017.

CRUZ MUNDET, J. R. **Archivística: gestión de documentos y administración de archivos**. Madrid: Alianza Editorial, 2012. Disponível em: https://www.academia.edu/35788112/Gesti%C3%B3n_de_documentos_y_Administraci%C3%B3n_de_archivos_textos_fundamentales_Jos%C3%A9_Ram%C3%B3n_Cruz_Mundet. Acesso em: 4 set. 2025.

CTS. CoreTrustSeal Standards and Certification Board. **Core Trust Seal Requirements 2023–2025 (V01.00)**. Zenodo, 2022. Disponível em: <https://www.coretrustseal.org/about/standards-and-certification-board/>. Acesso em: 4 set. 2025.

CUNHA, M. B. da; CAVALCANTI, C. R. de O. **Dicionário de Biblioteconomia e Arquivologia**. Brasília: Briquet de Lemos, 2008. xvi, 451 p.

CURTY, R. Abordagens de reuso e a questão da reusabilidade dos dados científicos. **Liinc em Revista**, v. 15, n. 2, p. 177-193, 2019. Disponível em: <https://doi.org/10.18617/liinc.v15i2.4777>. Acesso em: 4 set. 2025.

DA CUNHA MARQUES, A. A. Arquivologia e Ciência da Informação: de mãos dadas? **Informação & Sociedade**, v. 26, n. 3, 2016. Disponível em: [pdf-22.pdf](#). Acesso em: 4 set. 2025.

DA CUNHA MARQUES, A. A. A investigação científica em Arquivologia e a sua busca de identidade. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, 2017.

DA SILVA, F. I.; ARAKAKI, F. A.. Fampilia PROV-O e repositório de dados no contexto do Re3data. **Ciência da Informação Express**, v. 6, p. 1-20, 2025. Disponível em: [Família PROV-O e repositório de dados no contexto do Re3data | Ciência da Informação Express](#). Acesso em: 20 out. 2025.

DAMA INTERNATIONAL. **DAMA-DMBOK: The DAMA Guide to the Data Management**

- Body of Knowledge. 2. ed. Bradley Beach, New Jersey: Technics Publications, 2017. Disponível em: [DAMA-DMBOK | Guide books](#). Acesso em: 10 out. 2025.
- DAPPER, A.; GUENTHER, R. S.; PEYRARD, S. An introduction to the PREMIS Data Dictionary for Digital Preservation Metadata. In: **Digital Preservation Metadata for Practitioners**. Springer International Publishing, 2018.
- DARDOT, P.; LAVAL, C. Propriedade, apropriação social e instituição do comum. **Tempo Social**, v. 27, n. 1, p. 261-276, jun. 2015.
- DATA CITE. **DataCite**. Disponível em: <https://datacite.org/>. Acesso em: 20 out. 2025.
- DATAVERSE PROJECT. **Guides**. Disponível em: <https://dataverse.org>. Acesso em: 20 out. 2025.
- DAVANZO, L. **Construção e aplicação de tesauros funcionais na arquivologia: uma proposta teórica e metodológica a partir do modelo records continuum**. 2021. Disponível em: [Construção e aplicação de tesauros funcionais na arquivologia: uma proposta teórica e metodológica a partir do modelo records continuum](#). Acesso em: 20 out. 2025.
- DAVID, J. C. D. A. **Perspectivas contemporâneas da norma ISO 15489 no ensino de gestão de documentos na UFF e na UNIRIO**. Dissertação (Arquivologia) – Universidade Federal Fluminense, Instituto de Arte e Comunicação Social, 2018.
- DCC. Digital Curation Centre. **Diffuse Standards Frameworks**. Edinburgh, 2013. Disponível em: <http://www.dcc.ac.uk/resources/standards/diffuse/>. Acesso em: 10 out. 2025.
- DCC. Digital Curation Centre. **The digital curation lifecycle model**. Edinburgh: DCC, [2008]. Disponível em: <https://www.dcc.ac.uk/guidance/curation-lifecycle-model>. Acesso em: 20 out. 2025.
- DCC. Digital Curation Centre. **What is Digital Curation?** Briefing Paper. 2008. Disponível em: <https://www.dcc.ac.uk/sites/default/files/documents/resource/briefing-papers/what-is-digital-curation.pdf>. Acesso em: 20 out. 2025.
- DE ARAÚJO, D. G. *et al.* Contribuições para a gestão de dados científicos: análise comparativa entre modelos de ciclo de vida dos dados. **Liinc em Revista**, v. 15, n. 2, 2019.
- DE LUSIGNAN, S. *et al.* Key concepts to assess the readiness of data for International research: data quality, lineage and provenance, extraction and processing errors, traceability, and curation. **Yearbook of Medical Informatics**, v. 20, n. 1, p. 112-120, 2011.
- DE MOURA SOUSA, A. P. *et al.* Princípios da descrição arquivística: do suporte convencional ao eletrônico. **Arquivística.net**, v. 2, n. 2, p. 38-51, 2006. Disponível em: <http://arquivistica.fci.unb.br/wp-content/uploads/tainacan-items/476350/838861/56134-1.pdf>. Acesso em: 20 out. 2025.
- DE OLIVEIRA, A. C. S.; DA SILVA, E. M. Ciência aberta: dimensões para um novo fazer científico. **Informação & Informação**, v. 21, n. 2, p. 5-39, 2016.
- DE OLIVEIRA VEIGA, V. S. *et al.* Plano de gestão de dados FAIR: uma proposta para a Fiocruz. **Liinc em Revista**, v. 15, n. 2, 2019. Disponível em: <https://liinc.commscientia.com.br/index.php/liinc/article/view/5030>. Acesso em: 20 out. 2025.
- BUDAPEST OPEN ACCESS INITIATIVE (BOAI). **Budapest Open Access Initiative**. 2002. Disponível em: <https://www.budapestopenaccessinitiative.org/>. Acesso em: 20 jun. 2025.
- DECLARATION ON SCIENCE AND THE USE OF SCIENTIFIC KNOWLEDGE. [s. l.], 1999. Disponível em: <http://www.un-documents.net/sci-dec.htm>. Acesso em: 10 out. 2025.
- DELEUZE, G. Post-Scriptum sobre as sociedades do controle. In: DELEUZE, G. **Conversações: 1972-1990**. Rio de Janeiro: Editora 34, 1992.

DO NASCIMENTO, F. J. T. Segurança da informação. Aula 5. *In: Curso de Ciência Aberta* (FIOCRUZ). [S. l.], 2022. Material didático/slide.

DO NASCIMENTO, F. J. T.; DE CASTRO LOUREIRO, É.; DE ARRUDA JORGE, V. Sigilo e acesso: entendimentos e procedimentos para a adequada disponibilização de informações e dados governamentais abertos. *RECHIS*, v. 11, 2017.

DODEBEI, V. Memória do conhecimento: em busca de sustentabilidade para os objetos digitais. *Ciência da Informação*, v. 43, n. 1, 2014. Disponível em: <https://doi.org/10.18225/ci.inf.v43i1.1424>. Acesso em: 20 out. 2025.

DOLLAR, C. Archivists and Records Managers in the Information Age. *Archivaria*, n. 36, p. 37–52, outono 1993. Disponível em: [Archivists and Records Managers in the Information Age | Archivaria](#). Acesso em: 20 out. 2025.

DONALDSON, D. R.; ZEGLER-POLESKA, E.; YARMEY, L. Data managers' perspectives on OAIS designated communities and the FAIR principles: mediation, tools and conceptual models. *Journal of Documentation*, v. 76, n. 6, p. 1261-1277, 2020. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/JD-10-2019-0204/full/html>. Acesso em: 10 out. 2025.

DOORN, P. K. *Science Europe Practical Guide to the International Alignment of Research Data Management*. 2018.

DIGITAL COALITION PRESREvation (DPC). Disponível em <https://www.dpconline.org/handbook>. Acesso em: 15 set. 2025.

DURANTI, L. Concepts, principles, and methods for the management of electronic records. *The Information Society*, v. 17, n. 4, p. 271-279, 2001.

DURANTI, L. The archival bond. *Archives and Museum Informatics*, v. 11, p. 213–218, 1997.

DURANTI, L. *The long-term preservation of authentic electronic records: findings of the InterPARES Project*. 2005. Disponível em: <http://www.interpares.org/book/index.cfm>. Acesso em: 20 out. 2005.

ESANU, J. *et al.* Selection, appraisal, and retention of digital scientific data: highlights of an ERPANET/CODATA workshop. *Data Science Journal*, v. 3, p. 227-232, 2004.

ESTEVÃO, J. S. B.; ARNS, E. M.; STRAUHS, F. do R. Gestão de dados de pesquisa: uma prática para abrir a caixa preta da pesquisa científica. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, v. 17, p. e019031, 2019. Disponível em: [\(PDF\) Gestão de dados de pesquisa: uma prática para abrir a caixa preta da pesquisa científica](#). Acesso em: 20 out. 2025.

FACILITATE OPEN SCIENCE TRAINING FOR EUROPEAN (FOSTER). 2016. Disponível em: <https://www.fosteropenscience.eu/>. Acesso em: 10 jul. 2025.

FECHER, B.; FRIESIKE, S. Open science: one term, five schools of thought. *In: Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*. Cham: Springer International Publishing, 2013. p. 17-47.

FEDERER, L. The librarian as research informationist: a case study. *Journal of the Medical Library Association (JMLA)*, v. 101, n. 4, p. 298, 2013.

FELICIATI, P.; DURANTI, L. The responsible use of artificial intelligence in archives through the use of paradata. *JLIS.it*, v. 16, n. 2, p. 1-9, 2025. Disponível em: [The responsible use of Artificial Intelligence in archives through the use of paradata | JLIS.it](#). Acesso em: 4 set. 2025.

FERNANDES, D. F. L. *Digital preservation in e-science scenarios – an enterprise architecture approach*. 2012.

FERNANDES, H. D. H.; DE OLIVEIRA CORREIO, A. F. Gestão da preservação digital em repositórios de dados de pesquisa. **Revista Ibero-Americana de Ciência da Informação**, v. 11, n. 1, p. 255-273, 2018.

FIOCRUZ. Fundação Oswaldo Cruz. Presidência. Vice-presidência de Ensino, Informação e Comunicação. **Política de gestão, compartilhamento e abertura de dados para pesquisa: princípios e diretrizes**. Rio de Janeiro, 2020. 19 p. Disponível em: [Política de gestão, compartilhamento e abertura de dados para pesquisa: princípios e diretrizes](#). Acesso em: 20 out. 2025.

FIOCRUZ. Fundação Oswaldo Cruz. **O que é Ciência Aberta?** Formação modular. Curso Online. Disponível em: <https://campusvirtual.fiocruz.br/portal/?q=formacao-modular/ciencia-aberta>. Acesso em: 14 set. 2025.

FIOCRUZ. Fundação Oswaldo Cruz. Portal Fiocruz. Disponível em: <https://www.fiocruz.br>. Acesso em: 4 set. 2025.

FIOCRUZ. Fundação Oswaldo Cruz. Sistema de Gestão de Documentos e Arquivos da Fiocruz – SIGDA. **Plano de Gestão de Documentos Arquivísticos Digitais da Fiocruz**. 2024. Disponível em: <https://sigda.fiocruz.br>. Acesso em: 4 set. 2025.

FLORES, D.; PRADEBON, D. S.; CÉ, G. Análise do conhecimento teórico-metodológico da preservação digital sob a ótica da OAIS, SAAI, ISO 14721 e NBR 15472. **Brazilian Journal of Information Science**, v. 11, n. 4, p. 73-81, 2017. Disponível em: dialnet.unirioja.es/servlet/articulo?codigo=6234786. Acesso em: 20 out. 2025.

FOGLIENI, O. When preservation & conservation meet open science. **International Federation of Library Associations and Institutions**, Dublin, p. 1-9, 2022. Disponível em: <http://repository.ifla.org/bitstream/123456789/2077/1/s5-2022-%20foglieni-en.pdf>. Acesso em: 10 out. 2025.

FONSECA, J. J. S. **Metodologia da pesquisa científica**. Universidade Estadual do Ceará, 2002. Disponível em: [Apostila de metodologia da pesquisa científica - Google Livros](#). Acesso em: 20 out. 2025.

FONSECA, M. L. da. **A contribuição da Arquivologia para a gestão de dados de pesquisa**. Trabalho de Conclusão de Curso – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018. 68 f. Disponível em: <https://www.lume.ufrgs.br/handle/10183/189759?show=full>. Acesso em: 26 jan. 2025.

FONSECA, M. O. **Arquivologia e ciência da informação**. Rio de Janeiro: FGV, 2005.

FONSECA, M. **Arquivologia e Ciência da Informação: (re)definição de marcos interdisciplinares**. 2004. 181f. Tese (Doutorado em Ciência da Informação) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2004.

FORCE 11. **The Future of Research Communications and e-Scholarship**. Disponível em: [FORCE11 – The Future of Research Communications and e-Scholarship](#). Acesso em: 20 out. 2025.

FORMENTON, D.; GRACIOSO, L. S. Digital preservation challenges, requirements, strategies and scientific output. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, v. 18, p. e020012, 2023. Disponível em: [\(PDF\) Digital Preservation: challenges, requirements, strategies and scientific output](#). Acesso em: 20 out. 2025.

FOUCAULT, M. **A verdade e as formas jurídicas**. 3. ed. Rio de Janeiro: Nau Editora, 2002. p. 7-27.

FOUCAULT, M. Aula de 17 de março de 1976. In: FOUCAULT, M. **Em defesa da sociedade** (Curso no Collège de France – 1975-1976). São Paulo: Martins Fontes, 2005. p. 285-315.

FOUCAULT, M. O sujeito e o poder. In: FOUCAULT, M. **Genealogia da Ética, Subjetividade**

- e Sexualidade. **Ditos e Escritos IX**. Rio de Janeiro: Forense Universitária, 2014. p. 118-140.
- GAGNON-ARGUIN, L. Réflexions sur les revues professionnelles: le cas de l'archivistique. **Documentation et Bibliothèques**, v. 38, n. 4, p. 191-196, 1992.
- GAVA, T. B. S. *et al.* Dados de pesquisa na Arquivologia: uma reflexão. **Em Questão**, v. 30, p. e135857, 2024.
- GAVA, T. B. S.; FLORES, D. Preservação digital sistêmica. In: **Arquivo, Documento e Informação em Cenários Híbridos**: anais do Simpósio Internacional de Arquivos. São Paulo: Eventus, 2021.
- GEZELTER, D. **What, exactly, is open science?** [s. l.: s. n.], 2009. Disponível em: <http://www.openscience.org/blog/?p=269>. Acesso em: 30 ago. 2025.
- GIARETTA, D. *et al.* **LABDRIVE, a Petabyte scalable, OAIS/ISO 16363 conformant, for scientific research organisations to preserve documents, processed data, and software.** 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japão, 2022, p. 2515-2521. Disponível em: <https://doi.org/10.1109/BigData55660.2022.10020648>. Acesso em: 20 out. 2025.
- GIL, A. C. Métodos e técnicas de pesquisa social. 4. ed. São Paulo: **Atlas**, 1994. 207 p. Disponível em: https://www.academia.edu/28982188/M%C3%A9todos_e_T%C3%A9cnicas_de_Pesquisa_Social_Antonio_Carlos_Gil_6_ed_2008. Acesso em: 10 maio 2025.
- GLOBAL INDIGENOUS DATA ALLIANCE. **CARE Principles**. [S. l.: s. n.], [s. d.]. Disponível em: <https://www.gida-global.org/care>. Acesso em: 20 out. 2025.
- GLOBAL OPEN FINDABLE, ACCESSIBLE, INTEROPERABLE, REUSABLE (GOFAIR Initiative). Página institucional. Disponível em: <https://www.go-fair.org/>. Acesso em: 10 jan. 2025.
- GONÇALVES, J. **Como classificar e ordenar documentos de arquivo**. São Paulo: Arquivo do Estado, 1998. 37 p. : il. ; 23 cm. (Projeto como fazer; v. 2). Disponível em: https://www.arqsp.org.br/arquivos/oficinas_colecao_como_fazer/cf2.pdf. Acesso em: 20 jan. 2025.
- GRACY, K. F.; KAHN, M. B. Preservation in the digital age. **Library Resources & Technical Services**, v. 56, n. 1, p. 25-43, 2012. Disponível em: <https://journals.ala.org/index.php/lrts/article/view/5495>. Acesso em: 10 out. 2025.
- GUTMANN, M. *et al.* The selection, appraisal, and retention of social science data. **Data Science Journal**, v. 3, p. 209-221, dez. 2004. Disponível em: <https://datascience.codata.org/articles/308/files/submission/proof/308-1-599-1-10%2020150417.pdf>. Acesso em: 4 set. 2025.
- HAYNES, D. Metadata for information management and retrieval: understanding metadata and its use. London: **Facet Publishing**, 2018.
- HECKER, S. *et al.* Innovation in open science, society and policy – setting the agenda for citizen science. In: HECKER, S.; HAKLAY, M.; BOWSER, A.; MAKUCH, Z.; VOGEL, J.; BONN, A. (Eds.). **Citizen Science: Innovation in Open Science**. London: **UCL Press**, 2018. p. 1-23.
- HEDSTROM, M. Digital preservation: a time bomb for digital libraries. **Computers and the Humanities**, v. 31, n. 3, p. 189-202, 1997.
- HENNING, P. C. *et al.* GOFAIR 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, 2019. DOI: 10.19132/1808-5245252.389-412. Disponível em: <https://seer.ufrgs.br/index.php/EmQuestao/article/view/84753>. Acesso em: 15 dez. 2025.

HEREDIA HERRERA, A. **Archivística general: teoría y práctica**. 5. ed. Sevilla: Diputación Provincial de Sevilla, 1991.

HESS, C.; OSTROM, E. Introduction: an overview of the knowledge commons. In: HESS, C.; OSTROM, E. (Ed.). *Understanding Knowledge as a Commons: From Theory to Practice*. Cambridge, MA: MIT Press, 2007. p. 3-26.

HEY, T.; TANSLEY, S.; TOLLE, K. M. **Jim Gray on eScience: a transformed scientific method**. 2009. Disponível em: <https://languageolog.ldc.upenn.edu/myl/JimGrayOnE-Science.pdf>. Acesso em: 20 jan. 2025.

HIGGINS, S. The DCC curation lifecycle model. *International Journal of Digital Curation*, Edinburgh, v. 3, n. 1, p. 134–140, 2008. Disponível em: https://www.researchgate.net/publication/220924444_The_DCC_curation_lifecycle_model. Acesso em: 20 out. 2025.

HIGMAN, R.; BANGERT, D.; JONES, S. **Three camps, one destination: the intersections of research data management, FAIR and Open**. 2019. Disponível em: [Three camps, one destination: the intersections of research data management, FAIR and Open | GRO.publications](https://www.gro-publications.com/publication/Three-camps-one-destination-the-intersections-of-research-data-management-FAIR-and-Open). Acesso em: 20 out. 2025.

HODSON, S. *et al.* **Turning FAIR into reality: final report and action plan from the EC expert group on FAIR data**. Apresentação (slides). [S. l.: s. n.], 2018. Disponível em: [PowerPoint Presentation](#). Acesso em: 20 out. 2025.

INTERNATIONAL COUNCIL ON ARCHIVES (ICA). **ICA Multilingual Terminology**. Disponível em: [Multilingual Archival Terminology - MAT - ICA](#). Acesso em: 20 out. 2025.

INDOLFO, A. C. *Dimensões político-arquivísticas da avaliação de documentos na administração pública federal (2004-2012)*. 2013. Tese (Doutorado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Instituto Brasileiro de Informação em Ciência e Tecnologia; Universidade Federal do Rio de Janeiro, Escola de Comunicação, Rio de Janeiro, 2013. 312 f.

INNARELLI, H. C. Introdução aos dez mandamentos da preservação digital. **Sínteses**: Revista Eletrônica do SimTec, Campinas, SP, n. 2, p. 178–178, 2016. Disponível em: <https://econtents.bc.unicamp.br/inpec/index.php/simtec/article/view/8483>. Acesso em: 24 ago. 2025.

INNARELLI, H. C. *Preservação de documentos digitais*. São Paulo: Associação de Arquivistas de São Paulo – ARQ-SP, 2012. 64 p. (**Instrumenta**, 2).

INTERPARES PROJECT. **InterPARES Glossary Online**. [S. l.: s. n.], [s. d.]. Disponível em: http://www.interpares.org/ip2/ip2_term_pdf.cfm?pdf=glossary. Acesso em: 20 out. 2025.

INTERPARES PROJECT. Site do Projeto InterPARES. [S. l.: s. n.], [s. d.]. Disponível em: <https://interpares.org/>. Acesso em: 20 out. 2025.

INTERPARES TRUST. [S. l.: s. n.], [s. d.]. Disponível em: <https://interparestrust.org/>. Acesso em: 26 jan. 2025.

INTERPARES 2 PROJECT. **Diretrizes do preservador**: a preservação de documentos arquivísticos digitais: diretrizes para organizações. Disponível em: [Diretrizes do preservador - A PRESERVAÇÃO DE DOCUMENTOS ARQUIVÍSTICOS DIGITAIS: DIRETRIZES PARA ORGANIZAÇÕES](#). Acesso em: 20 out. 2025.

INTERPARES 2 PROJECT. **Diretrizes do produtor**: a elaboração e a manutenção de materiais digitais: diretrizes para indivíduos. Disponível em: https://www.interpares.org/display_file/ip2_creator_guidelines_booklet--portuguese.pdf. Acesso em: 15 set. 2025.

ISO. International Organization for Standardization (ISO). ISO 15489-1. **Information and**

documentation — Records management — Part 1: Concepts and principles. 2016. Disponível em: [ISO 15489-1:2016 - Information and documentation — Records management — Part 1: Concepts and principles](#). Acesso em: 20 out. 2025.

ISO. International Organization for Standardization (ISO). ISO 16363. **Space data and information transfer systems — Audit and certification of trustworthy digital repositories.** Disponível em: [ISO 16363:2025 - Space data and information transfer systems — Audit and certification of trustworthy digital repositories](#). Acesso em: 20 out. 2025.

JARDIM, J. M. Caminhos e perspectivas da gestão de documentos em cenários de transformações. **Acervo**, v. 28, n. 2, p. 19–50, 2015. Disponível em: <https://revista.an.gov.br/index.php/revistaacervo/article/view/607>. Acesso em: 4 out. 2025.

JARDIM, J. M. O conceito e a prática de gestão de documentos. **Acervo**, v. 2, n. 2, p. 35-42, 1987.

JISC. **Research Data Management Toolkit Guide.** Disponível em: <https://www.jisc.ac.uk/guides/research-data-management-toolkit>. Acesso em 05 set. 2025.

JISC. **Research lifecycle.** [S. l.: s. n.], [s. d.]. Disponível em: <https://webarchive.nationalarchives.gov.uk/>. Acesso em: 05 out. 2025.

JISC. **Technical Specification Requirements Checklist.** Disponível em: <https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Frepository.jisc.ac.uk%2F8435%2F1%2FResearch-data-management-checklist.docx&wdOrigin=BROWSELINK>. Acesso em: 06 out. 2025.

JOHNSTON, L. R. Curating Research Data: Practical Strategies for Your Digital Repository. Chicago: **Association of College and Research Libraries**, 2017. Disponível em: https://conservancy.umn.edu/bitstream/handle/11299/185334/Intro_CuratingResearchData_v1.pdf?sequence=1&isAllowed=y. Acesso em: 26 jan. 2025.

JORGE, Vanessa de Arruda. **Abertura e compartilhamento de dados para pesquisa nas situações de emergência em saúde pública: o caso do vírus zika.** 2018. 263 f. Tese (Doutorado em Ciência da Informação) - Universidade Federal do Rio de Janeiro; Instituto Brasileiro de Informação em Ciência e Tecnologia, Rio de Janeiro, 2018. Disponível em: <https://arca.fiocruz.br/handle/icict/32604>. Acesso em: 10 out. 2025.

KENNEY, A.; MCGOVERN, N. **The three-legged stool: institutional response to digital preservation.** Cornell University Library, 2003. Disponível em: <http://www.nancemcgobern.org/digitalpreservation/DPMcheatsheet-DPMThree-leggedStool.pdf>. Acesso em: 17 out. 2025.

KOHWALTER, S. **Proveniência de Dados.** CEFET. Rio de Janeiro. 2021. Apresentação em power point. Disponível em: http://profs.ic.uff.br/~troy/presentations/2021_CEFET_Prov_Dados.pdf. Acesso em: 23 out. 2025.

LAFLAMME, M.; POETZ, M.; SPICHTINGER, D. Seeing oneself as a data reuser: how subjectification activates the drivers of data reuse in science. **PLoS One**, v. 17, n. 8, p. e0272153, 2022.

LAKATOS, E. M.; MARCONI, M. A. Fundamentos de metodologia científica. 5. ed. São Paulo: **Atlas**, 2003.

LAVOIE, B. The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition). **DPC Technology Watch Report 14-02.** York, UK: Digital Preservation Coalition. Disponível em: <https://dx.doi.org/10.7207/twr14-02>. Acesso em: 20 out. 2025.

LAZZARATO, M. Sujeição e servidão no capitalismo contemporâneo. **Cadernos de Subjetividade**, PUC/SP, p. 168-179, 2010. Disponível em:

<https://doi.org/10.2354/cs.v0i12.38458>. Acesso em: 10 out. 2025.

LEE, Y. W.; STRONG, D. M. Knowing-why about data processes and data quality. **Journal of Management Information Systems**, v. 20, n. 3, p. 13-39, 2003. Disponível em: <https://web.mit.edu/tdqm/www/tdqmpub/Knowing-why.pdf>. Acesso em 10 out. 2025.

LI, C.; SUGIMOTO, S. Provenance description of metadata using PROV with PREMIS for long-term use of metadata. In: Proceedings of the International Conference on Dublin Core and Metadata Applications. **Dublin Core Metadata Initiative**, 2014. Disponível em: [Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata](#). Acesso em: 10 out. 2025.

LIN, D. *et al.* The TRUST principles for digital repositories. **Scientific Data**, v. 7, p. 144, 2020. Disponível em: <https://doi.org/10.1038/s41597-020-0486-7>. Acesso em: 20 out. 2025.

LINDLAR, M. *et al.* “You say potato, I say potato”: mapping digital preservation and research data management concepts towards collective curation and preservation strategies. **International Journal of Digital Curation**, v. 15, n. 1, p. 1-26, 2020. Disponível em: <https://oa.tib.eu/renate/handle/123456789/6101>. Acesso em: 10 out. 2025.

LIU, S. Embedding library services in research stages: Chinese subject service and the research lifecycle model. **International Journal of Librarianship**, v. 2, n. 1, p. 16-31, 2017. Disponível em: <https://doi.org/10.23974/ijol.2017.vol2.1.2>. Acesso em: 20 out. 2025.

LONGO, D. L.; DRAZEN, J. M. Data sharing. **New England Journal of Medicine**, v. 374, n. 3, p. 276-277, 2016. Disponível em: [Data Sharing | New England Journal of Medicine](#). Acesso em: 20 out. 2025.

LOPES, L. C. Vers une archivistique internationale à l'ère de l'information. **Archives, Quebec**, v. 29, n. 2, 1997-1998.

MACHADO, J. G. N. **Documentos arquivísticos audiovisuais da Fiocruz: adaptações de padrões de metadados para atender a requisitos de preservação digital**. 2022. Dissertação (Mestrado Profissional em Preservação e Gestão do Patrimônio Cultural das Ciências e da Saúde) - Casa de Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, 2022. Disponível em: <https://arca.fiocruz.br/handle/iciict/61751>. Acesso em: 10 out. 2025.

MACHADO, J. G. N. **Planejamento para estruturação de Pacotes de Informações OAIS e o seu gerenciamento através do software de preservação para repositórios digitais**. Apresentação de aula (slides). Slide 71. Fundação Casa de Rui Barbosa, 2023. Enviado por e-mail à autora.

MACIEL, M. L.; ABDO, A. H.; ALBAGLI, S. **Open Science, open issues**. 2015. Disponível em: [Open Science open issues_Digital.pdf](#). Acesso em: 11 out. 2025.

MARCONI, M. de A.; LAKATOS, E. M. Fundamentos de metodologia científica. 5. ed. São Paulo: **Atlas**, 2003.

MÁRDERO ARELLANO, M. Á. **Critérios para a preservação digital da informação científica**. 2008. Tese (Doutorado) – Universidade de Brasília, Departamento de Ciência da Informação, 2008. Disponível em: <https://core.ac.uk/download/pdf/11884842.pdf>. Acesso em: 11 out. 2025

MÁRDERO ARELLANO, M. Á. Preservação de documentos digitais. **Ciência da Informação**, v. 33, n. 2, p. 15-27, maio/ago. 2004. Disponível em: <http://www.scielo.br/pdf/ci/v33n2/a02v33n2.pdf>. Acesso em: 10 out. 2025.

MARÍN-ARRAIZA, P.; VIDOTTI, S. Implementação de serviços institucionais de administração de dados. **Liinc em Revista**, v. 15, n. 2, 2019.

MARQUES, A. A. **A arquivologia brasileira: busca por autonomia científica no campo da informação e interlocuções internacionais**. Associação dos Arquivistas Brasileiros, 2013.

MATLALA, M. E.; MAPHOTO, A. R. Application of the records life-cycle and records continuum models in organizations in the 21st century. **ESARBICA Journal**, v. 39, p. 79-98, 2020.

MCHUGH, A. **DCC Digital Curation Manual**. 2005. Disponível em: open-source.pdf. Acesso em: 10 out. 2025.

MCKEMMISH, S. Traces: document, record, archive, archives. In: **Archives: Recordkeeping in Society**. v. 24, p. 1-20, 2005.

MEDEIROS, C. B. *et al.* **Promoção e Implementação da Ciência Aberta no Brasil**: resposta ao texto “Ciência Aberta: uma visão desapaixonada”. 2025. Zenodo. Disponível em: <https://doi.org/10.5281/zenodo.14754044>. Acesso em: 20 out. 2025.

MELO, Í. F. **As funções arquivísticas à luz do princípio da proveniência: um habitus em construção**. 2021. Disponível em: <http://repositorio.unb.br/handle/10482/41694>. Acesso em: 15 out. 2025.

MENEZES, D.; GALVÃO, R.. **Ciência Aberta: uma visão desapaixonada**. Brasília, DF: Conselho Nacional de Desenvolvimento Científico e Tecnológico, 2025. Disponível em: [Ciência Aberta: uma visão desapaixonada — Conselho Nacional de Desenvolvimento Científico e Tecnológico](https://cienciaaberta.org.br/conselho-nacional-de-desenvolvimento-cientifico-e-tecnologico). Acesso em: 02 ago. 2025.

METS. **Metadata Encoding & Transmission Standard (METS)**: METS Version 2 – an overview & tutorial. [S. l.: s. n.], [s. d.]. Disponível em: <https://www.loc.gov/standards/mets/METSOOverview.v2.html>. Acesso em: 20 out. 2025.

MICHAELIS. **Dicionário online em português**. Disponível em: <https://michaelis.uol.com.br>. Acesso em: 20 out. 2025

MILLAR, L. A. A morte dos fundos e a ressurreição da proveniência: o contexto arquivístico no espaço e no tempo. **Informação Arquivística**, v. 4, n. 1, p. 144-162, 2015.

MINAYO, M. C. de S. (org.). **Pesquisa social: teoria, método e criatividade**. 18. ed. Petrópolis: Vozes, 2001. Disponível em: <https://pesquisa.bvsalud.org/portal/resource/pt/cla-1168>. Acesso em: 10 maio 2025.

MINAYO, M. C. de S.; SANCHES, O. Quantitativo-qualitativo: oposição ou complementaridade? **Cadernos de Saúde Pública**, v. 9, p. 237-248, 1993.

MONS, B. **Data stewardship for open science: implementing FAIR principles**. Chapman and Hall/CRC, 2018.

MOREIRA, J. L. R. *et al.* Towards findable, accessible, interoperable and reusable (FAIR) data repositories: improving a data repository to behave as a FAIR data point. **Liinc em Revista**, v. 15, n. 2, 2019. Disponível em: [4318-libre.pdf](https://liinc.org.br/revista/4318-libre.pdf). Acesso em: 20 out. 2025.

MOTA, M. B. da. Apresentação. In: FOUCAULT, M. **Estratégia, poder-saber (Ditos e Escritos IV)**. 2. ed. Rio de Janeiro: Forense Universitária, 2006. p. V-XXXIII; LVIII-LXVI.

NASA. **Meteorite Landings** – About data. Disponível em: https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh/about_data. Acesso em: 15 set. 2025.

NASA. **Meteorite Landings** – Visualization. Disponível em: <https://data.nasa.gov/d/gh4g-9sfh/visualization>. Acesso em: 15 set. 2025.

NASCIMENTO, L. **A preservação da organicidade da informação arquivística**. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal Fluminense, Programa de Pós-Graduação em Ciência da Informação, Niterói, RJ, 2012. Orientador: Sérgio Conde de Albite Silva.

NAUMANN, F.; ROLKER, C. **Assessment methods for information quality criteria**. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für

- Informatik, 2005. Disponível em: <G:\PAPERS\IQ2000\IQ2000NaRo.pdf.prn.pdf>. Acesso em: 10 out. 2025.
- NESMITH, T. Archives from the bottom up: social history and archival scholarship. *Archivaria*, p. 5-26, 1982. Disponível em: [View of Archives from the Bottom Up: Social History and Archival Scholarship](#). Acesso em: 23 out. de 2025.
- OLETO, R. R. Percepção da qualidade da informação. *Ciência da Informação*, v. 35, p. 57-62, 2006.
- OLIVEIRA, A. C. S.; SILVA, E. M. da. Ciência aberta: dimensões para um novo fazer científico. *Informação & Informação*, Londrina, v. 21, n. 2, p. 5–39, 2016. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/27666>. Acesso em: 10 out. 2025.
- OLIVEIRA, L. M. V. de. Política de aquisição: uma reflexão em torno das questões que orientam o processo de ampliação dos acervos institucionais. In: Políticas de Aquisição de Preservação de Acervos em Universidades e Instituições de Pesquisa. Rio de Janeiro: **Museu de Astronomia e Ciências Afins**, 2012. p. 117-130.
- OPEN KNOWLEDGE FOUNDATION. 2004. Disponível em: <https://okfn.org/en/>. Acesso em: 10 out. 2025.
- ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). **Making Open Science a Reality**. (OECD Science, Technology and Industry Policy Papers, n. 25). 2015. Disponível em: [Making Open Science a Reality | OECD](#). Acesso em: 11 out. 2025.
- ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). **OECD Principles and Guidelines for Access to Research Data from Public Funding**. 2007. Disponível em: <http://www.oecd.org/science/inno/38500813.pdf>. Acesso em: 25 jan. 2025.
- PAES, M. L. **Arquivo: teoria e prática**. 3. ed. Rio de Janeiro: FGV Editora, 2004.
- PAMPEL, H.; DALLMEIER-TIESSEN, S. Open research data: from vision to practice. In: BARTLING, S.; FRIESIKE, S. (eds.). *Opening Science*. Cham: **Springer**, 2014. p. 213–224. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-00026-8_14. Acesso em: 20 out. 2025.
- PASQUETTO, I. V.; RANDLES, B. M.; BORGMAN, C. L. **On the reuse of scientific data**. 2017. Disponível em: [On the Reuse of Scientific Data](#). Acesso em: 11 out. 2025.
- PAZIN, M. C. de C. Conceitos fundamentais de classificação e ordenação de documentos de arquivo. In: CARVALHO, A. C. D. (org.). *Memória da Saúde: desafios e possibilidades do trabalho em arquivos e museus de ciência*. Ribeirão Preto: **FUNPEC**, 2006.
- PERINI, E. O que move as fake news e o negacionismo científico? [Entrevista cedida a] Marco Weissheimer. *Sul* 21, 27 nov. 2019. Disponível em: <https://outraspalavras.net/outrasmidias/o-que-move-as-fake-news-e-negacionismocientifico/>. Acesso em: 10 set. 2025.
- PESET MANCEBO, M. F.; GONZÁLEZ-MORENO, L. M. Electronic laboratory notebooks (ELN), gestores do ciclo de vida de objetos de investigação. *Anuario ThinkEPI*, v. 10, p. 269-272, 2016. Disponível em: <https://core.ac.uk/display/275624456?source=2>. Acesso em: 10 out. 2025.
- PHILIPSON, J.; HASAN, A.; MOA, H.. **Making data management plans machine actionable: templates and tools**. 2023. Disponível em: [Making Data Management Plans Machine Actionable: Templates and Tools | Data Science Journal](#). Acesso em 10 out. 2025.
- PICCOLO, D. M. *et al.* Qualidade de dados em gestão de dados de pesquisa: um estudo bibliométrico. *Em Questão*, v. 28, n. 1, p. 159-184, 2022. Disponível em: [64988-libre.pdf](#). Acesso em: 4 set. 2025.

PIRES, C. F. de O.; ROCHA, R. P. da. Propriedades significativas na curadoria digital e suas relações com a representação da informação do modelo OAIS. In: **Fórum de Estudos em Informação, Sociedade e Ciência** (3.: 2020 out., Porto Alegre, RS). Resumos [recurso eletrônico]. Porto Alegre: UFRGS/PPGCIN, 2020.

POSNER, E. Archives in the Ancient World. Cambridge: **Harvard University Press**, 1972.

PRAXEDES, K. V. **Diretrizes para um programa de preservação e de gestão de documentos arquivísticos nas organizações militares do Comando da Aeronáutica**. 2015. Dissertação (Mestrado Profissional em Gestão de Documentos e Arquivos) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2015. Disponível em: http://www.repositorio-bc.unirio.br:8080/xmlui/bitstream/handle/unirio/11751/KVG%20-%20TCC_final.pdf?sequence=1&isAllowed=y. Acesso em: 10 out. 2025.

PRAXEDES, K. V.; RANGEL, K. da S. Relações entre o vínculo arquivístico e a autenticidade de documentos nato digitais: alguns apontamentos a respeito dos metadados. In: Congresso de Archivología del Mercosur, 12., 2017, Córdoba. Actas del Congreso. Córdoba: **Redes**, 2017. p. 19-32. Disponível em: <http://redarchiveroscordoba.com/wp-content/uploads/2017/10/TOMO-II.pdf>. Acesso em: 10 out. 2025.

PREMIS EDITORIAL COMMITTEE. **PREMIS Data Dictionary for Preservation Metadata, version 3.0. June 2015** – revised November 2015. Disponível em: [PREMIS Data Dictionary for Preservation Metadata, Version 3.0 \(Library of Congress\)](#). Acesso em: 20 out. 2025.

RABELO, N. B.; SCHMIDT, C. M. dos S. Identificação de documentos arquivísticos digitais em sistemas de negócio: perspectivas nacionais e internacionais. **Perspectivas em Ciência da Informação**, [S. l.], v. 30, p. e55641, 2025. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/55641>. Acesso em: 19 fev. 2026.

RANGEL, K. da S. **Revisitando o princípio da proveniência: percepções sobre a organicidade**. 2015. Dissertação (Mestrado Profissional em Gestão de Documentos e Arquivos) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2015. Disponível em: http://www.repositorio-bc.unirio.br:8080/xmlui/bitstream/handle/unirio/11753/Kissila_Versao%20Final.pdf?sequence=1. Acesso em: 10 out. 2025.

RANGEL, K.; SILVA, M. C. S. de M. e. Princípios e características de documentos arquivísticos: algumas questões terminológicas. **Ágora: Arquivologia em Debate**, v. 31, n. 62, p. 1–21, 2021. Disponível em: [Princípios e características de documentos arquivísticos: algumas questões terminológicas | ÁGORA: Arquivologia em debate](#). Acesso em: 04 set. 2025.

RE3DATA. **Registry of Research Data Repositories**. Disponível em: <https://www.re3data.org>. Acesso em: 20 out. 2025.

RFC. RFC 8493/IETF. **The BagIt File Packaging Format**. Disponível em: [RFC 8493: The BagIt File Packaging Format \(V1.0\)](#). Acesso em: 20 out. 2025.

RHEE, H. L. A New Lifecycle Model Enabling Optimal Digital Curation. **Journal of Librarianship and Information Science**, London, 2022. Disponível em: <https://journals.sagepub.com/doi/full/10.1177/09610006221125956>. Acesso em: 10 out. 2025.

RIBEIRO, N. C.; SILVEIRA, L.; SANTOS, S. R. O. **Taxonomia da Ciência Aberta**. 2020. DOI: <https://doi.org/10.6084/m9.figshare.12124002.v4>. Disponível em: <http://repositorio.ufla.br/jspui/handle/1/39498>. Acesso em: 04 set. 2025.

RLG-OCLC. **Trusted Digital Repositories: Attributes and responsibilities**. 2002. Disponível em: [Trusted Digital Repositories - final published version 050... \(oclc.org\)](#). Acesso em 15 out. 2025.

RLG-OCLC. **Trusted Digital Repositories: Requirements and Audit Checklist**. 2002.

Disponível em: [Trustworthy repositories audit & certification: criteria and checklist – OCLC, CRL, NARA – Hipátia](#). Acesso em: 20 set. 2025.

ROCHA, C. L. Glossário multilíngue do Projeto InterPARES 3. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, 2011.

ROCHA, C. L.; RONDINELLI, R. C. Gestão e preservação de documentos arquivísticos digitais: revisitando alguns dos conceitos que as precedem. **Acervo**. 2016. Disponível em: [Gestão e preservação de documentos arquivísticos digitais: revisitando alguns dos conceitos que as precedem. | Acervo](#). Acesso em: 20 out. 2025.

ROCHA, C. L.; SILVA, M. da. Padrões para garantir a preservação e o acesso aos documentos digitais. **Acervo**, v. 20, n. 1, p. 113–124, 2007. Disponível em: <https://revista.an.gov.br/index.php/revistaacervo/article/view/76>. Acesso em: 20 out. 2025.

ROCHA, R. P. da *et al.* Análise dos sistemas DSpace e Dataverse para repositórios de dados de pesquisa com acesso aberto. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 17, p. 1–25, 2021.

RODRIGUES, A. C. **Diplomática contemporânea como fundamento metodológico da identificação de tipologia documental em arquivos**. 2008. Tese (Doutorado) – Universidade de São Paulo, São Paulo, 2008. Disponível em: <http://www.teses.usp.br/teses/disponiveis/8/8138/tde-27112008-151058/>. Acesso em: 17 out. 2025.

RODRIGUES, G. M. **A representação da informação em arquivística: uma abordagem a partir da perspectiva da norma internacional de descrição arquivística**. 2003. Disponível em: <https://repositorio.unb.br/handle/10482/1442>. Acesso em: 20 out. 2025.

ROUSSEAU, J.; COUTURE, C.; ARÈS, F. Os fundamentos da disciplina arquivística. Lisboa: **Publicações Dom Quixote**, 1998.

SÁ, I. P. de. **A Zika enquanto objeto de fronteira: diretrizes para identificação de metadados de dados de pesquisa**. 2019. 334 f. Tese (Doutorado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Universidade Federal Fluminense, 2019. Disponível em: <https://app.uff.br/riuff/handle/1/13854>. Acesso em: 20 out. 2025.

SÁ, I. P. de; ARAÚJO, R. O. de. **Orientações gerais para seleção, avaliação e retenção de dados de pesquisa na Fiocruz**. Rio de Janeiro: Fiocruz; Casa de Oswaldo Cruz, 2024. Em processo de aprovação.

SÁ, I. P. de *et al.* Metodologia para identificação de tipos de dados de pesquisa: a experiência da Fundação Oswaldo Cruz (Fiocruz). **Acervo**, S.l., v. 34, n. 3, p. 1–15, 2021. Disponível em: <https://revista.arquivonacional.gov.br/index.php/revistaacervo/article/view/1764>. Acesso em: 10 out. 2023.

SADIQ, S.; INDULSKA, M. Open data: quality over quantity. **International Journal of Information Management**, v. 37, n. 3, p. 150–154, 2017.

SALES, L. F.; SAYÃO, L. F. A ciência invisível: revelando os dados da cauda longa da pesquisa. In: **XIX Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB)**, 2018, Londrina. Anais [...]. Londrina: [s.n.], 2018.

SALES, L. F.; SAYÃO, L. F. Aumentando a transparência e a colaboração na pesquisa científica: proposta de modelo matricial para apoiar a formulação de programas de ação para Ciência Aberta. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**. Disponível em: <https://www.scielo.br/j/rdbci/a/PBH3PTnk8MZ9gJKfjyw9LBD/?format=pdf&lang=en>. Acesso em: 20 out. 2025.

SALES, L. *et al.* Um panorama histórico da iniciativa GOFAIR: da Europa para o Brasil. In:

Princípios FAIR aplicados à gestão de dados de pesquisa, p. 9–22, 2021.

SANJUAN, J. L. La norma ISO 15489:2001 y su evolución hacia un cambio de estrategia en las organizaciones. *Acervo*, v. 28, n. 2, p. 51–71, jul./dez. 2015.

SANSONE, S. *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, v. 37, n. 4, p. 358–367, 2019.

SANTOS, A. D. S. C. dos *et al.* **Novos cadernos de laboratório e novas culturas epistêmicas: entre a política do experimento e o experimento da política**. 2016. Disponível em: <http://ridi.ibict.br/handle/123456789/943>. Acesso em: 20 out. 2025.

SANTOS, P. R. E. dos. **Arquivística no laboratório: história, teoria e métodos de uma disciplina**. Rio de Janeiro: **Teatral**; Faperj, 2010.

SANTOS, P. R. E. dos. Arquivologia, laboratórios e ciência aberta: contribuições e desafios para a gestão de documentos e dados. *Acervo*, S.l., v. 34, n. 3, p. 1–22, 2021. Disponível em: <https://revista.an.gov.br/index.php/revistaacervo/article/view/1772>. Acesso em: 9 out. 2025.

SANTOS, H. M. dos; FLORES, D. Infraestrutura organizacional necessária ao repositório arquivístico digital confiável: um diálogo com a ISO 16363. *Revista Brasileira de Biblioteconomia e Documentação*, v. 16, p. 1–29, 2020.

SANTOS, H. M. dos; FLORES, D. **Introdução aos conceitos básicos do modelo Open Archival Information System no contexto da arquivística**. 2019. Disponível em: [Introdução aos conceitos básicos do modelo Open Archival Information System no contexto da arquivística | Acervo](https://www.acervo.org.br/index.php/revistaacervo/article/view/1772). Acesso em: 20 out. 2025.

SANTOS, H. M. dos; FLORES, D. O documento digital no contexto das funções arquivísticas. **Páginas a&b: Arquivos e Bibliotecas**, p. 165–177, 2016. Disponível em: <https://ojs.letras.up.pt/index.php/paginasab/article/view/1477>. Acesso em: 20 out. 2025.

SANTOS, H. M. dos; FLORES, D. Modelo lógico da informação no Open Archival Information System: uma reflexão arquivística sobre o pacote de informação para arquivamento. **Perspectivas em Gestão & Conhecimento**, João Pessoa, v. 10, n. 1, p. 23–38, 27 abr. 2020. Disponível em: <https://periodicos.ufpb.br/ojs2/index.php/pgc/article/view/46433>. Acesso em: 10 out. 2025.

SARAMAGO, M. de L. Metadados para a preservação digital e aplicação do Modelo OAIS. In: **Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas**, 8., 2004, Estoril. Anais [...]. Estoril: [s.n.], 2004. Disponível em: <https://www.bad.pt/publicacoes/index.php/congressosbad/article/view/640>. Acesso em: 10 out. 2025.

SAYÃO, L. F. Uma outra face dos metadados: informações para a gestão da preservação digital. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, v. 15, n. 30, p. 1–31, 2010.

SAYÃO, L. F.; SALES, L. F. Afinal, o que é dado de pesquisa? **Biblos: Revista do Instituto de Ciências Humanas e da Informação**, Rio Grande, v. 34, n. 2, p. 32–51, jul./dez. 2020. Disponível em: <https://periodicos.furg.br/biblos/article/view/11875/8426>. Acesso em: 25 jan. 2025.

SAYÃO, L. F.; SALES, L. F. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Informação & Sociedade: Estudos**, v. 22, n. 3, p. 179–191, 2012. Disponível em: http://www.cnen.gov.br/images/CIN/PDFs/GUIA_DE_DADOS_DE_PESQUISA.pdf. Acesso em: 21 jan. 2025.

SAYÃO, L. F.; SALES, L. F. **Guia de Gestão de Dados de Pesquisa para Bibliotecários e Pesquisadores**. Rio de Janeiro: CNEN/IEN, 2015. Disponível em: [\(PDF\) Guia de gestão de](#)

[dados de pesquisa para bibliotecários e pesquisadores](#). Acesso em: 20 out. 2025.

SAYÃO, L. F.; SALES, L. F. Nanopublicações: abordagens de representação do conhecimento como inovação para a comunicação científica. **Informação & Informação**, v. 29, n. 4, p. 247-266, 2024. Disponível em: [Nanopublicações: abordagens de representação do conhecimento como inovação para a comunicação científica | Informação & Informação](#). Acesso em: 20 out. 2025.

SAYÃO, L. F.; SALES, L. F. Plataformas de gestão de dados de pesquisa: expandindo o conceito de repositórios de dados. **Palavra chave**, v. 12, n. 1, p. 171-171, 2022. Disponível em: [Plataformas de gestão de dados de pesquisa: expandindo o conceito de repositórios de dados](#). Acesso em: 20 out. 2025.

SAYÃO, L. F.; SALES, L. F.; CARVALHO SEGUNDO, W. L. R. de. Objeto digital FAIR: ampliando as fronteiras da gestão de dados. **Informação & Sociedade**, João Pessoa, v. 33, 2024. Disponível em: <https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/68646>. Acesso em: 13 out. 2025.

SCHELLENBERG, T. R. **Arquivos modernos**. Rio de Janeiro: FGV Editora, 1974. 2. tiragem.

SCHISLER, M. Propostas para uma preservação digital: trajetórias. In: II Seminário Serviços e Informação em Museus: o trabalho da informação em instituições culturais: em busca de conceitos, métodos e políticas de preservação. São Paulo. **Pinacoteca do Estado de São Paulo**, 2014. Disponível em: [Grátis: II seminário serviços de informação em museus - Material Claro e Objetivo em PDF para Estudo Rápido](#). Acesso em: 20 out. 2025.

SCHMIDT, C. **Documento arquivístico digital e gestão de documentos**: considerações na perspectiva da Arquivística. 2015. Disponível em: [Documento arquivístico digital e gestão de documentos: considerações na perspectiva da Arquivística](#). Acesso em: 20 out. 2025.

SILVA, J. T.; SANTOS, J. A. dos. A Arquivística Brasileira e o Modelo Australiano do Records Continuum. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 14, n. 2, 2019.

SILVA, M. C. S. de M. **Visitando laboratórios**: o cientista e a preservação de documentos. 2007. Tese (Doutorado) – Universidade de São Paulo, São Paulo, 2007.

SILVA, M. H. F. X. da S. *et al.* Competências dos bibliotecários na gestão dos dados de pesquisa. **Ciência da Informação**, Brasília, v. 48, n. 3, 2019.

SILVA, S. C. de A. **A preservação da informação arquivística governamental nas políticas públicas do Brasil**. Rio de Janeiro: Associação dos Arquivistas Brasileiros; FAPERJ, 2008.

SILVEIRA, L. da *et al.* Taxonomia da Ciência Aberta: revisada e ampliada. **Encontros Bibli**: Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 28, e91712, 2023. Disponível em: <https://brapci.inf.br/index.php/res/v/225742>. Acesso em: 10 out. 2025.

SIMMHAN, Y. L.; PLALE, B.; GANNON, D. A survey of data provenance in e-science. **ACM SIGMOD Record**, v. 34, n. 3, p. 31–36, 2005.

SOEHNER, C.; STEEVES, C.; WARD, J. E-science e serviços de suporte de dados: um estudo de instituições membros da ARL. **Associação de Bibliotecas de Pesquisa**, 2010.

SOUSA, R. T. B. Os princípios arquivísticos e o conceito de classificação. In: RODRIGUES, Georgete Medleg; LOPES, Ilza Leite (org.). Organização e representação do conhecimento na perspectiva da Ciência da Informação. Brasília: **Thesaurus**, 2003. v. 2, p. 240–269.

STRASSER, K. **Primer on research data management**. 2015. Disponível em: <https://www.niso.org/publications/primer-research-data-management>. Acesso em: 4 set. 2025.

STRONG, D. M.; LEE, Y. W.; WANG, R. Y. Data quality in context. **Communications of the ACM**, v. 40, n. 5, p. 103–110, 1997.

STRONG, M. A. Digital curation: a how-to-do-it manual. **Journal of the Medical Library Association**, Chicago, v. 100, n. 2, p. 148, 2012. Disponível em: [Harvey, Ross. Digital Curation: A How-To-Do-It Manual - Document - Gale Academic OneFile](#). Acesso em: 10 out. 2025.

THOMASSEN, T. The development of archival science and its European dimension. In: Seminar for Anna Christina Ulfspärre. Stockholm: **Swedish National Archives**, fev. 1999.

TOGNOLI, N. B. **A contribuição epistemológica canadense para a construção da arquivística contemporânea**. 2010. 119 f. Dissertação (Mestrado) – Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2010.

TURKYILMAZ-VAN DER VELDEN, Y.; DINTZNER, N.; TEPERER, M. Reproducibility starts from you today. **Patterns**, Amsterdam, v. 1, n. 6, 2020. Disponível em: [https://www.cell.com/patterns/pdf/S2666-3899\(20\)30133-1.pdf](https://www.cell.com/patterns/pdf/S2666-3899(20)30133-1.pdf). Acesso em: 10 out. 2025.

UNESCO. Organização das Nações Unidas para Educação, Ciência e Cultura. **Declaração sobre a Ciência e o Uso do Conhecimento Científico**. Budapeste, 1999. Disponível em: <http://www.dominiopublico.gov.br/download/texto/ue000111.pdf>. Acesso em: 2 fev. 2025.

UNESCO. Organização das Nações Unidas para Educação, Ciência e Cultura. **Recomendação da Unesco sobre ciência aberta**. S.l.: Unesco, 2022. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000379949_por. Acesso em: 10 out. 2025.

UNIVERSIDADE DE SÃO PAULO (USP). Instituto de Relações Internacionais. **Repositório de dados científicos da USP**. [S. l.: s. n.], [202?]. Disponível em: <https://www.iri.usp.br/br/pesquisa/repositorio-de-dados-cientificos-da-usp>. Acesso em: 20 out. 2025.

UPWARD, F. H. The records continuum. In: **Archives: recordkeeping in society**. Centre for Information Studies, Charles Sturt University, 2005. p. 197–222.

URE, J. *et al.* Curating complex, dynamic and distributed data: telehealth as a laboratory for strategy. **International Journal of Digital Curation**, v. 6, n. 2, p. 128–145, 2011. Disponível em: [Curating Complex, Dynamic and Distributed Data: Telehealth as a Laboratory for Strategy | International Journal of Digital Curation](#). Acesso em: 4 set. 2025.

UTRECHT UNIVERSITY. **Research Data Management Support: Guide**. [S.l.]: Utrecht University. Disponível em: [Guides - Research Data Management Support - Utrecht University](#). Acesso em: 20 out. 2025.

VAN DE SANDT, S. *et al.* The definition of reuse. **Data Science Journal**, v. 18, p. 22–22, 2019.

VARNIENÉ-JANSSEN, R.; KUPRIENĖ, J. Authenticity and provenance in long-term digital preservation: analysis of the scope of content. **Informacijos Mokslai**, v. 82, p. 131–160, 2018. Disponível em: [Authenticity and Provenance in Long-Term Digital Preservation: Analysis of the Scope of Content](#). Acesso em: 20 out. 2025.

VIDOTTI, S. A. B. G.; TORINO, E.; CONEGLIAN, C. S. #SejaJUSTOeCUIDADOSO: princípios FAIR e CARE na gestão de dados de pesquisa. In: **Princípios FAIR aplicados à gestão de dados de pesquisa**, 2021. Disponível em: [principiosfairdadospesquisa.pdf](#). Acesso em: 20 out. 2025.

VILELA, M. L.; SELLES, S. E. É possível uma educação em ciências crítica em tempos de negacionismo científico? **Caderno Brasileiro de Ensino de Física**, v. 37, n. 3, p. 1722–1747, 2020. Disponível em: <https://doi.org/10.5007/2175-7941.2020v37n3p1722>. Acesso em: 20 out. 2025.

WA out. ND, Y.; WANG, R. Y. Anchoring data quality dimensions in ontological foundations. **Communications of the ACM**, v. 39, n. 11, p. 86–95, 1996. Disponível em: <https://doi.org/10.1145/240455.24047>. Acesso em: 20 out. 2025.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: what data quality means to data consumers. **Journal of Management Information Systems**, v. 12, n. 4, p. 5–33, 1996. Disponível em: [Beyond Accuracy: What Data Quality Means to Data Consumers: Journal of Management Information Systems: Vol 12, No 4](#). Acesso em: 20 out. 2025.

WEBB, C. Practical digital preservation: a how-to guide for organizations of any size. **Archives and Manuscripts**, v. 42, n. 1, p. 110–113, 2014. Disponível em: <https://doi.org/10.1080/01576895.2014.888126>. Acesso em: 20 out. 2025.

WEBSTER, F. **Theories of the Information Society**. London: Routledge, 2014. Disponível em: <https://doi.org/10.4324/9781315867854>. Acesso em: 20 out. 2025.

WILKINSON, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 1, 2016. Disponível em: <https://doi.org/10.1038/sdata.2016.18>. Acesso em: 20 out. 2025.

YEO, G. Contexts, original orders, and item-level orientation: responding creatively to users' needs and technological change. **Journal of Archival Organization**, v. 12, n. 3–4, p. 170–185, 2015. Disponível em: <https://doi.org/10.1080/15332748.2015.1048626>. Acesso em: 20 out. 2025.

ZUIDERWIJK, A.; SHINDE, R.; JENG, W. What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption. **PLoS One**, v. 15, n. 9, p. e0239283, 2020. Disponível em: <https://doi.org/10.1371/journal.pone.0239283>. Acesso em: 20 out. 2025.

APÊNDICE A – STATUS DE IDENTIFICAÇÃO

Datasets	Categorias	Tipos encontrados por observação de registro de itens	Status
Dataset 1	Proveniência		A
	Contexto		A
	Fixidez		A
Dataset 2	Proveniência		A
	Contexto	Relacionamento (URL alternativo, Outro ID)	P
	Fixidez		A
Dataset 3	Proveniência		A
	Contexto		A
	Fixidez		A
Dataset 4	Proveniência		A
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A
Dataset 5	Proveniência		A
	Contexto		A
	Fixidez		A
Dataset 6	Proveniência		A
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A
Dataset 7	Proveniência		A
	Contexto		A
	Fixidez		A
Dataset 8	Proveniência		A
	Contexto		A
	Fixidez		A
Dataset 9	Proveniência		A
	Contexto		A
	Fixidez		A
Dataset 10	Proveniência		A
	Contexto		A
	Fixidez		A
Dataset 11	Proveniência		A
	Contexto	Relacionamento (Classificação de tópicos)	P
	Fixidez		A
Dataset 12	Proveniência	Fonte ou origem (Produtor) Histórico (Período coberto)	P
	Contexto	Relações com ambiente (Tipo de dados, <i>software</i>) Relacionamento (URL alternativo)	P
	Fixidez		A

Datasets	Categorias	Tipos encontrados por observação de registro de itens	Status
Dataset 13	Proveniência	Fonte ou origem (Produtor) Histórico (Data de coleta)	P
	Contexto	Relações com ambiente (Tipo de dados) Relacionamento (Classificação de tópicos)	P
	Fixidez		A
Dataset 14	Proveniência	Fonte ou origem (Produtor)	P
	Contexto	Relações com ambiente (Tipo de dados) Relacionamento (Classificação de tópicos)	P
	Fixidez		A
Dataset 15	Proveniência	Fonte ou origem (Produtor) Histórico (Data de coleta, Período coberto)	P
	Contexto	Relações com ambiente (Tipo de dados)	P
	Fixidez		A
Dataset 16	Proveniência	Histórico (Data de produção)	P
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A
Dataset 17	Proveniência	Histórico (Data de coleta, Cobertura geográfica, Classificação taxonômica, Métodos)	P
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A
Dataset 18	Proveniência	Histórico (Data de coleta, Cobertura geográfica, Caixa delimitadora geográfica)	P
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A
Dataset 19	Proveniência	Histórico (Notas)	P
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A
Dataset 20	Proveniência		A
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A
Dataset 21	Proveniência		A
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A
Dataset 22	Proveniência		A
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A
Dataset 23	Proveniência		A
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A

Datasets	Categorias	Tipos encontrados por observação de registro de itens	Status
Dataset 24	Proveniência	Histórico (Notas, Data de coleta)	P
	Contexto	Relacionamento (Publicação relacionada)	P
	Fixidez		A

APÊNDICE B – TABELA DE VERIFICAÇÃO DE ARQUIVOS DE DADOS DOS DATASETS

Dataset	Qtde arquivos	Tipos de arquivo	Formatos	Tipo de acesso	Verificação de itens de informação	Observação (caso haja)
D1	1	Documento	.tab	Restrito	Não foi possível verificar	N/A
D2	75	Documento Imagem Áudio	.xlsx; .mp3; .pdf; .png	Público sob contrato de licença/uso de dados	a) Foi verificada a configuração dos arquivos em .xlsx e .pdf. Na descrição disponível nas propriedades dos documentos consta: Nome do arquivo; Tamanho do arquivo; autor; criado em; modificado em; criador. b) Não foi possível verificar os arquivos dos outros formatos.	Constam como dados: postos, atas de reunião, relatórios, apresentações, relatos, programação, convites, planejamento, entre outros
D3	1	Documento	.tab	Restrito	Não foi possível verificar	N/A
D4	2	Documento	.xlsx	Público sob contrato de licença/uso de dados	Foi verificada a configuração de cada arquivo. Na descrição disponível nas propriedades dos documentos constam: Datas (última modificação e criado em) e Pessoas relacionadas (Autor e última modificação).	N/A
D5	1	Documento	.tab	Restrito	Não foi possível verificar	N/A
D6	1	Documento	.pdf	Público	Foi verificada a configuração do arquivo. Na descrição disponível nas propriedades do documento consta: Nome do arquivo; Tamanho do arquivo; criado em; modificado em; criador.	Na descrição do arquivo são correlacionados os autores às suas atividades
D7	13	Documento	.pdf .tab	Restrito	Não foi possível verificar	Pelo título do arquivo em pdf, trata-se de um dicionário de termos

Dataset	Qtde arquivos	Tipos de arquivo	Formatos	Tipo de acesso	Verificação de itens de informação	Observação (caso haja)
D8	1	Sem especificação	.accdB	Público	Não foi possível verificar	Banco de dados desenvolvido no Microsoft Access para registrar os documentos encontrados na pesquisa
D9	9	Documento e imagem	.pdf .png	Público	Foi verificada a configuração de cada arquivo. Na descrição disponível nas propriedades dos documentos constam: Nome do arquivo; Tamanho do arquivo; autor; criado em; modificado em; criador.	São 8 arquivos do tipo documento. 7 são tabelas correlacionadas (com a <i>tag</i> documentation) e 1 dado (com a <i>tag</i> (data))
D10	30	Sem especificação	.vcf.gz	Restrito	Não foi possível verificar	N/A
D11	1	Sem especificação	.tab	Restrito	Não foi possível verificar	Consta na descrição se tratar de banco de dados
D12	25	Dados tabulares	.tab	Público	Necessário download. Abertura do arquivo mediante uso da ferramenta Bloco de notas. Não foi possível verificar configurações do documento nessa ferramenta.	N/A
D13	1	Textual	.tab	Restrito	Não foi possível verificar	N/A
D14	1	Textual	.tab	Público	Necessário download. Abertura do arquivo mediante uso da ferramenta Bloco de notas. Não foi possível verificar configurações do documento nessa ferramenta.	N/A

Dataset	Qtde arquivos	Tipos de arquivo	Formatos	Tipo de acesso	Verificação de itens de informação	Observação (caso haja)
D15	2	Textual	.tab .pdf	Público	a) Necessário download. Abertura do arquivo mediante uso da ferramenta Bloco de notas. Não foi possível verificar configurações do documento nessa ferramenta; b) O arquivo em pdf tem o título "metadados", no entanto, não foi possível conferi-lo (estava dando erro no download).	N/A
D16	5	Textual	.txt	Público	Abertura dos arquivos como bloco de notas no próprio sistema, portanto, não foi possível verificar configurações dos documentos. Informações dos documentos disponíveis, exclusivamente, na aba "metadados" dos arquivos.	4 Dados + 1 <i>Read.me</i>
D17	2	Textual	.txt	Público	Abertura dos arquivos como bloco de notas no próprio sistema, portanto, não foi possível verificar configurações dos documentos. Informações dos documentos disponíveis exclusivamente na aba "metadados" dos arquivos.	1 Dado + 1 <i>Read.me</i>
D18	2	Textual	.csv .txt	Público	Abertura dos arquivos como bloco de notas no próprio sistema, portanto, não foi possível verificar configurações dos documentos. Informações dos documentos disponíveis, exclusivamente, na aba "metadados" dos arquivos.	1 Dado + 1 <i>Read.me</i>

Dataset	Qtde arquivos	Tipos de arquivo	Formatos	Tipo de acesso	Verificação de itens de informação	Observação (caso haja)
D19	2	Documento Textual	.doc .txt	Público	a) Foi verificada a configuração do arquivo em .doc. Na descrição disponível nas propriedades do documento consta: Datas (última modificação e criado em) e Pessoas relacionadas (Autor e última modificação); b) O arquivo em .txt foi aberto como bloco de notas no próprio sistema, portanto, não foi possível verificar configurações do documento. Informações do documento disponíveis, exclusivamente, na aba "metadados" dos arquivos.	1 Dado + 1 <i>Read.me</i>
D20	2	Textual	.tab .txt	Público	Abertura dos arquivos como bloco de notas no próprio sistema, portanto, não foi possível verificar configurações dos documentos. Informações dos documentos disponíveis, exclusivamente, na aba "metadados" dos arquivos.	1 Dado + 1 <i>Read.me</i>
D21	2	Documento	.pdf	Público	Abertura dos arquivos como bloco de notas no próprio sistema, portanto, não foi possível verificar configurações dos documentos. Informações dos documentos disponíveis, exclusivamente, na aba "metadados" dos arquivos.	1 Dado + 1 <i>Read.me</i>

Dataset	Qtde arquivos	Tipos de arquivo	Formatos	Tipo de acesso	Verificação de itens de informação	Observação (caso haja)
D22	2	Textual	.pdf .docx	Público sob contrato de licença/uso de dados	a) Aparentemente, pelo título, o arquivo em formato .docx é uma versão do documento em .pdf, no entanto, ele não está disponível para ser aberto. b) O arquivo .pdf é aberto no próprio sistema. Não foi possível verificar configurações do documento. Informações do documento disponíveis, exclusivamente, na aba "metadados" do arquivo.	N/A
D23	2	Dado tabular Textual	.tab .txt	Público	Abertura dos arquivos como bloco de notas no próprio sistema, portanto, não foi possível verificar configurações dos documentos. Informações dos documentos disponíveis, exclusivamente, na aba "metadados" dos arquivos.	1 Dado + 1 <i>Read.me</i>
D24	2	Documento Textual	.xls .txt	Público	a) Foi verificada a configuração do arquivo em .xls. Na descrição disponível nas propriedades do documento consta: Datas (última modificação e criado em) e Pessoas relacionadas (Autor e última modificação); b) O arquivo em .txt foi aberto como bloco de notas no próprio sistema, portanto, não foi possível verificar configurações do documento. Informações do documento disponíveis, exclusivamente, na aba "metadados" dos arquivos.	1 Dado + 1 <i>Read.me</i>

ANEXO A - PROPOSTA DE TEMPLATE DE PLANO DE GESTÃO DE DADOS DO ICICT/FIOCRUZ

Seção A: Informação Administrativa
A1. Informe o nome completo do pesquisador principal
A2. Informe o e-mail de Contato
A3. Qual o seu telefone?
A4. Informe o link do Currículo Lattes do Pesquisador Principal
A5. Informe o link do ORCID do Pesquisador Principal.
A6. Informe o título do seu projeto
A7. Descreva o resumo do projeto
A8. A qual Unidade da Fiocruz o seu projeto está vinculado?
A9. Qual a data de início do projeto?
A10. Indique a previsão de término do projeto
A11. O projeto tem financiamento?
A12. Especifique o financiamento
A13. Qual a versão deste Plano de Gestão de Dados?
A14 Quem será o responsável pelo preenchimento e atualização do Plano de Gestão de Dados?
Seção B: Descrição dos dados coletados ou reuso de dados existentes
B1. Como os dados serão coletados, produzidos ou como os dados existentes serão reutilizados?
B2. Quais os tipos de dados que serão coletados ou produzidos?
B3. Quais os formatos de dados que serão coletados ou produzidos?
B4. Qual o volume aproximado dos dados coletados ou produzidos?
Seção C: Documentação e Qualidade dos Dados
C1. Indique os metadados adotados
C2. Indique o padrão de metadados adotado
C3. Indique os documentos que acompanharão os dados
C4. Indique as medidas adotadas para a organização e controle de qualidade dos dados
Seção D: Armazenamento e Backup durante o processo de pesquisa
D1. Como os dados serão armazenados e como serão feitas as cópias de segurança durante a pesquisa?
D2. Como a segurança dos dados e a proteção dos dados sensíveis serão tratadas durante a pesquisa?
Seção E: Requisitos Legais, Éticos e de Códigos de Conduta
E1. Se os dados pessoais forem tratados, como será assegurado o cumprimento da legislação relativa aos dados pessoais e à proteção dos dados?
E2. Como as questões legais referentes aos direitos de propriedade intelectual serão gerenciadas? Qual legislação se aplica?
E3. Indique a licença que será aplicada.
E4. Como as possíveis questões éticas e os códigos de conduta serão levados em consideração?

Seção F: Compartilhamento de Dados e Preservação a longo prazo
F1. Como e quando os dados serão compartilhados? Existem possíveis restrições ao compartilhamento e/ou à abertura de dados ou motivos para embargo?
F2. Como os dados para preservação serão selecionados e onde os dados serão preservados a longo prazo (por exemplo, em um repositório de dados)?
F3. Quais métodos ou ferramentas de software serão necessários para acessar e usar os dados?
F4. Como será assegurado o registro de um identificador único e persistente (como um DOI - Digital Object Identifier) para cada conjunto de dados?
Seção G: Responsabilidades e Recursos Financeiros para a Gestão de Dados de Pesquisa
G1. Quem será o responsável pela gestão dos dados?
G2. Quais recursos (ex.: financeiros, tempo) serão dedicados à gestão de dados e à garantia que eles sejam FAIR (localizável, acessível, interoperável, reutilizável)?
G3. Para projetos colaborativos, explique como será a coordenação da gestão de dados e as responsabilidades de cada um dos parceiros.

Fonte: VEIGA *et al.* (2019, p.281).¹²²

¹²² O formulário online para preenchimento pelos pesquisadores está disponível em: <https://fiodmp.fiocruz.br/>

ANEXO B – METODOLOGIA PARA IDENTIFICAÇÃO DA TIPOLOGIA DE DADOS DA COC/FIOCRUZ

Bloco 1: Aspectos gerais
1. Questão de pesquisa.
2. Resumo das funções e atividades realizadas na pesquisa.
3. Propósito científico da coleta de dados (função do dado na pesquisa). Por que os dados foram produzidos?
4. Domínio do conhecimento em que os dados foram produzidos.
5. A quem se destina, qual é o público-alvo dos dados na pesquisa?
6. Quais as formas de uso dos dados?
7. Como percebem o uso dos dados em novas pesquisas por outros pesquisadores?
8. Consideram que os conjuntos de dados possuem valor imediato e duradouro?
9. Consideram que os conjuntos de dados são passíveis de reprodução?
Bloco 2: Características descritivas dos conjuntos de dados de pesquisa
1. Todos os dados foram produzidos no formato digital (nato digital) ou existem dados analógicos? Em caso afirmativo, estes serão digitalizados?
2. Nome/título da coleção de dados (o que).
3. Autores/produtores dos dados (quem).
4. Resumo da coleção de dados.
5. Período em que os dados foram coletados: período de tempo que cobre (quando).
6. Escopo geográfico onde os dados foram coletados (onde).
7. Como os dados foram coletados? Descrição dos métodos, práticas e instrumentos utilizados na coleta dos dados (cópia do formulário de consentimento, configuração das entrevistas, critérios para seleção dos sujeitos das entrevistas, instruções dadas aos entrevistadores, instrumento de coleta, como os questionários/roteiros das entrevistas, lista de entrevistas).
8. Origem dos dados: computacionais, experimentais, observacionais.
9. Tipos (natureza) de dados: áudio, vídeo, imagem, texto, entre outros.
10. Status do fluxo de trabalho da pesquisa segundo sua fase: dados brutos, crus ou preliminares, derivados, canônicos ou referenciais.
11. Qual a quantidade de conjuntos de dados produzidos por tipos (natureza)?
12. O que identificam como peculiar (características muito específicas) nos dados produzidos?
Bloco 3: Características técnicas e organizacionais dos conjuntos de dados de pesquisa
1. Qual é o esquema de organização dos conjuntos de dados de pesquisa?

2. Organização do armazenamento dos conjuntos de dados: nomes dos arquivos que compõem a coleção. Como os conjuntos de dados, arquivos, pastas, diretórios e tabelas de banco de dados foram nomeados e como se relacionam? Utilizam algum padrão para nomenclatura?
3. Formato dos arquivos.
4. Foram realizadas conversões de formato de arquivo? Em caso afirmativo, por quê?
5. Tamanho dos arquivos.
6. Existe relacionamento entre os conjuntos de dados?
7. Unidades de medida utilizadas.
8. Como os dados foram processados? Métodos e procedimentos utilizados para o processamento de dados, por exemplo, softwares
9. Foram realizadas modificações nos arquivos dos dados? Em caso afirmativo, essas modificações foram registradas e datadas?
10. Software (incluindo o número da versão) usado para preparar e ler a coleção de dados.
11. Propósito científico da coleta de dados (função do dado na pesquisa). Por que os dados foram produzidos?
Bloco 4: Documentação
1. Que tipos de documentos foram criados que podem apoiar o entendimento sobre como os dados foram produzidos, processados e de como estão estruturados, organizados, nomeados? Por exemplo, de campo, questionários, roteiros de entrevistas, modelos de descrição etc.
2. Existem dados armazenados em bases de dados?
3. Quais metadados são utilizados para a descrição dos dados e para o registro da proveniência? Utilizam algum padrão ou esquema de metadados específico?
4. Fazem uso de taxonomias, vocabulários controlados?
5. Utilizaram dados de outras pesquisas ou de outras fontes?
Bloco 5: Qualidade
1. Quais verificações (ações de controle e inspeção) foram utilizadas para encontrar valores inválidos, respostas inconsistentes, registros incompletos etc.?
2. Quais verificações são e serão usadas para gerenciar as versões dos dados?
Bloco 6: Preservação
1. Consideram que todos os dados devem ser preservados?
2. Quais seriam os dados mais relevantes e que deveriam ser priorizados em um planejamento de preservação?
3. Existe alguma orientação específica da área em relação a preservação dos dados?

Fonte: PEREIRA *et al.* (2021, p. 10).