



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Débora Gomes de Araújo Rodrigues

**ELEMENTOS DE CICLOS DE VIDA DOS DADOS NO PERCURSO
METODOLÓGICO DAS TESES BRASILEIRAS DA ÁREA DE CIÊNCIA DA
INFORMAÇÃO: Um estudo diagnóstico**

João Pessoa
2021

DÉBORA GOMES DE ARAÚJO RODRIGUES

**ELEMENTOS DE CICLOS DE VIDA DOS DADOS
NO PERCURSO METODOLÓGICO DAS TESES BRASILEIRAS DA ÁREA DE
CIÊNCIA DA INFORMAÇÃO: Um estudo diagnóstico**

Tese apresentada ao Programa de Pós-Graduação em
Ciência da Informação da Universidade Federal da
Paraíba, como requisito parcial para a obtenção do
título de Doutora em Ciência da Informação.

Linha de pesquisa: Organização, Acesso e Uso da
Informação.

Orientador: Prof. Dr. Guilherme Ataíde Dias

João Pessoa

2021

Catálogo na publicação
Seção de Catalogação e Classificação

R696e Rodrigues, Débora Gomes de Araújo.

Elementos de ciclos de vida dos dados no percurso metodológico das teses brasileiras da área de Ciência da Informação : um estudo diagnóstico / Débora Gomes de Araújo Rodrigues. - João Pessoa, 2021.
183 f. : il.

Orientação: Guilherme Ataíde Dias.
Tese (Doutorado) - UFPB/CCSA.

1. Dados científicos. 2. Ciclos de vida dos dados. 3. Ciência da Informação. 4. Metodologia científica. 5. eScience. I. Dias, Guilherme Ataíde. II. Título.

UFPB/BC

CDU 001.103(043)



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

ATA DE DEFESA DE TESE

Defesa nº 53

Ata da Sessão Pública de Defesa de Tese da
Doutoranda **DÉBORA GOMES DE ARAÚJO
RODRIGUES** como requisito para obtenção do
grau de Doutora em Ciência da Informação, Área
de Concentração em Informação, Conhecimento
e Sociedade e com Linha de Pesquisa em
Organização, Acesso e Uso da Informação.

Aos vinte e sete dias do mês de setembro de dois mil e vinte e um (27/09/2021), às catorze horas (a defesa terminou às dezessete horas), na sala virtual do Google Meet, conectaram-se via videoconferência a banca examinadora designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação para avaliar a candidata ao Grau de Doutora em Ciência da Informação na Área de Concentração Informação, Conhecimento e Sociedade, a doutoranda **DÉBORA GOMES DE ARAÚJO RODRIGUES**. Devido à pandemia do novo Coronavírus (COVID-19), considerando as estratégias de distanciamento social para contenção pandêmica e a Portaria N° 323/GR/REITORIA/UFPB, de 16 de outubro de 2020, a videoconferência da defesa ocorreu com acesso por meio do link: (<https://meet.google.com/icg-dads-juj>) A banca examinadora foi composta pelos(as) professores(as): Dr. Guilherme Ataíde Dias – PPGCI/UFPB (Presidente/Orientador); Dr. Marckson Roberto Ferreira de Sousa – PPGCI/UFPB (Examinador Interno); Dra. Marynice de Medeiros Matos Autran – PPGCI/UFPB (Examinadora Interna); Dra. Sandra de Albuquerque Siebra – UFPE (Examinadora Externa); Dr. Luis Fernando Sayão – UFRJ (Examinador Externo); Dr. Fabiano Couto Corrêa da Silva – UFRGS (Examinador Externo); Dr. Wagner Junqueira Araújo – PPGCI/UFPB (Suplente Interno) e Dra. Ana Carolina Simionato Arakaki – UFSCAR (Suplente Externa). Dando início aos trabalhos, o Professor Dr. Guilherme Ataíde Dias, Presidente da Banca Examinadora, explicou aos presentes a finalidade da sessão e passou a palavra à discente para que fizesse oralmente a apresentação do trabalho de tese intitulado: **ELEMENTOS DE CICLOS DE VIDA DOS DADOS NO**

PERCURSO METODOLÓGICO DAS TESES BRASILEIRAS DA ÁREA DE CIÊNCIA DA INFORMAÇÃO: Um estudo diagnóstico. Após a apresentação, a candidata foi arguida na forma regimental pelos examinadores. Respondidas todas as arguições, o Professor Dr. Guilherme Ataíde Dias, Presidente da Banca Examinadora, acatou todas as observações da banca e procedeu para o julgamento do trabalho, concluindo por atribuir-lhe o conceito:

(X)Aprovado ()Indeterminado ()Reprovado.

Proclamados os resultados e encerrados os trabalhos, eu, o Professor Dr. Guilherme Ataíde Dias, Presidente da Banca Examinadora, lavrei a presente ata que segue assinada por mim, como representante dos participantes da banca, juntamente com os pareceres de avaliação da TESE e da defesa de tese da doutoranda, devidamente assinados por seus respectivos avaliadores e em formato digital.

João Pessoa, 27 de setembro de 2021.



Prof. Dr. Guilherme Ataíde Dias

Presidente da Banca/Orientador – PPGCI/UFPB

DÉBORA GOMES DE ARAÚJO RODRIGUES

**ELEMENTOS DE CICLOS DE VIDA DOS DADOS
NO PERCURSO METODOLÓGICO DAS TESES BRASILEIRAS DA ÁREA DE
CIÊNCIA DA INFORMAÇÃO: Um estudo diagnóstico**

Tese apresentada ao Programa de Pós-Graduação em
Ciência da Informação da Universidade Federal da
Paraíba, como requisito parcial para a obtenção do
título de Doutora em Ciência da Informação.

Aprovada em 27/10/2021

BANCA EXAMINADORA

Prof. Dr. Guilherme Ataíde Dias
(Orientador/Universidade Federal da Paraíba)

Profa. Dra. Marynice de Medeiros Matos Autran
(Membro Interno/Universidade Federal da Paraíba)

Prof. Dr. Marckson Roberto Ferreira de Sousa
(Membro Interno/Universidade Federal da Paraíba)

Profa. Dra. Sandra de Albuquerque Siebra
(Membro Externo/Universidade Federal de Pernambuco)

Prof. Dr. Luís Fernando Sayão
(Membro Externo/Universidade Federal do Rio de Janeiro)

Prof. Dr. Fabiano Couto Corrêa da Silva
(Membro Externo/ Universidade Federal do Rio Grande do Sul)

Prof. Dr. Wagner Junqueira de Araújo
(Membro Suplente Interno/Universidade Federal da Paraíba)

Profa. Dra. Ana Carolina Simionato Arakaki
(Membro Suplente Externo/Universidade Federal de São Carlos)

Dedico este trabalho a Deus.

Ao meu esposo, Dr. Francisco Rodrigues, por todo suporte.

À minha mãe, Delma Gomes, a quem devo as condições de continuar os estudos.

À minha avó, Maria Joanete (*in memoriam*). Eternas saudades.

Ao meu orientador, prof. Dr. Guilherme Ataíde Dias, por todo conhecimento compartilhado e apoio na minha vida acadêmica.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por me capacitar nessa jornada. Quando as forças me faltaram, Ele me fez ver que o Seu poder se aperfeiçoa nas minhas fragilidades e limitações.

À minha mãe, por ter me proporcionado as condições para que eu me dedicasse com exclusividade ao doutorado. Agradeço pelos ensinamentos, carinho, amor e compreensão.

Ao meu esposo, Francisco Rodrigues, por todo amor, companheirismo, suporte e encorajamento. As palavras não são suficientes para expressar o quanto ele faz parte da concretização desse sonho.

Aos meus avós paternos, Antônio Arnaldo (*in memoriam*) e Maria Joanete (*in memoriam*) que, desde a minha infância, me acolheram com tanto amor. Meu avô, conhecido como Totô das charadas, foi um amante do conhecimento, um exemplo para a minha vida. Minha avó, que sempre foi uma amiga inseparável, partiu na construção desse trabalho e deixou uma saudade irreparável.

Ao meu pai, por ter se aproximado mais nesse período e por todos os seus atos de amor e carinho.

À minha tia, Ana Maria, por todas as orações, palavras de força, conselhos e apoio emocional.

Aos meus irmãos, Daiany, Marcos, Jocasta e Ana Júlia, que se alegram com as minhas conquistas e me dão força nas minhas adversidades.

A toda minha família, materna e paterna, que sempre se fez presente, alegrando-se com minhas conquistas.

À minha sogra, exemplo de mulher de fé, pelos ensinamentos e palavras de ânimo.

Aos meus amigos, que sempre torceram por mim, em especial Esthelina, Milena, Renata e Ana Cecylia (Cecy). Esta me deu a honra de corrigir o português do meu texto e tem me acompanhado desde a minha graduação.

À Adenice Ferreira que, desde o mestrado, vem me oferecendo um suporte no tratamento dos dados estatísticos.

A todos os que fazem parte do Programa de Pós-Graduação em Ciência da Informação da UFPB. Sou grata por toda a ajuda concedida e ensinamentos compartilhados.

Aos membros da banca, que me apontaram novos caminhos, me fizeram vislumbrar novos horizontes que enriqueceram este trabalho. Gratidão por todas as considerações feitas visando à melhoria desta tese.

Ao meu orientador, Prof. Dr. Guilherme Ataíde, pela orientação desde o mestrado, por toda compreensão e apoio demonstrados durante as atividades acadêmicas e de pesquisa. A quem devo grande parte das minhas conquistas acadêmicas.

À Profa. Dra. Marynice Autran, sempre atenciosa, sou grata por todos os seus ensinamentos dentro e fora da sala de aula.

À Profa. Dra. Sandra Siebra, pelas contribuições enriquecedoras que tanto ajudaram no desenvolvimento desse trabalho e por todo conhecimento transmitido nas aulas de preservação da informação digital.

Ao prof. Dr. Luis Fernando Sayão, a minha gratidão por aceitar fazer parte deste trabalho e por todas as contribuições e ideias sugeridas.

Ao prof. Dr. Fabiano Couto Corrêa da Silva, pelas contribuições e ideias propostas.

Ao prof. Dr. Marckson, sempre disposto a ajudar, gratidão por todas as contribuições na sala de aula e neste trabalho de tese.

A todos que acreditaram no meu potencial, muito obrigada! Essa conquista também é de vocês.

RESUMO

Ações associadas ao tratamento de dados científicos podem ser observadas com frequência nas etapas metodológicas inerentes aos processos de investigações científicas. Neste contexto, o objetivo do presente estudo é analisar possíveis intersecções entre os elementos componentes dos ciclos de vida dos dados e os percursos metodológicos apresentados nas teses brasileiras, na área da Ciência da Informação. A tese defende que esses elementos associados aos ciclos de vida dos dados pertencentes às iniciativas do DataONE, do *Digital Curation Centre* e da *Data Documentation Initiative* estão presentes nas metodologias das teses brasileiras da referida área. Para tanto, foi desenvolvido um modelo de referência da interseção dessas iniciativas para realização do estudo. O percurso metodológico da pesquisa possui uma abordagem quali-quantitativa; caracteriza-se quanto aos meios como bibliográfica e documental; e quanto aos fins como um estudo descritivo. A técnica de análise de conteúdo foi utilizada para analisar os dados com o suporte do *software* QDAMiner. Os resultados apontaram que alguns Programas de Pós-graduação em Ciência da Informação possuem dificuldades em dar acesso às próprias teses e na identificação das metodologias nos respectivos trabalhos. Foi verificado que os dados da área se distinguem quando a natureza por serem textuais, visualizações, multimídia, números e *software*, além disso, por apresentarem as origens de registro, observacional, computacional e experimental. Em relação à presença dos elementos dos ciclos de vida dos dados foi constatado que assim como na literatura que trata sobre as metodologias nas ciências sociais, algumas etapas como planejar, coletar e analisar são abordadas de forma mais clara. Porém, não foi possível identificar que os pesquisadores adotaram o plano de gestão de dados no momento do planejamento do estudo. Outra questão identificada é que apesar da etapa descrever ter apresentado um percentual notório de participação nas metodologias, não significa dizer que existe uma documentação robusta para descrever esses dados, pois nesse momento foram codificados os trechos aleatórios que correspondiam a algum elemento de metadados. As etapas organizar e assegurar a qualidade também foram identificadas nos trabalhos, as etapas encontrar e reutilizar foram vistas de forma muito sutil e a etapa preservar não foi evidenciada. A realidade demonstrou que os pesquisadores cumprem algumas ações referentes às etapas dos ciclos de forma instintiva e, por isso, há a necessidade de um melhor acesso aos resultados dos estudos da área. Espera-se que essa pesquisa, ao diagnosticar os desafios encontrados na Ciência da Informação em relação à gestão dos dados científicos, por meio do ciclo de vida de dados proposto, possa despertar nos pesquisadores da área a urgência em trabalhar de forma mais detalhada esses elementos nos percursos metodológicos dos seus trabalhos científicos, de forma a contribuir com o acesso, compartilhamento e reúso dos dados. Dessa forma, esse estudo apresenta um impacto inovador, pelo ineditismo da proposta; um impacto econômico e social, por incentivar a reutilização e o acesso dos dados; e um impacto cultural, por propor mudanças nas práticas dos pesquisadores em relação ao uso e reúso dos seus dados científicos.

Palavras-chave: dados científicos; ciclos de vida dos dados; Ciência da informação; metodologia científica; *eScience*.

ABSTRACT

Actions associated with the processing of scientific data can be frequently observed in the methodological steps inherent to the processes of scientific investigations. In this context, the aim of this study is to analyze possible intersections between the components of data life cycles and the methodological paths presented in Brazilian theses, in the field of Information Science. The thesis argues that these elements associated with the data life cycles belonging to the initiatives of DataONE, Digital Curation Centre and Data Documentation Initiative are present in the methodologies of the Brazilian theses in that area. For this purpose, a reference model of the intersection of these initiatives was developed to carry out the study. The methodological path of the research has a quali-quantitative approach; characterized in terms of means as bibliographic and documentary; and as for the purposes as a descriptive study. The content analysis technique was used to analyze the data with the support of the QDAMiner software. The results showed that some Postgraduate Programs in Information Science have difficulties in giving access to their own theses and in identifying the methodologies in their respective works. It was verified that the area data are distinguished when the nature for being textual, visualizations, multimedia, numbers, and software, in addition, for presenting the registration, observational, computational, and experimental origins. Regarding the presence of data life cycle elements, it was found that some steps, such as planning, collecting, and analyzing, are addressed more clearly in the literature dealing with methodologies in social sciences. However, it was not possible to identify that the researchers adopted the data management plan when planning the study. Another issue identified is that despite the step describing having presented a notorious percentage of participation in the methodologies, it does not mean that there is a robust documentation to describe these data, since at that time the random excerpts that corresponded to some metadata elements were coded. The steps to organize and ensure quality were also identified in the work. The steps to find and reuse were seen in a very subtle way and the step to preserve was not evidenced. Data has shown that researchers carry out some actions related to the stages of cycles instinctively and, therefore, there is a need for better access to the results of studies in the area. It is hoped that this research, by diagnosing the challenges found in Information Science in relation to the management of scientific data, through the proposed data life cycle, may awaken in researchers in the field the urgency to work on these elements in more detail in the methodological paths of their scientific work to contribute to the access, sharing and reuse of data. Thus, this study has an innovative impact, due to the originality of the proposal; an economic and social impact, by encouraging the reuse and access of data; and a cultural impact, as it proposes changes in researchers' practices regarding the use and reuse of their scientific data.

Keywords: scientific data; data life cycle; Information science; scientific methodology; eScience.

LISTA DE FIGURAS

Figura 1 - Tipologia dos dados no tocante a natureza e a origem	30
Figura 2 - Dados da cauda longa da Ciência	37
Figura 3 - Ciclo de vida dos dados - DataONE	48
Figura 4 - Ciclo de vida dos dados do <i>Digital Curation Centre</i>	56
Figura 5 - Modelo de ciclo de vida dos dados do DDI	60
Figura 6 - Repositório Institucional da UnB/Ciência da Informação/Teses.....	84
Figura 7 - ECA USP/Ciência da Informação/Teses PPGCI	84
Figura 8 - UNESP/Programa de Pós- Graduação em Ciência da Informação/Teses	85
Figura 9 - Programa de Pós-Graduação em Ciência da Informação/UFBA/Trabalhos de Conclusão	86
Figura 10 - Repositório Institucional da Universidade Federal da Paraíba.....	86
Figura 11 - PPGCI/UFMG/Teses defendidas	87
Figura 12 - PPGGOC/UFMG/Teses defendidas	88
Figura 13 - PPGCin/UFSC/Teses e Dissertações	88
Figura 14 - PPGCI/IBICT/UFRJ/Teses de Doutorado	89
Figura 15 - PPGCI/ Teses defendidas.....	89
Figura 16 - Universidade/FUMEC	90
Figura 17 - Resumo dos principais pontos para atingir os resultados da pesquisa.....	93
Figura 18 - Teses disponíveis em acesso aberto do período de 2012 a 2020 por estado da federação.....	98
Figura 19 - Ciclo de vida dos dados científicos.....	112
Figura 20 - Representação dos códigos identificados nas metodologias da Ciência da Informação.....	133

LISTA DE QUADROS

Quadro 1 - Lista de periódicos de dados com os respectivos <i>links</i> de acesso	41
Quadro 2 - Iniciativas de modelos de Ciclos de Vida dos Dados.....	46
Quadro 3 - Ciclo de vida dos dados – DataONE	49
Quadro 4 - Descrição das fases do ciclo de vida de dados do DDI versão 3.0	60
Quadro 5 - Fases metodológicas dos dados e os ciclos de vida do DataONE, DCC e DDI.....	73
Quadro 6 - Programas de Pós-Graduação brasileiros com curso de doutorado na área da Ciência da Informação.....	79
Quadro 7 - Teses brasileiras da Ciência da Informação	81
Quadro 8 - Categorias e códigos referentes às teses brasileiras da Ciência da Informação	92
Quadro 9 - Descrição dos dados	95
Quadro 10 - Acesso às teses da Ciência da Informação no Brasil	100
Quadro 11 - Correlação do ciclo de vida de dados do DataONE com o DCC e o DDI	105
Quadro 12 - Correlação entre os ciclos de vida dos dados do DDI, DataONE e o DCC	108
Quadro 13 - Etapas sugeridas da interseção do DCC, DataONE e DDI	110
Quadro 14 - Detalhamento das etapas do ciclo de vida dos dados científicos	113
Quadro 15 - Dados textuais da Ciência da Informação no Brasil	117
Quadro 16 - Dados em visualizações da Ciência da Informação no Brasil.....	118
Quadro 17 - Dados numéricos da Ciência da Informação no Brasil	120
Quadro 18 - Dados de multimídia da Ciência da Informação no Brasil.....	121
Quadro 19 - Dados em <i>software</i> da Ciência da Informação no Brasil	122
Quadro 20 - Coocorência entre a origem dos dados e a preservação e curadoria das teses da CI	124
Quadro 21 - Dados de Registro da Ciência da Informação no Brasil.....	125
Quadro 22 - Coocorência entre a origem de Registro e a coleta de dados	126
Quadro 23 - Dados observacionais da Ciência da Informação no Brasil	128
Quadro 24 - Dados computacionais da Ciência da Informação no Brasil.....	129
Quadro 25 - Dados experimentais da Ciência da Informação no Brasil	130
Quadro 26 - Métodos e técnicas utilizados nas teses brasileiras da Ciência da Informação ..	136
Quadro 27 - Etapa planejar nas teses brasileiras da Ciência da Informação	137
Quadro 28 - Etapa coletar nas teses brasileiras da Ciência da Informação	140

Quadro 29 - Correlação entre os códigos da etapa coletar	142
Quadro 30 - Etapa organizar nas teses brasileiras da Ciência da Informação	142
Quadro 31 - Correlação entre as etapas organizar, descrever e analisar	144
Quadro 32 - Etapa descrever nas teses brasileiras da Ciência da Informação.....	145
Quadro 33 - Etapa assegurar a qualidade nas teses brasileiras da Ciência da Informação.....	147
Quadro 34 - Etapa descobrir nas teses brasileiras da Ciência da Informação	149
Quadro 35 - Etapa reutilizar nas teses brasileiras da Ciência da Informação.....	151
Quadro 36 - Etapa analisar nas teses brasileiras da Ciência da Informação.....	152

LISTA DE GRÁFICOS

Gráfico 1 – Quantitativo da amostra das teses dos programas de Pós-Graduação brasileiros da Ciência da Informação, do período de 2012 a 2020	83
Gráfico 2 - Localização das metodologias nas teses brasileiras da CI	103
Gráfico 3 - Natureza dos dados da Ciência da Informação no Brasil.....	116
Gráfico 4 - Distribuição da origem dos dados das teses da Ciência da Informação no Brasil	123
Gráfico 5 - Interseção entre o ciclo de vida dos dados de referência e as metodologias das teses da Ciência da Informação no Brasil.....	131
Gráfico 6 - Etapa planejar e as metodologias da ciência da informação no Brasil	134

LISTA DE ABREVIATURAS E SIGLAS

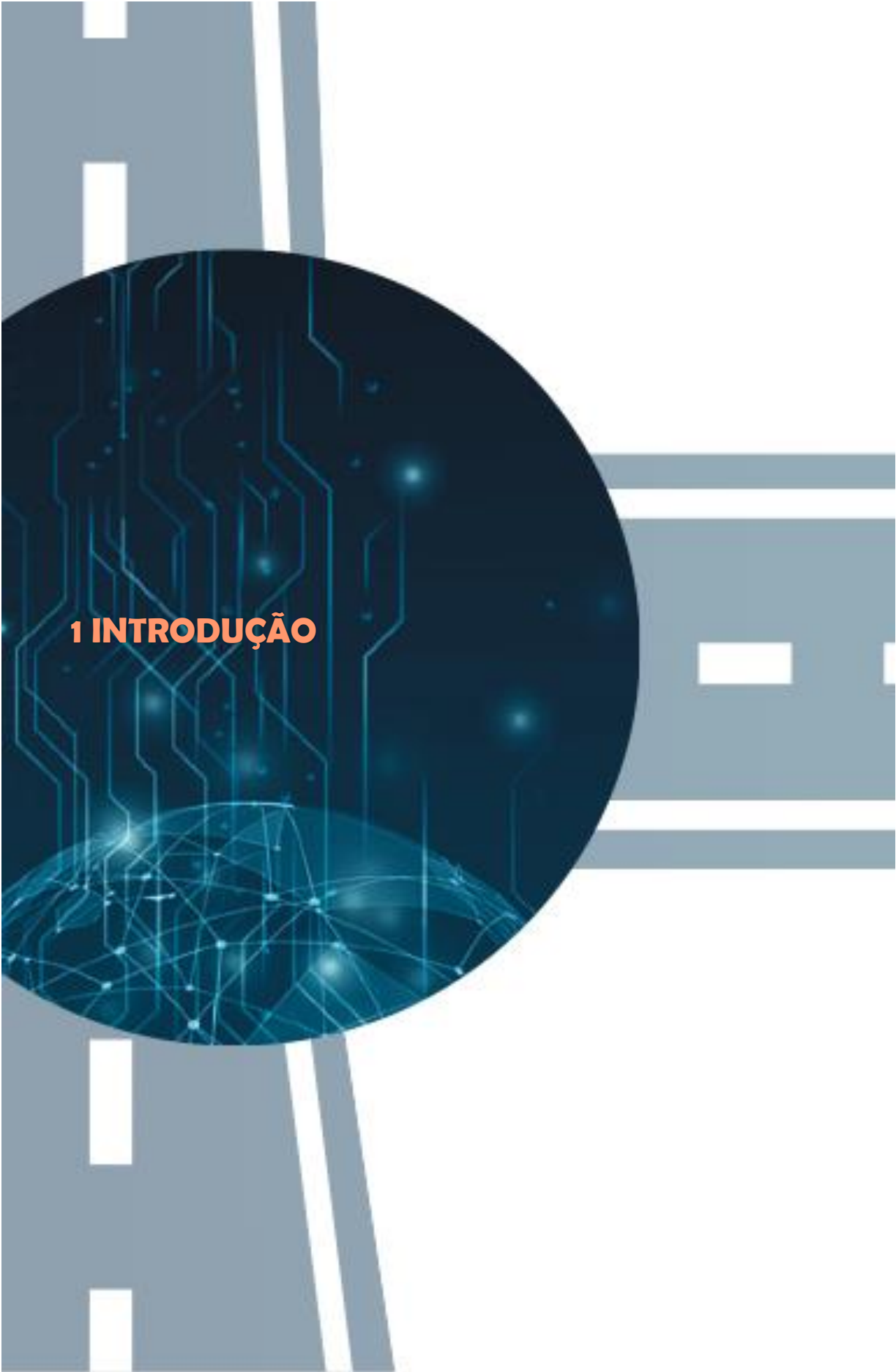
BENANCIB	Base dos Encontros Nacionais de Pesquisa e Pós-Graduação em Ciência da Informação
BDTD	Biblioteca Digital Brasileira de Teses e Dissertações
BRAPCI	Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CI	Ciência da Informação
CSV	<i>Comma-Separated-Values</i>
CVD	Ciclo de vida dos dados
CVD-CI	Ciclo de Vida dos Dados para a Ciência da Informação
DataONE	<i>Data Observation Network for Earth</i>
DCC	<i>Digital Curation Centre</i>
DDI	<i>Data Documentation Initiative</i>
DMPTool	<i>Management Plan Tool</i>
DTL	<i>Dutch Techcentre for the Life Science</i>
FAIR	<i>Findable, Accessible, Interoperable, Reusable</i>
FAPESP	FUNDAÇÃO DE AMPARO A PESQUISA DO ESTADO DE SÃO PAULO
FNS	<i>National Science Foundation</i>
FUMEC	Universidade FUMEC
IBGE	Instituto Brasileiro de Geografia e Estatística
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
LISA	<i>Library and Information Science Abstracts</i>
LISTA	<i>Library, Information Science & Technology Abstracts</i>
MIT	<i>Massachusetts Institute of Technology</i>
NIH	<i>National Institutes of Health</i>
NSB	<i>National Science Board</i>
OAIS	<i>Open Archival Information System</i>
OECD	<i>Organisation for Economic Co-operation and Development</i>
PDF	<i>Portable Document Format</i>
PGD	Plano de Gestão de Dados

PPGCI	Programa de Pós-Graduação em Ciência da Informação
PPGCin	Programa de Pós-Graduação em Ciência da Informação
PPGOC	Programa de Pós-Graduação Gestão e Organização do Conhecimento
PPGSIGC	Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento
RIDI	Repositório Institucional do Instituto Brasileiro de Informação em Ciência e Tecnologia
TCLE	Termo de Consentimento Livre e Esclarecido
TIC	Tecnologias da Informação e Comunicação
UCL	<i>University College London</i>
UEL	Universidade Estadual de Londrina
UFAL	Universidade Federal de Alagoas
UFBA	Universidade Federal da Bahia
UFC	Universidade Federal do Ceará
UFES	Universidade Federal do Espírito Santo
UFF	Universidade Federal Fluminense
UFMG	Universidade Federal de Minas Gerais
UFPA	Universidade Federal do Pará
UFPB	Universidade Federal da Paraíba
UFPE	Universidade Federal de Pernambuco
UFRJ	Universidade Federal do Rio de Janeiro
UFRN	Universidade Federal do Rio Grande do Norte
UFSC	Universidade Federal de Santa Catarina
UNB	Universidade de Brasília
UNESP	Universidade Estadual Paulista Júlio de Mesquita Filho
USP	Universidade de São Paulo
5Vs	Volume, velocidade, variedade, veracidade e valor
WoS	Web of Science
WRCO	<i>Web</i> , Representação do Conhecimento e Ontologias
WWW	<i>Word Wide Web</i>
XML	<i>Extensible Markup Language</i>

SUMÁRIO

1	INTRODUÇÃO	16
1.1	CONTEXTUALIZAÇÃO DO PROBLEMA DE PESQUISA	21
1.2	HIPÓTESE	22
1.3	OBJETIVOS	22
1.3.1	Objetivo Geral	23
1.3.2	Objetivos Específicos	23
1.4	JUSTIFICATIVA	23
1.5	ESTRUTURA DA TESE	25
2	DADOS CIENTÍFICOS	27
2.1	PUBLICAÇÃO DE DADOS	34
2.1.1	Repositório de Dados	38
2.1.2	Publicações Ampliadas	39
2.1.3	Periódicos de Dados	40
3	CICLO DE VIDA DOS DADOS	43
3.1	INICIATIVAS DE CICLOS DE VIDA DOS DADOS	46
3.1.1	Ciclo de Vida dos Dados – DataONE	48
3.1.2	Ciclo de Vida dos Dados – <i>Digital Curation Centre</i>	54
3.1.3	Ciclo de Vida dos Dados – <i>Data Documentation Initiative</i>	59
4	A METODOLOGIA NO CONTEXTO DOS TRABALHOS CIENTÍFICOS NA ÁREA DAS CIÊNCIAS SOCIAIS	64
4.1	A INVESTIGAÇÃO CIENTÍFICA E A PRESENÇA DOS DADOS NOS PERCURSOS METODOLÓGICOS	66
4.1.1	Descrição da Fase de Obtenção dos Dados	67
4.1.2	Descrição da Fase de Coleta dos Dados	71
4.1.3	Procedimentos de Análise dos Dados	72
5	PROCEDIMENTOS METODOLÓGICOS	76
5.1	TIPO E NATUREZA DA PESQUISA	77
5.2	<i>CORPUS</i> DA PESQUISA	79
5.3	TÉCNICAS E INSTRUMENTOS PARA A COLETA DE DADOS	83
5.4	MÉTODOS E TÉCNICAS DE ANÁLISE DE DADOS	91
5.5	CICLO DE VIDA DOS DADOS CIENTÍFICOS NO PRESENTE ESTUDO	94
6	ANÁLISE E DISCUSSÃO DOS RESULTADOS	97
6.1	A DESCOBERTA DAS TESES DA CIÊNCIA DA INFORMAÇÃO NO BRASIL	98
6.1.1	O Acesso às Teses Brasileiras da Ciência da Informação	100

6.1.2	A Descoberta das Metodologias nas Teses Brasileiras da Ciência da Informação ..	102
6.2	INTERSECÇÕES ENTRE OS MODELOS DE CICLOS DE VIDA DE DADOS DO DATAONE, DCC E DDI.....	104
6.2.1	Modelo de Referência da Intersecção do DCC, DataONE e DDI.....	110
6.3	A TIPOLOGIA DOS DADOS DAS TESES BRASILEIRAS DA CIÊNCIA DA INFORMAÇÃO, DE ACORDO COM A SUA NATUREZA E ORIGEM.....	115
6.3.1	Natureza dos Dados.....	115
6.3.2	Origem dos Dados.....	122
6.3.2.1	<i>Categoria de origem dos dados – Registro</i>	125
6.3.2.2	<i>Categoria de origem dos dados – Observacional</i>	127
6.3.2.3	<i>Categoria de origem dos dados – Computacional</i>	129
6.3.2.4	<i>Categoria de origem dos dados – Experimental</i>	130
6.4	INTERSECÇÕES ENTRE O CICLO DE VIDA DOS DADOS DE REFERÊNCIA E AS METODOLOGIAS DAS TESES BRASILEIRAS DA CIÊNCIA DA INFORMAÇÃO	131
6.4.1	Etapa “Planejar”	133
6.4.2	Etapa “Coletar”.....	139
6.4.3	Etapa “Organizar”	142
6.4.5	Etapa “Descrever”	144
6.4.6	Etapa “Assegurar a Qualidade”	146
6.4.7	Etapa “Preservar”	148
6.4.8	Etapa “Descobrir”	149
6.4.9	Etapa “Reutilizar”.....	150
6.4.10	Etapa “Analisar”.....	152
7	CONSIDERAÇÕES FINAIS	154
	APÊNDICE A - PLANO DE GESTÃO DE DADOS	169
	APÊNDICE B - LISTA CÓDIGO/REFERÊNCIA	174



1 INTRODUÇÃO

Em nossa contemporaneidade, há um discurso crescente sobre a relevância dos dados que compõem os fluxos intelectuais associados ao fazer científico. O valor dos dados está diretamente relacionado com o que pode ser obtido a partir deles, constituindo-se como uma fonte de conhecimento, quando tratados de forma adequada.

Os pesquisadores das diversas áreas do conhecimento, no exercício de suas atividades de pesquisa, lidam constantemente com dados científicos, de forma notória nos procedimentos metodológicos dos seus trabalhos de investigação, porém constitui um desafio significativo incorporar nas suas práticas de pesquisa as dinâmicas inerentes às novas provocações estabelecidas pela temática.

As inquietações que surgem neste contexto despertam o interesse de investigar ferramentas que auxiliem o pesquisador a fazer uma boa gestão dos seus dados científicos, assim como incentivam o anseio por disseminar a importância do tema, contribuindo para que o assunto seja conhecido especialmente na comunidade dos profissionais da informação e que se possa difundir diretamente a mensagem para outros pesquisadores, configurando-se, desta forma, como elementos fundamentais nesse processo.

A realidade mostra que a produção de dados nos meios acadêmicos, empresariais e governamentais atravessa um crescimento exponencial em termos de volume, variedade e velocidade, fato este denominado de megadados por Silva Segundo e Araújo (2019) ao remeterem à questão do *Big Data*. Para Moraes *et al.* (2018), tal fenômeno se refere a imensos conjuntos de dados que, por sua vez, demandam ferramentas adequadas para trabalhar com eles, possibilitando que em tempo hábil, a partir dos dados, as informações possam ser encontradas, analisadas e aproveitadas.

Prá Netto, Moro e Ferreira (2020) apresentam também a veracidade e o valor para compor o conceito do *Big Data* (megadados). Para os autores, a veracidade está relacionada às análises dos dados condizerem com a realidade do momento e o valor remete a empenhar esforços na coleta e análise de dados significantes. Assim sendo, atualmente na literatura o fenômeno em questão envolve 5 (cinco) Vs: volume, variedade, velocidade, veracidade e valor (PRÁ NETTO; MORO; FERREIRA, 2020).

No âmbito dos dados científicos, as formas tradicionais de fazer ciência estão permeadas por transformações significativas, seja nas ciências naturais, ciências sociais ou nas ciências humanas (JANKOWSKI, 2007). Estas transformações são causadas pela tendência de crescimento da disponibilidade dos dados. Por isso é fundamental o entendimento sobre as mudanças que estão acontecendo no processo de desenvolvimento científico e tecnológico, assim como compreender que os aspectos conceituais que surgem nesse contexto remetem a

um novo paradigma de se fazer ciência, denominado de *eScience*, cuja base está na computação intensiva de dados (COSTA, 2017). Os paradigmas científicos anteriores como: a ciência experimental, a ciência teórica e a ciência baseada em simulação são reconfigurados na literatura pela *eScience*, considerado como o quarto paradigma da ciência, diante de um cenário rico em dados (GRAY, 2009; SAYÃO; SALES, 2019). A ciência como um todo está se modificando em razão do impacto da tecnologia da informação, e estão sendo influenciadas pelo dilúvio de dados que está atrelado à emergência do quarto paradigma científico. Neste contexto, a finalidade é que a literatura científica e os dados científicos estejam *online* e que seja possível a interoperabilidade entre eles (GRAY, 2009).

Para Gray (2009), a *eScience* caracteriza-se por dados capturados através de instrumentos ou gerados por meio de simuladores, processados por produtos de *software*, cujos resultados são armazenados em computador. Neste processo, os cientistas analisam as bases de dados e arquivos gerados através das ferramentas de gestão e estatística.

O momento que se descortina apresenta consideráveis desafios de ordem tecnológica para possibilitar que os dados estejam acessíveis e, conseqüentemente, contribuam não apenas para aumentar a velocidade de disponibilização da informação científica, mas também a sua qualidade, oportunizando florescerem novas descobertas que atendam às necessidades da sociedade.

As práticas acadêmicas são direcionadas para a ciência aberta, que também é um fenômeno que faz parte desse momento, sendo um movimento que permite um maior acesso aos resultados das pesquisas científicas. Inclui vários aspectos como a publicação aberta, dados abertos, *software* de código aberto, registros científicos abertos, revisão por pares aberta, disseminação aberta e materiais abertos (BEZJAK *et al.*, 2018). Para Vicente-Saez e Martinez-Fluentes (2018), trata-se de um conhecimento cujo desenvolvimento ocorre por meio de redes colaborativas, sendo ele transparente, acessível e compartilhado – o que demonstra um ambiente favorável ao compartilhamento de dados, pois é essencial manter os dados científicos disponíveis.

Todo este cenário associado ao processo de tornar os dados científicos reutilizáveis requer um sistema de gestão que permita uma adequada seleção, tratamento, representação, recuperação e reúso dos dados. Sales (2020) mostra que o processo de gestão dos dados perpassa várias instâncias, envolve a infraestrutura tecnológica, a sustentabilidade dos dados, questões políticas e as pessoas que estão envolvidas no ecossistema associado ao tratamento dos dados.

A emergência desta realidade envolve reflexões sobre o termo curadoria digital que, segundo Sayão e Sales (2012), está relacionado com a garantia da preservação dos dados para utilização futura, além do valor imediato para os seus criadores e usuários. Esta situação é possível por meio de recursos políticos, administrativos, estratégicos metodológicos e das tecnologias utilizadas no processo de curadoria digital, que contribuem com o acesso persistente a dados digitais com credibilidade através do aprimoramento da qualidade dos dados, do seu contexto e da verificação de autenticidade.

Neste sentido, Borgman (2015) revela que os dados não possuem significado isoladamente. Eles existem dentro de uma infraestrutura de conhecimento, de uma ecologia de pessoas, práticas, instituições, objetos materiais e relacionamentos. O assunto é relevante não apenas para os pesquisadores atuais e do futuro, mas também para aqueles que irão usar o conhecimento criado por eles. O valor dos dados reside dentro de um contexto que precisa estar registrado por meio da sua adequada preservação e gestão.

A forma como esses dados são operacionalizados impactam diretamente nos processos associados à geração dos estoques informacionais e posterior apropriação pelos seres humanos (BARRETO, 2002). Apesar disso, muitos pesquisadores ainda obtêm os seus dados apenas para chegar aos resultados dos problemas de pesquisa particulares e, posteriormente, promovem o descarte de forma "consciente" ou não, levando ao desperdício, não sendo possível, assim, compartilhá-los futuramente com outros pesquisadores.

Os dados inacessíveis limitam os impactos positivos na ciência, provocam o retrabalho e o atraso de novas descobertas, deixando de atender de forma proativa às urgências de problemáticas de pesquisas que poderão surgir e, conseqüentemente, corre-se o risco de não ter ao dispor conjuntos de dados que poderiam trazer respostas científicas pontuais. Assim, o prejuízo de manter os dados “engavetados” pode ser incalculável.

Existem iniciativas catalisadoras de mudanças, tais como as novas formas de comunicar a ciência apresentadas por Curty e Aventurier (2017), ao mostrarem três abordagens sobre a publicação de dados: repositório de dados, publicações ampliadas e artigos de dados. Neste âmbito, já existem periódicos que publicam dados científicos, a exemplo da revista de dados *Scientific Data* (2020), da *Springer Nature* (2020), e do periódico brasileiro intitulado *Encontros Bibli* (2020), que também passou a adotar a publicação de artigos de dados (*data papers*). Assim sendo, um novo cenário se descortina no horizonte da pesquisa científica.

Diante do surgimento de um contexto inovador voltado para gestão de dados como requisito fundamental para as boas práticas de pesquisa, surgiram algumas iniciativas que tratam do ciclo de vida dos dados científicos, com o objetivo de proporcionar uma curadoria

adequada para eles. No cenário internacional, destacamos o *Digital Curation Centre – DCC*, que apresenta a *Curation Lifecycle Model* (DCC, 2020); o *Data Observation Network for Earth – DataONE*, que propôs o *Data Life Cycle* (DATAONE, 2020); e o *Data Documentation Initiative* (DDI), que oferece o *DDI-Lifecycle* (DDI, 2020). Sayão e Sales (2015), ao tratarem da temática da gestão de dados científicos, no que tange às operações que demandam uma estrutura adequada para o registro dos dados no decorrer da sua vida, de forma que eles possam ter o uso otimizado por longo prazo, revelam que os ciclos de vida dos dados das iniciativas do DCC, do DataONE e do DDI, se tornaram referência para pesquisadores, bibliotecários e gestores de dados.

Para Sanchez, Vidotti e Vechiato (2017), as funções encontradas no modelo de ciclo de vida de dados do DCC apresentam-se como uma solução primordial para que ocorra a preservação digital. Os autores entendem que a iniciativa tem como foco dados, objetos digitais e base de dados, porém é predominantemente mencionado na literatura como processos de curadoria voltados aos dados científicos, o que contribui para assegurar confiabilidade a esses dados.

De acordo com Anjos e Dias (2019), grande parte das competências dos profissionais da informação estão alinhadas com as etapas do ciclo de vida dos dados do DataONE, podendo esses profissionais atuarem na gestão de dados científicos por meio desta ferramenta. Para corroborar com esta realidade, Sayão e Sales (2015), ao elaborarem um Guia intitulado de “Gestão de dados de pesquisa para bibliotecários e pesquisadores”, utilizaram o modelo DataONE como exemplo por se alinhar com os objetivos do documento.

No que se refere à iniciativa do DDI, Vardigan, Heus e Thomas (2008) mostram que o modelo possibilita a descrição do ciclo de vida dos dados e aumenta o número de metadados que acompanham os dados para ampliar o contexto. Esta característica favorece a validação da qualidade dos dados. Neste cenário, o *UK Data Service* (Serviço de dados do Reino Unido), que tem ao longo do tempo desenvolvido procedimentos e protocolos para a gestão de dados, documentação e publicação utiliza o padrão de metadados do DDI, que é voltado para as ciências sociais (VAN DEN EYNDEN; CORTI, 2017). Diante da relevância dos três ciclos de vida dos dados supracitados para a gestão dos dados científicos, eles foram adotados como referência para realizar esta pesquisa.

Vale ainda destacar que, no contexto brasileiro, evidenciamos o modelo de Sant’Ana (2016), denominado de Ciclo de Vida dos Dados para a Ciência da Informação (CVD-CI), desenvolvido no âmbito do Programa de Ciência da Informação da Universidade Estadual Paulista (UNESP).

A área da Ciência da Informação (CI) apresenta-se como forte aliada no processo de lidar adequadamente com os dados provenientes das pesquisas. De acordo com Sant’Ana (2016), a CI apresenta técnicas que possibilitam um melhor acesso e uso dos dados. Neste contexto, supõe-se que os pesquisadores desta área do conhecimento, na construção dos procedimentos metodológicos de suas teses, explicitam a forma de captura e tratamento dos seus dados, seja de cunho qualitativo ou quantitativo. Logo, indica-se a possibilidade de que as etapas associadas ao tratamento desses dados estejam de alguma forma representadas nos ciclos de vida dos dados científicos. Este estudo ora proposto tem interesse em desvendar as dinâmicas associadas a esta situação.

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA DE PESQUISA

No cenário internacional, mais notadamente no Reino Unido e nos Estados Unidos, as instituições de fomento à pesquisa solicitam que os pesquisadores, quando da solicitação de recursos financeiros para a condução dos seus estudos científicos, apresentem estratégias para a gestão dos seus conjuntos de dados. Segundo Zimmerman *et al.* (2009), um plano de compartilhamento dos dados começou a ser exigido pelo *National Institutes of Health (NIH)* em 2005 para as propostas que recebiam recursos públicos. Esta prática já chegou ao Brasil e tende a se tornar cada vez mais comum, não obstante ainda seja incipiente. A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), que financia a pesquisa científica, no contexto deste estado, exige que para os projetos temáticos, o pesquisador inclua um documento indicando como será feita a gestão dos dados ao longo de sua pesquisa (FAPESP, 2020).

Sales (2020) explica que a gestão de dados científicos, assim como o trabalho do bibliotecário estão presentes em todas as etapas da investigação científica, a partir do planejamento do projeto de pesquisa e com a elaboração do plano de gestão de dados. A gestão continua presente no decorrer da pesquisa, no momento que ocorre a coleta, o processamento, a análise, o controle de qualidade e representação dos dados por meio de metadados, assim como no armazenamento seguro. Perpassa ainda a ocasião que a pesquisa é finalizada quando os dados são catalogados, contextualizados, no momento que ocorrem as decisões referentes a publicações, preservação de longo prazo e sobre compartilhamento, acesso e reúso.

Para dar suporte ao processo de gestão de dados, é preciso um arcabouço de diversas facetas que envolvam todo o ciclo de vida dos dados (CVD) (SAYÃO; SALES, 2018). Segundo Lenhardt *et al.* (2014), o objetivo da ferramenta do CVD, que dá suporte à descoberta científica, é assegurar que os dados científicos sejam obtidos com rigor de modo a atender às necessidades

do uso pretendido, com o propósito de suportar o gerenciamento de dados, possibilitar a reutilização e redirecionamento dos dados e permitir a preservação a longo prazo.

A temática referente à gestão dos dados científicos se configura como uma prática nova no Brasil (DIAS; ANJOS; RODRIGUES, 2019), apesar dessa realidade, é iminente considerar a necessidade de realizar a gestão dos científicos por meio do ciclo de vida dos dados mais adequados à realidade concreta. Desta forma, supõe-se que muitas das práticas relacionadas aos elementos componentes dos ciclos de vida dos dados já são feitas intrinsecamente pelos pesquisadores nos seus projetos de pesquisa, nas dissertações de Mestrado e nas teses de Doutorado, de forma mais evidente nas estratégias metodológicas, uma vez que este momento da pesquisa lida diretamente com os dados, seja ao planejar a sua obtenção, na coleta propriamente dita, e nas decisões de técnicas e métodos para a sua análise.

A partir dessas considerações iniciais relacionadas aos ciclos de vida dos dados científicos e aos percursos metodológicos, indagamos: **de que forma os elementos componentes dos ciclos de vida dos dados emergem nos percursos metodológicos das teses brasileiras, na área da Ciência da Informação?**

1.2 HIPÓTESE

Os procedimentos metodológicos das investigações científicas comumente apresentam a forma como os dados são gerados, tratados e analisados, ou seja, apresentam o fluxo dos dados nas pesquisas. Levando em consideração que a gestão dos dados científicos e o CVD estão presentes nos debates da área da CI, foi elaborada a seguinte hipótese: Elementos das etapas associadas aos ciclos de vida dos dados pertencentes às iniciativas do DataONE, do DCC e da DDI podem estar presentes nos percursos metodológicos das teses brasileiras, na área da Ciência da Informação. A tese defende que há a presença dos elementos dessas iniciativas nas referidas metodologias.

1.3 OBJETIVOS

Nesta subseção, são apresentados os objetivos de estudo, divididos em objetivo geral e objetivos específicos.

1.3.1 Objetivo Geral

A pesquisa tem como objetivo geral analisar possíveis intersecções entre os elementos componentes dos ciclos de vida dos dados e os percursos metodológicos apresentados nas teses brasileiras, na área da Ciência da Informação.

1.3.2 Objetivos Específicos

Em consonância com o objetivo geral proposto nesta pesquisa, delineou-se os seguintes objetivos específicos:

- a) levantar as teses disponíveis dos Programas de Pós-Graduação em Ciência da Informação brasileiros;
- b) identificar etapas que podem emergir da intersecção entre os ciclos de vida dos dados do DataONE, do DCC e o DDI;
- c) propor um modelo de referência para descrição dos dados científicos dentro do contexto de metodologias das teses da Ciência da Informação, tendo como parâmetro a análise baseada nesses três modelos;
- d) classificar a tipologia dos dados usados nas teses brasileiras, na área da Ciência da Informação, de acordo com a sua natureza e origem.

1.4 JUSTIFICATIVA

Na ciência da atualidade, a importância do tema compartilhamento de dados científicos é ancorada no seu potencial de reúso. Este, por sua vez, envolve a identificação e acesso aos dados preexistentes que estão de acordo com os interesses dos pesquisadores e requer dados que estejam alinhados com condições mínimas de reusabilidade (CURTY, 2019).

Esse cenário remete a trazer novos usos aos dados científicos pelos próprios autores, como por outros pesquisadores no futuro, ao oportunizar respostas às novas problemáticas de pesquisas. Assim, justifica-se a importância de estudar ferramentas que não apenas reflitam, mas que tragam uma aplicabilidade à curadoria dos dados e, conseqüentemente, motivem os pesquisadores a disponibilizarem os seus dados, sustentados em uma adequada gestão ao longo do CVD.

Para Medeiros (2018), a gestão dos dados científicos envolve todo o ciclo de vida dos dados, da fase de coleta até o armazenamento de longo prazo, perpassando alguns momentos

como processos de limpeza, curadoria, anotação, indexação e transformação. Para a autora, a investigação científica da atualidade requer comumente algum tipo de análise e processamento de dados. Desta forma, o planejamento da gestão dos dados usados e gerados em uma pesquisa passou a integrar a metodologia científica, chegando a ser tido como um dos itens fundamentais às boas práticas de pesquisa.

Diante da realidade, os pesquisadores dos diversos domínios do conhecimento, possivelmente, já realizam ações direcionadas ao CVD em seus percursos metodológicos. No tocante à área da CI, é possível que exista alguma evidência da necessidade de documentar melhor os dados. Desta forma, o presente estudo se justifica pela importância de visualizar se há evidências da ausência de previsão nas metodologias sobre o tratamento, preservação e compartilhamento dos dados.

Por isso, é fundamental saber sobre os tipos de dados utilizados na CI, para adotar medidas de preservação adequadas. Além disso, o presente estudo permitirá identificar se as teses da área vêm sendo disponibilizadas e se há dificuldade de encontrá-las. É provável que haja indícios no atraso do compartilhamento desses próprios trabalhos, sendo difícil chegar até eles, quanto mais descobrir os seus dados.

Este estudo visa contribuir com o esclarecimento sobre como os pesquisadores brasileiros da CI, no contexto das teses, estão fazendo a gestão dos dados científicos. A partir deste diagnóstico, alguns caminhos são apontados, com base na literatura, de forma que a área possa se fortalecer nesse campo e disseminar a temática para pesquisadores de outras áreas do conhecimento. Desta forma, apresenta-se um modelo de ciclo de vida dos dados científicos, baseado nas iniciativas do DataONE, DCC e DDI, que pode ser utilizado como referência para os pesquisadores da área realizarem nos procedimentos metodológicos dos seus trabalhos de pesquisa a gestão dos dados.

No que se refere aos impactos desta pesquisa, ela tem um caráter inovador, pois através da revisão da literatura não foram encontrados trabalhos na área que abordassem a mesma temática. Apresenta um impacto econômico e social, pois busca trazer a reflexão sobre o investimento eficiente em pesquisa, quanto a evitar o desperdício econômico, de tempo e trabalho na obtenção de dados que já foram alcançados e que podem ser reaproveitados em novos estudos. A busca por essa conscientização ocorre principalmente porque muitas pesquisas são financiadas por meio de recursos públicos e, por conseguinte, os resultados devem ser acessíveis aos usuários interessados.

Por ser um tema ainda novo no Brasil, esta pesquisa também remete a um impacto cultural, pois ensina os pesquisadores a refletirem e incorporarem novas práticas, ao lidarem

com os dados obtidos em suas pesquisas. Essa mudança pode gerar resistência inicial, por demandar esforços extras e, por isso, os pesquisadores precisam ser bem esclarecidos sobre os benefícios da curadoria, bem como ser capacitados para tal.

A opção pela temática se deu pela experiência acadêmica e científica como pesquisadora, membro do grupo de pesquisa cadastrado no CNPq, intitulado de *Web, Representação do Conhecimento e Ontologias (WRCO)*, que tem como escopo de pesquisa a representação do conhecimento e o estudo dos dados. Desta forma, pela compreensão da importância de fazer o uso dos dados, não apenas para atender à necessidade do momento de uma determinada pesquisa, há o entendimento de que eles podem enriquecer outras investigações científicas e trazer maior credibilidade e compreensão ao estudo em questão.

A familiaridade com a temática também parte da percepção particular da relevância desta para o avanço da ciência, de forma que evita o retrabalho da obtenção de dados e a otimização do seu uso e acesso, além do entendimento de que a Ciência da Informação pode ser uma área que abre caminhos para esse processo e, portanto, pode ser uma referência nesse assunto para a comunidade científica.

1.5 ESTRUTURA DA TESE

O trabalho está estruturado em 7 (sete) seções. A primeira seção, constituída pela introdução, apresenta o contexto do fenômeno do *Big Data*, o quarto paradigma da ciência (*eScience*), o movimento da ciência aberta e as iniciativas dos ciclos de vida dos dados. Essa seção ainda está subdividida pela problemática da pesquisa, hipótese levantada, pelos objetivos do trabalho, pela justificativa e a estrutura da tese.

Na segunda seção, são abordadas as definições dos dados no contexto científico, principalmente a tipologia dos dados. Nesse momento, também são discutidas as alternativas para a publicação e compartilhamento dos dados, tais como: repositório de dados, publicações ampliadas e periódicos de dados.

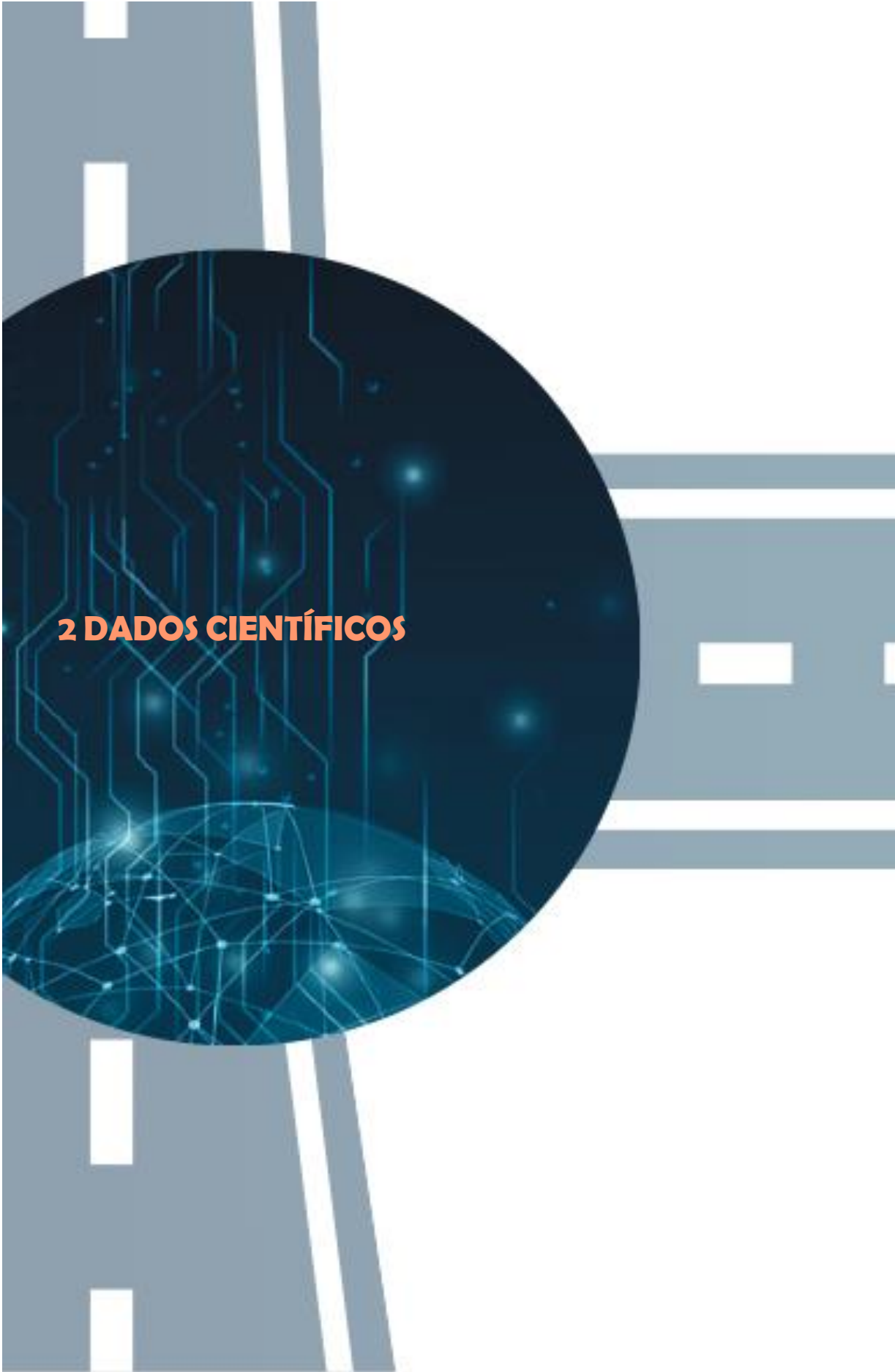
A temática acerca dos ciclos de vida dos dados é discutida na terceira seção. A princípio, são apresentados conceitos sobre o tema e enfatizada a importância do bibliotecário neste contexto. Na sequência, os CVDs das iniciativas do DataONE, DCC e DDI são discutidos individualmente, de forma detalhada.

Na quarta seção, o tema sobre a metodologia é contextualizado no âmbito das ciências sociais. Neste capítulo, são apresentados os conceitos de metodologia e são destacadas as fases metodológicas dos dados encontradas na literatura.

Os procedimentos metodológicos das teses estão presentes na quinta seção, em que são apresentados o tipo e natureza da pesquisa, o *corpus* da pesquisa, as técnicas para a coleta e as técnicas e métodos de análise dos dados. Neste capítulo, também foi incluída uma síntese do caminho metodológico dos dados científicos obtidos no trabalho.

Na sexta seção, estão evidenciadas as análises e discussões dos resultados. A princípio, apresenta-se a descoberta das teses da Ciência da Informação no Brasil. Em seguida, enfatiza-se as intersecções entre as iniciativas dos ciclos de vida dos dados do DataONE, DCC e DDI e o desenvolvimento de um novo modelo de CVD. Neste mesmo capítulo, são discutidas as tipologias dos dados encontradas nas teses da CI, no tocante à origem e à natureza dos dados e, por fim, foram levantadas as intersecções entre os elementos do novo CVD com as metodologias da CI.

As considerações finais encontram-se na sétima seção e remetem à problemática da pesquisa, à hipótese e às discussões e análises dos resultados, com a finalidade de fechar a discussão da tese, com as sugestões de pesquisas futuras que podem ser desenvolvidas a partir do estudo em questão.



2 DADOS CIENTÍFICOS

A temática associada à questão dos dados científicos, objeto do presente estudo, avança na literatura. A ciência na atualidade tem colocado os dados em um lugar de destaque, considerando-os não como mero subproduto das atividades científicas, mas como recurso valioso e fonte primária de novos saberes (SAYÃO; SALES, 2018).

Silva (2019) esclarece sobre o que torna os dados tão valiosos ao revelar que o conhecimento é o motor do avanço científico, ao passo que os dados correspondem ao seu combustível. Porém o interesse pelos dados é recente. Essa percepção de valor vem sendo obtida ao longo do tempo e continua a crescer em todo âmbito científico. Entre 1950 e 1960, o entendimento dos dados era limitado a ser um resultado de uma pesquisa, posteriormente passaram a ser percebidos como subprodutos. Só a partir do período de 1990 a 2010 a reutilização de dados ganhou importância para a promoção do desenvolvimento de novos estudos. A partir de então os dados passaram a ser considerados como substrato, ou seja, elemento primordial para o avanço da ciência (SILVA, 2019).

Curty e Aventurier (2017) destacam que os dados de pesquisa são matérias-primas fundamentais para a ecologia da ciência e são indispensáveis para a criação de novos ciclos de conhecimento científico, uma vez que são fontes de insumos para um processo iterativo no ciclo de vida da pesquisa, possibilitando que avance a descoberta científica e o desenvolvimento tecnológico.

A Agência USP de Gestão da Informação Acadêmica da Universidade de São Paulo (AGUIA) (2020), ao conceituar dados de pesquisa, revela que compreendem elementos cruciais pertencentes ao processo de pesquisa. São registros científicos que fundamentam os resultados de pesquisa publicados na forma de dissertações, teses, artigos, patentes e trabalhos científicos.

Assante *et al.* (2016) mostram que não existe uma definição partilhada sobre dados de pesquisa. O termo geralmente remete aos diversos materiais produzidos nas atividades desenvolvidas no decorrer da pesquisa. Neste cenário, é preciso compreender que o conceito de dados é complexo, ele pode se apresentar de múltiplas formas. Pode variar entre os pesquisadores e de modo mais acentuado entre os diversos domínios do conhecimento (SAYÃO; SALES, 2015).

É possível visualizar entre as definições que mesmo não existindo um conceito único para os dados de pesquisa, os autores são unânimes em considerar que o termo é intimamente ligado à produção científica. Os dados compreendem o ponto de partida para a geração de conhecimento e sem eles não é possível legitimar os resultados das investigações científicas.

Silva (2019) esclarece que os dados científicos podem possuir diversas definições por envolverem todos os domínios do conhecimento. Porém, de modo geral, o autor define dados

científicos como informações registradas ou produzidas por intermédio de alguma forma ou meio, no decorrer de uma pesquisa, e acrescenta que eles também compreendem as evidências necessárias ao pesquisador para validar suas conclusões depois de ter realizado uma pesquisa.

O relatório do National Science Board (NSB) de 2005, intitulado *Long-Lived Digital Data Collections*, ao se referir ao termo “dados” mostra que ele pode se distinguir em relação a sua natureza, os dados podem ser textos, números, imagens, vídeos, áudios, *software*, algoritmos, equações, animações, modelos, simulações. A *Organisation for economic co-operation and development* – Organização para cooperação e desenvolvimento econômico (OECD) (2007, p.13, tradução nossa) corrobora com a definição ao conceituar dados de pesquisa como: “[...] registros factuais (pontuações numéricas, registros textuais, imagens e sons) usados como fontes primárias de pesquisa científica e que são comumente aceitos na comunidade científica como necessários para validar resultados da pesquisa.”¹.

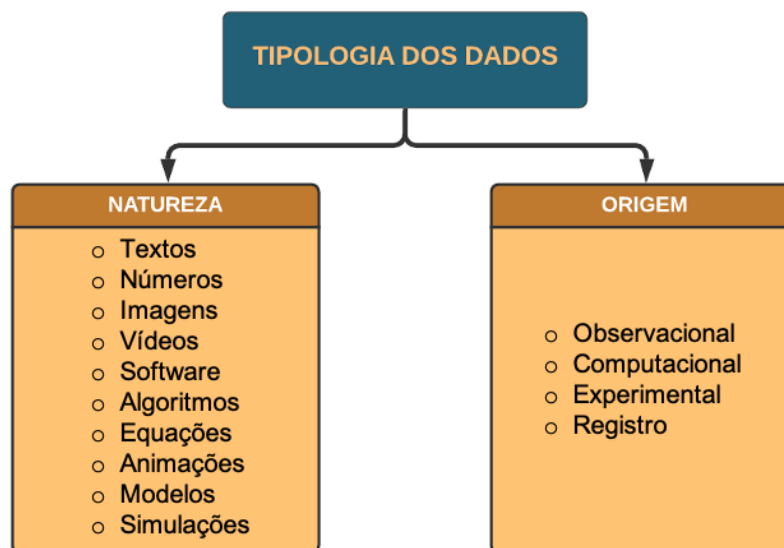
Sales e Sayão (2019) realizaram um estudo empírico voltado à área multidisciplinar das ciências nucleares. Nesse contexto, os autores agruparam os tipos de dados com relação a natureza em alguns grupos diversificados, tais como: números, multimídias, *software*, visualizações, textuais, artefatos, processos e outros. Para os autores, os dados numéricos corresponderam a medidas, resultados de levantamentos, resultados de experimentos, fórmulas, equações e algoritmos. Os dados multimídia se referiram a imagens, vídeos, fotografias, áudios, animações e filmes. No que se tange aos dados, como *software*, foram considerados bases de dados, simulações e códigos nucleares. No grupo das visualizações, foram incluídos tabelas, diagramas, modelos em 3D, modelos reduzidos e desenhos. Os dados textuais corresponderam aos metadados, questionários, entrevistas anotações, normas, padrões, certificados, caderno de laboratório, transcrição, correspondência, diário e caderno de campo. No grupo de artefatos foram considerados os espécimes, amostras e maquete. Os dados como processos referiram-se aos procedimentos operacionais padronizados, *workflows*, protocolos e testes. Por fim, o grupo outros foi relacionado a phantom ou manequim.

O relatório do NSB (2005) mostra que os dados podem se diferenciar conforme a sua origem, a saber: observacional, computacional e experimental. Borgman (2015) acrescenta uma quarta categoria de origem dos dados, a de registros. A compreensão sobre esta diferença é fundamental para as decisões a respeito do arquivamento e preservação. Para Sayão e Sales (2015), os tipos de dados podem ser caracterizados por sua natureza, origem ou *status* no fluxo

¹[...] *factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings.*

de trabalho da pesquisa. Para fins desta tese, o foco será na origem e natureza dos dados, conforme ilustrado na Figura 1.

Figura 1 - Tipologia dos dados no tocante a natureza e a origem



Fonte: Dados da pesquisa (2021)

Segundo o relatório do NSB (2005), os dados de origem observacional são registros históricos que não podem ser replicados, logo podem necessitar de um arquivamento indefinido. A proveniência desses dados remete a observações de fenômenos e eventos que só podem acontecer uma vez, e, portanto, necessitam ser arquivados de forma fidedigna, autêntica e íntegra em repositórios confiáveis que assegurem uma preservação adequada no longo prazo (SALES; SAYÃO, 2019). Neste sentido, para Borgman (2015), os dados observacionais são considerados os mais importantes para preservar, porque são os menos reproduzíveis. Silva (2019) exemplifica que em ciências sociais as observações podem remeter ao comportamento de diversos grupos ou indivíduos. Assim sendo, pela razão que os dados desse tipo de origem não podem ser obtidos mais de uma vez, é fundamental aprofundar as medidas de curadoria para posterior reutilização desses dados.

O relatório NSB (2005) revela que os dados computacionais são resultados da execução de um modelo ou simulação de computador (inclui uma documentação robusta do *hardware*, *software* e dados de entrada). A preservação dos dados de saída pode não ser essencial, pois os dados podem ser reproduzidos, porém é necessário o arquivamento do próprio modelo e de um conjunto completo de metadados sobre ele. Borgman (2015) complementa que comumente eles

são encontrados nas ciências físicas e da vida, porém também podem ser encontrados nas ciências sociais e humanas.

No que se refere aos dados experimentais (exemplo das taxas de reações químicas), segundo o relatório NSB (2005), pode não ser possível reproduzir de forma precisa todas as condições experimentais, quando algumas variáveis não são conhecidas e quando os custos da reprodução do experimento são proibitivos. Desta forma, as questões de custo e reprodutibilidade são essenciais para a implementação de políticas para preservação de dados experimentais.

Borgman (2015) enfatiza que a categoria de registro envolve formas de dados que não se enquadrem facilmente nas categorias de observação, experimentação e computação ou que sejam resultado de qualquer uma destas categorias. Para a autora, registros de qualquer fenômeno ou atividade humana podem ser tratados como dados para a pesquisa. Eles podem incluir a documentação de governos, empresas, públicas e atividades privadas; livros e outros textos; materiais de arquivo; documentação na forma de áudio e gravações de vídeo, placas de vidro, papiros, cuneiformes, bambu, e assim por diante.

O relatório NSB (2005) ainda destaca os níveis de dados dentro de cada categoria de origem dos dados, ao falar dos dados brutos e dos dados derivados. Assim, os dados podem ser coletados de forma bruta, a partir de então, sua forma passa a ser refinada durante versões sucessivas. Apesar dos dados brutos serem os mais completos, os dados derivados são os mais facilmente usados por outros, por já terem recebido um tratamento.

Silva (2019), ao discutir sobre os dados primários, secundários e terciários mostra que os dados primários correspondem a trabalhos originais de pesquisa e/ou dados brutos ausentes de interpretação. Os dados secundários estão disponibilizados para serem acessados porque foram coletados, tabulados, ordenados e colocados à disposição das pessoas. Os dados terciários constituem-se como uma forma de dados derivados como recontagens, categorias e resultados de dados estatísticos. Desta forma, a obtenção dos dados primários está estritamente relacionada com o objetivo da pesquisa. Os dados secundários são disponibilizados publicamente, normalmente em repositórios, bases de dados estatísticos, dentre outras publicações. É comum os pesquisadores fazerem a combinação dos dados primários, secundários e terciários para a produção de dados derivados com mais valor.

Os dados científicos compõem o processo dinâmico da investigação científica. Neste contexto, Schöpfel *et al.* (2016) apresentam duas funções dos dados: **Dados como materiais (insumos)** em um primeiro estágio do processo de pesquisa – os dados são coletados e analisados de diversificadas fontes e de distintas formas e formatos; **Dados como resultado**

(saída) – trata-se dos dados produzidos no decorrer do processo de pesquisa e no final, juntamente com publicações, como resultados de pesquisa. Os autores complementam que os dados brutos, que são os de entrada, normalmente são transformados e processados em produtos (métricas, gráficos, etc.). Ao passo que os dados provenientes da própria pesquisa, especialmente os de saída, seguem sua própria dinâmica, o que pode ser materializado em um ciclo de vida dos dados, conforme veremos adiante.

Um fato que chama atenção na literatura é que frequentemente o termo “dados de pesquisa” está associado ao termo “dados científicos”, eles são muitas vezes empregados como sinônimos, pois em um mesmo trabalho acadêmico/científico os pesquisadores usam as duas terminologias para o mesmo propósito, a exemplo da investigação realizada por Curty e Aventurier (2017). Na obra intitulada de “Dados científicos: perspectivas e desafios”, organizada pelos pesquisadores Dias e Oliveira (2019), cujo documento reúne textos de diversos autores, alguns destes como Sayão e Sales (2019) utilizam a expressão dados de pesquisa ao passo que outros, como Dias, Anjos e Rodrigues (2019), usam o termo dados científicos. Segundo Costa (2017), ainda não há na literatura um consenso sobre a utilização das expressões “dados científicos” e “dados de pesquisa”. Apesar do presente trabalho não se configurar como um estudo terminológico, considera-se pertinente a reflexão de que, embora os significados das terminologias sejam parecidos, podem apresentar diferenças sutis.

Compreendemos que as definições supracitadas referentes a dados de pesquisa sejam correspondentes a dados científicos, que são dados cujo processo de coleta, geração, obtenção, interpretação ou análise requer a adoção de uma metodologia que inclua habilidades acadêmicas dos pesquisadores, métodos, técnicas e procedimentos aprovados pela ciência, ou seja, que perpassa o rigor científico. Esse caminho da busca da verdade pode ocorrer na construção de um trabalho de conclusão de curso, seja graduação, mestrado ou doutorado, na elaboração de artigos científicos e investigações acadêmicas de modo geral.

De fato, entendemos a terminologia dados de pesquisa como algo mais abrangente que os dados científicos. Em nossa concepção, eles podem ser usados com finalidade científica ou não científica, já o segundo se restringem às pesquisas de caráter científico. Os dados de pesquisa, além de incluírem os dados científicos, são dados que podem ser gerados ou coletados com fins diversos, seja para fazer o levantamento de uma situação, por exemplo, por meio de um censo (dados governamentais), seja para identificar uma realidade empresarial (dados de mercado). O que vai dar um caráter científico aos dados de pesquisa é a utilização de uma metodologia para lidar com os dados que passe pelas diretrizes confiáveis estabelecidas pela ciência.

O entendimento é que todo dado científico é um dado de pesquisa, mas nem todos os dados de pesquisa são científicos. Desta forma, é possível adotar uma ou outra nomenclatura para se referir aos dados que constituem matéria-prima ou produtos das pesquisas científicas. O termo “dados acadêmicos” também é encontrado na literatura, porém em menor proporção, para este entendemos ter o mesmo significado de “dados científicos”, logo, faz parte dos dados de pesquisa. No estudo em questão será utilizada predominantemente a expressão “dados científicos”.

O processo de geração de dados normalmente envolve o investimento de recursos como tempo, dinheiro e esforços necessários, porém muitos dados estão perdendo a possibilidade de terem seu devido valor, por se encontrarem armazenados nos computadores dos pesquisadores ou em outros dispositivos de armazenamento que não os tornem públicos. Inúmeros são os motivos pelos quais os cientistas não compartilham os dados. Segundo Stall *et al.* (2019), os criadores dos dados dificilmente recebem crédito, e quando isso acontece, o reconhecimento se restringe a citações. Além do mais, a infraestrutura para a curadoria de dados é limitada. Esses impasses são inerentes a todas as áreas de conhecimento, elas mesmas não apresentam um discurso alinhado sobre a temática.

Para buscar abordar desafios como esses, no início de 2014, mediante uma solicitação do *Netherlands e-Science Center* e do *Dutch Techcentre for the Life Science* (DTL) no Centro Lorentz em Leiden, na Holanda, representantes de diversas partes interessadas se encontraram para discutir sobre como contribuir com o ecossistema da *eScience*, cujo desafio é possibilitar o uso ideal de dados e métodos de pesquisa. Diante do debate, surgiu a ideia de que através da definição e do apoio de um conjunto mínimo de princípios e práticas direcionadores, alinhados com a comunidade, fornecedores de dados e usuários de dados seria possível descobrir, acessar, interoperar e reutilizar a imensa quantidade de dados gerados pela ciência contemporânea (FORCE11, 2020).

Diante do contexto de uma necessidade premente de aprimorar a infraestrutura que sustenta a reutilização de dados provenientes das pesquisas acadêmicas, o termo FAIR, que corresponde ao acrônimo de *Findable* (Localizável), *Accessible* (Acessível), *Interoperable* (Interoperável) e *Reusable* (Reutilizável), foi criado em 2014, após dois anos, Wilkinson *et al.* (2016) publicaram um artigo de periódico em que são apresentados os princípios de dados FAIR (*FAIR Data Principles*).

Neste sentido, os princípios sugerem que os dados científicos sejam: localizáveis por qualquer indivíduo que faça uso de ferramentas de pesquisa comuns (STALL *et al.*, 2019), assim como devem ser fáceis de serem encontrados por computadores (GO FAIR, 2020).

Devem ser acessíveis, para ser possível examinar dados e metadados; interoperáveis, para que dados comparáveis consigam ser analisados e integrados por meio da utilização de vocabulário e formatos comuns; e reutilizável por outros pesquisadores ou pelo público, como consequência de metadados robustos e informações claras de proveniência e licenças (STALL *et al.*, 2019).

Segundo o GO FAIR (2020), o alvo desses princípios está atrelado à ação da máquina, de forma que essa seja capaz, com nenhuma ou mínima intervenção humana, de encontrar, acessar, interoperar e reutilizar dados, uma vez que diante do volume e velocidade em que os dados são gerados, as pessoas demandam cada vez mais os serviços dos sistemas computacionais. Assim, Wilkinson *et al.* (2016) corroboram com a ideia e revela que os Princípios FAIR estão voltados a aperfeiçoar a capacidade das máquinas, de forma a descobrir os dados e usá-los automaticamente e acrescentam que eles podem auxiliar a reutilização dos dados pelos indivíduos.

Essas diretrizes discutidas no cenário atual, caracterizado pelo uso intenso dos dados, contribuem com o compartilhamento dos dados científicos. Além de se preocupar com os interesses humanos, envolve uma preocupação com os agentes computacionais. O que é fundamental para a agregação de valor, pois beneficia a reutilização dos dados por outros usuários, além daqueles que originalmente geraram os seus dados, eliminando assim as barreiras que impedem esse processo.

2.1 PUBLICAÇÃO DE DADOS

A reflexão sobre a importância de o pesquisador publicar e compartilhar os seus dados científicos para que eles possam ser úteis em novas investigações é um assunto discutido na comunidade científica. As ciências sociais fazem parte deste cenário que trata da reutilização de dados.

Neste contexto, é importante compreender que “a pesquisa social trabalha com gente e com suas realizações, compreendendo-os como atores sociais em relação a grupos específicos ou perspectivas, produtos e exposição de ações, no caso de documentos.” (MINAYO, 2007, p. 62). A autora complementa que os questionamentos deste tipo de investigação estão relacionados à política, economia, relações, funcionamento das instituições, determinados problemas concernentes a segmentos sociais, cultura geral ou local, entre outros. Logo, os dados advindos da pesquisa social são produzidos em torno dessas temáticas.

Segundo Van den Eynden e Corti (2017), a publicação e o compartilhamento de dados científicos têm crescido em âmbito mundial, cuja realidade é fruto dos avanços tecnológicos,

assim como as necessidades de pesquisas e exigências de financiadores, editores e governos no tocante à abertura, transparência e investimento eficiente em pesquisa. Nas ciências sociais, temos o exemplo da *UK Data Archive*, que visando à curadoria, preservação e publicação de conjunto de dados valiosos busca cada vez mais aperfeiçoar as práticas de publicação de dados e aprimorar as habilidades dos pesquisadores.

Neste sentido, Monteiro e Sant’Ana (2018) revelam que a gestão adequada dos dados é uma condição prévia indispensável para o seu compartilhamento eficaz entre uma comunidade científica específica e comunidades não pertencentes ao grupo alvo.

Os dados das pesquisas sociais podem ser recursos de acesso comum, por meio de arquivos abertos para alguns pesquisadores ou bens do clube, restrito aos membros de um consórcio. Também podem ser particulares, quando compartilhados apenas entre os estudiosos individuais ou podem ser bens públicos no momento que são partilhados abertamente sem restrição de uso. Neste contexto, a obtenção dos dados científicos pode custar caro, mas a reutilização pode economizar recursos como dinheiro, tempo e experiência para expandir a diversidade de recursos disponíveis voltados à pesquisa (BORGMAN, 2015). Além disso, os pesquisadores geralmente utilizam recursos públicos para obter os seus dados, o que seria outro motivo para disponibilizá-los publicamente (SILVA, 2019).

Em estudo realizado por Veiga *et al.* (2019), sobre os benefícios percebidos pelos pesquisadores da instituição Fiocruz, no tocante ao compartilhamento e abertura dos dados provenientes das investigações científicas, foi evidenciado que a maioria dos pesquisadores elegeram a transparência na condução das pesquisas e a otimização do avanço da ciência como principais vantagens dessas práticas, além do aumento da acessibilidade, confiabilidade, visibilidade e o reconhecimento profissional. A principal barreira manifestada pelos pesquisadores foi a desinformação dos tipos de compartilhamento das plataformas existentes e das normas vigentes. A realidade em questão levou a instituição a elaborar um plano de gestão de dados. Fato este que demonstra uma iniciativa da instituição para atender de forma pontual às limitações do processo.

Para a adoção do compartilhamento de dados, é necessário levar em consideração aspectos inerentes à sociologia da ciência, tais como: existe a necessidade de ficar claro para os pesquisadores as vantagens desta prática e qual o incentivo ele receberá em troca. Diante do tempo que é investido para tornar os dados disponíveis, por deixar de ter o acesso exclusivo e, conseqüentemente, perder a vantagem competitiva de não ter os dados exclusivos (SILVA, 2019).

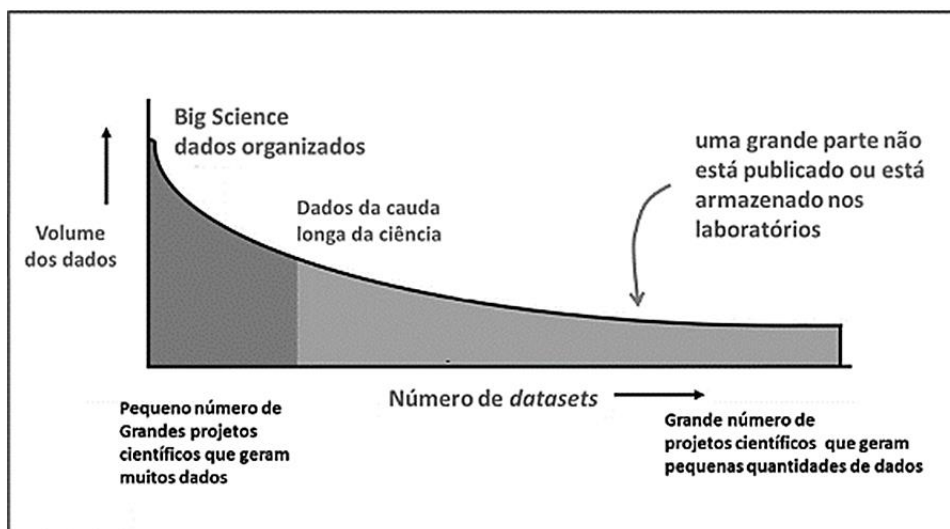
Piwowar e Vision (2013) alertam que disponibilizar os dados publicamente envolve desafios e custos. Alguns destes são financiados pela sociedade, porém diversos custos são investimentos dos próprios pesquisadores, realizados durante a coleta dos seus dados. Os autores destacam o desafio dos pesquisadores terem receio de outros investigadores identificarem erros nos seus resultados ou aplicarem análises adicionais que seriam usadas por eles em um momento posterior.

Segundo Candela *et al.* (2015), diversas barreiras prejudicam a publicação dos dados, tais como: a ausência de atribuição e recompensas, práticas limitadas de citação, a falta de uma cultura de compartilhamento de dados. Desta forma, para que os dados possam estar disponíveis, são necessárias práticas e políticas de compartilhamento de dados.

Neste contexto, Sayão e Sales (2019) mostram que a gestão e o compartilhamento de dados ocorrem de forma diferente entre os dois universos da ciência, caracterizados de “*big sciences*” (grande ciência) e “*small sciences*” (pequena ciência). Para Borgman (2012), A primeira produz grandes volumes de dados e referem-se a ciências como a Astronomia, Astroinformática, Bioinformática e Humanidades digitais. Por outro lado, existem os pequenos laboratórios que compõe as estruturas de pesquisa das universidades que também se destacam na geração e uso de dados, com a utilização de novos métodos e instrumentos que possibilitam a coleta de volumes sem precedentes de observações por parte dos pesquisadores individuais e pequenas equipes.

Os segmentos científicos da grande e pequena ciência, no que se refere à análise e distribuição da geração de dados, geralmente têm base no conceito estatístico de “cauda longa” (*long tail*). Essa forma de distribuição estatística, comumente voltada para as áreas de *marketing*, principalmente, de produtos digitais, mostra que em uma pequena parte da população há diversas ocorrências (cabeça), ao passo que em uma grande parcela existem poucos casos (a cauda) (SAYÃO; SALES, 2019), conforme está expresso na Figura 2.

Figura 2 - Dados da cauda longa da Ciência



Fonte: Sayão e Sales (2019, p.40)

De acordo com Ferguson *et al.* (2014), em relação ao mundo da pesquisa, grandes esforços científicos apresentam dados homogêneos e bem-organizados, o que corresponde apenas a uma pequena proporção de dados coletados por pesquisadores. Do outro lado existe uma grande quantidade de dados científicos que é distribuída na cauda longa, com inúmeros esforços de pesquisa independente e com a produção de uma rica diversidade de conjuntos de dados científicos. A parte da extrema direita da cauda inclui dados que não estão publicados, como banco de dados isolados, descobertas nulas, notas de laboratório, dentre outros. Para os autores, esses dados obscuros possuem um grande potencial de conhecimento, mas comumente não estão acessíveis ao público.

A grande ciência relativamente trabalha com um número pequeno de conjunto de dados volumosos, que são facilmente detectáveis. Já a pequena ciência é produtora de grandes quantidades de dados amplamente dispersos, mantidos em inúmeros arquivos de computador, difíceis de serem encontrados (HEDSTROM; MYERS, 2014). Sayão e Sales (2019) complementam que, neste contexto, o fluxo de compartilhamento de dados se limita aos colaboradores mais próximos ou através de *websites* dos projetos das instituições. A gestão ocorre com base em ferramentas como *off-the-shelves*², que de acordo com o Cambridge Dictionary (2021), consiste em um produto disponível para uso imediato, a exemplo de planilhas eletrônicas. Todavia, grande parte desses dados é mantida sem ações de gestão e curadoria e, com o tempo, desaparece em razão da obsolescência tecnológica dos dispositivos

² *used to describe a product that is available immediately and does not need to be specially made to suit a particular purpose.*

portáteis de armazenamento e dos discos dos computadores dos pesquisadores. Esse cenário também remete à grande parte dos dados obscuros, difíceis de serem visíveis, e, consequentemente, reutilizáveis.

Diante das reflexões apresentadas, observa-se que a ciência é produtora de uma grande quantidade de dados e se encontra distribuída pela cauda longa. Ela envolve um universo de dados valiosos, que apesar de não possuir os mesmos investimentos da grande ciência, pode enriquecer o conhecimento científico a partir da gestão, curadoria e a publicação dos dados, ao permitir que eles sejam visualizados por outros e reutilizados em novos projetos. A seguir, são apresentadas alternativas que possibilitam as descobertas dos dados como os repositórios de dados, as publicações ampliadas e os periódicos de dados.

2.1.1 Repositório de Dados

Disponibilizar os dados de forma que sua reutilização seja possível constitui um desafio, pois é preciso uma gestão adequada para democratizar o acesso aos dados científicos de forma universal. Para buscar atender às mudanças do cenário atual da ciência existem, atualmente, algumas opções na literatura como solução para o compartilhamento de dados, tais como: repositório de dados específicos da área, repositório de dados institucionais, repositórios genéricos (BEZJAK *et al.*, 2018). Por meio do OpenDoar³, um diretório global de repositórios de acesso aberto, é possível encontrar repositórios que publicam conjuntos de dados de diferentes tipos.

Van den Eynden e Corti (2017) apontam alguns benefícios em preservar os dados em um repositório de dados específicos, a saber: a garantia da qualidade de que os dados estão alinhados com os padrões estabelecidos; preservação de dados no longo prazo em formatos de arquivos acessíveis; o depósito de dados em um local seguro que possibilita, quando necessário, o controle do acesso; encontro de recursos online e acesso a dados através de catálogos de dados; apoio ao usuário; e oportunidades promocionais e de treinamentos voltadas à coleta de dados, proporcionando uma maior visibilidade na temática dos dados. Neste contexto, os autores citam o repositório de dados intitulado de *Reshare*⁴, que se tornou o principal sistema de publicação de dados científicos no Reino Unido, na área de ciências sociais.

A Universidade de Cambridge e a Universidade de Leeds, ambas localizadas na Inglaterra, possuem repositórios de dados institucionais com foco na descoberta e

³ Disponível em: <https://v2.sherpa.ac.uk/opensoar/>

⁴ Disponível em: <https://reshare.ukdataservice.ac.uk/>

compartilhamento dos dados. Repositórios de dados como o *Dryad*⁵, *Figshare*⁶, *Zenodo*⁷ e o *Dataverse*⁸ são do tipo generalista.

Monteiro (2017) enfatiza que os repositórios de dados científicos têm uma parcela crucial no processo de gestão de dados. Eles tornam viável a representação, o processamento, gerenciamento, disseminação e o acesso aos dados que neles foram depositados, envolvendo métodos de curadoria e fortalecendo a colaboração científica internacional ao propiciar o compartilhamento, acesso, uso e reúso dos dados.

No entanto, conforme alertam Curty e Aventurier (2017), mesmo que os repositórios de dados venham investindo em ferramentas a fim de contribuir com a descoberta e a citação dos dados depositados, essa modalidade de publicação não é considerada tão atrativa para aqueles que compartilham os dados, principalmente em relação às limitações inerentes à recompensa e ao crédito científico. Logo, os pesquisadores geralmente escolhem meios e formatos que possibilitem maior visibilidade e, por consequência, gerem mais créditos e citações. Neste contexto, outras abordagens como as publicações ampliadas vêm se destacando. Assim sendo, os esforços voltados à publicação do registro e da documentação dos dados através do repositório necessitam proporcionar ao pesquisador uma maior notoriedade para se tornar uma abordagem mais atrativa.

2.1.2 Publicações Ampliadas

No final da década de 90, Meadows (1999) vislumbrou que, em razão do desenvolvimento da Tecnologia da Informação (TI), os dados brutos poderiam ser publicados como parte de um relatório de pesquisa juntamente com a análise dos dados, o que promoveria o aumento da circulação da informação científica. Posteriormente, surgiram as primeiras tentativas para dar suporte à publicação de dados. Neste cenário, os dados passaram a fazer parte dos artigos ou residir em arquivos suplementares anexados aos artigos (CANDELA *et al.*, 2015). Essa realidade remete às publicações ampliadas que são utilizadas atualmente.

De acordo com Santos e Rockembach (2018), em detrimento do uso intenso dos dados no contexto atual da ciência aberta, as publicações ampliadas correspondem a uma das práticas de comunicar a ciência. As características dessa modalidade ultrapassam o escopo de uma

⁵ Disponível em: <https://datadryad.org/stash>

⁶ Disponível em: <https://figshare.com/>

⁷ Disponível em: <https://zenodo.org/>

⁸ Disponível em: <https://dataverse.org/>

publicação tradicional, interligando-a aos seus próprios dados, além de outros documentos e objetos digitais relacionados. Mucheroni, Silva e Paletta (2015) mostram que enquanto forma de publicação eletrônica, elas podem ser aproveitadas para divulgar e compartilhar os resultados de uma pesquisa.

Sayão e Sales (2016) conceituam uma publicação ampliada como uma instância de um objeto digital complexo, que se caracteriza pela ligação de artigos aos dados que subsidiaram a sua geração, assim como a outras informações que ajudem na contextualização da pesquisa, por exemplo, informações a respeito de autores, sobre os equipamentos utilizados, projetos, dentre outros.

Curty e Aventurier (2017) enfatizam que, nesta situação, existe o vínculo entre o manuscrito e os dados, estes não são protagonistas, mas sim auxiliares e dependentes do artigo científico. Para os autores, esta abordagem de publicação não permite a citação independente de dados, pela carência de identificadores e metadados pertinentes, o que pode provocar uma barreira no reuso dos dados. Ao tratar das limitações dessa abordagem, Candela *et al.* (2015) acrescenta que nesse processo são necessárias a curadoria e a preservação desses arquivos. Outra questão apontada é que não permitem aos usuários descobrir e vincular dados independentemente da publicação principal.

Assim sendo, as publicações ampliadas necessitam de um identificador único e informações pertinentes com metadados descritivos, para que assim venham a contribuir com os domínios do conhecimento científico, geralmente compostos por um conjunto de partes vinculadas, referente aos registrados em diversos tipos de comunicação (por exemplo, conjuntos de dados, vídeos, imagens, folhas de estilo, bancos de dados, slides de apresentações, entre outros) e descrições textuais da pesquisa (por exemplo, documentos, capítulos, seções, tabelas) (MUCHERONI; SILVA; PALETTA, 2015).

2.1.3 Periódicos de Dados

O artigo de dados corresponde a outra abordagem de publicação que torna o compartilhamento de dados possível. Essa abordagem apresenta uma maior ênfase nos dados, quando comparada com as publicações ampliadas, pois o seu objetivo é descrever as coleções de dados de forma exaustiva, apresentando a descrição do contexto, dos procedimentos metodológicos da investigação e das possíveis aplicabilidades dos dados, podendo ser expressa através da publicação dos artigos de dados em periódicos científicos híbridos, receptivos aos *data papers*, ou periódicos voltados à publicação de dados (CURTY; AVENTURIER, 2017).

Segundo Monteiro (2017), periódicos que publicam dados científicos surgiram como mais um canal de comunicação científica para que pesquisadores publiquem seus dados. Periódicos de dados são tendências ímpares para tornar públicos, de forma mais ampliada, conjuntos de dados. O autor apresenta alguns periódicos com essa natureza, conforme apresentados no Quadro 1.

Quadro 1 - Lista de periódicos de dados com os respectivos *links* de acesso

Periódicos de Dados	Endereço eletrônico
Biodiversity Data Journal	https://bdj.pensoft.net/
Data in Brief da Elsevier	https://www.journals.elsevier.com/data-in-brief
Earth system Science Data	https://www.earth-system-science-data.net/
GeoScience Data Journal	https://rmets.onlinelibrary.wiley.com/journal/20496060
Journal of Open Archeology Data	https://openarchaeologydata.metajnl.com/
Journal of Open Psychology	https://openpsychologydata.metajnl.com/
Journal of Physical and chemical Research Data	https://aip.scitation.org/journal/jpr
Journal of Open Research Software	https://openresearchsoftware.metajnl.com/
Open Health Data	https://openhealthdata.metajnl.com/

Fonte: Adaptado de Monteiro (2017)

Neste contexto, como já mencionado anteriormente, citamos como exemplo a revista de dados Scientific Data, da Springer Nature. Este periódico é de acesso aberto, revisado por pares, voltado para descrições de conjuntos de dados com valor científico. O seu objetivo é fomentar o compartilhamento e a reutilização de dados mais amplos e creditar aqueles que compartilham os dados compilados em suas respectivas pesquisas. Outro exemplo é o Data Science Journal, que também consiste em um periódico eletrônico de acesso aberto, revisado por pares, que publica artigos sobre gerenciamento, disseminação, uso e reutilização de dados e bancos de dados. Este periódico envolve todas as áreas do conhecimento científico.

Van den Eynden e Corti (2017) tratam da realidade das ciências sociais e revelam que cada vez mais conjuntos de dados, desta área do conhecimento, são publicados em periódicos de dados, como Scientific Data, Research Data Journal for the Humanities e Social Sciences ou Journal of open psychology data. Neste contexto, o artigo de dados descreve a metodologia de geração de dados, a proveniência e o potencial de reutilização.

Segundo Curty e Aventurier (2017), é importante destacar que a elaboração de um *data paper* para ser submetido em uma revista de dados requer o investimento de recursos como esforço e tempo dos pesquisadores, o que pode ser uma barreira para os envolvidos neste processo, que podem optar por produzir vários artigos científicos a partir de dados coletados, ao invés de tornarem públicos os *data papers*.

Kratz e Strasser (2014) revelam que o tamanho e a estrutura de um *data paper* diferem entre os periódicos, porém são unânimes na tendência em apresentar um formato curto e bem estruturado. Para os autores, eles são compreendidos com maior clareza, não pela presença de qualquer informação em particular, mas pela inexistência de análises e conclusões. Portanto, sua definição está relacionada com a ausência de uma publicação prévia sobre a análise dos dados.

Os repositórios de dados científicos, as publicações ampliadas e os *data papers* que são encontrados nas revistas de dados apresentam suas vantagens, assim como desafios, porém são iniciativas promissoras, compreendendo canais através dos quais os dados podem ser compartilhados. Essas abordagens podem ser aperfeiçoadas para atender prontamente aos anseios dos pesquisadores que passaram a demandar mais recursos, como tempo e esforço, para disponibilizar os seus dados, logo, o fator social não pode ser esquecido.

De modo geral, o capítulo 2 refletiu os conceitos de dados científicos encontrados na literatura; mostrou as diferenças tipológicas dos dados em relação a natureza e a origem; discutiu os princípios de dados *FAIR* e apresentou os repositórios de dados, as publicações ampliadas e os periódicos de dados como opções para publicar e compartilhar os dados provenientes das investigações científicas. Nesse contexto, o capítulo seguinte aborda a ferramenta de gestão ciclo de vida dos dados, que possibilita as condições para que os dados científicos sejam reutilizados.



3 CICLO DE VIDA DOS DADOS

De acordo com Rice e Southall (2016), o ciclo de vida de pesquisa rastreia a atividade do pesquisador durante o curso de um projeto de pesquisa. De modo similar, um CVD segue o caminho dos dados ou as ações necessárias nos dados para mover a pesquisa para o próximo estágio. Goben e Raszewski (2015) complementam que o ciclo de vida dos dados compreende os estágios pelos quais os dados atravessam, desde o momento em que são criados até serem descartados.

Anjos e Dias (2019), ao visualizarem o ciclo de vida dos dados como um instrumento que ajuda o pesquisador no decorrer das investigações científicas, enfatizam que cada ação que corresponde as suas etapas é primordial para o processo de curadoria, assegurando uso otimizado e sustentável aos dados. Seguindo o mesmo posicionamento, para o DataONE (2020), o CVD proporciona uma visão abrangente dos estágios relacionados ao processo de gestão e preservação adequados dos dados para uso e reúso.

É possível observar que os autores supracitados são unânimes em destacar o ciclo de vida de dados como uma ferramenta que oferece suporte aos pesquisadores através dos seus estágios, com ações necessárias que garantam que os dados se estendam para atender às necessidades pontuais e futuras, o que possibilita o avanço das pesquisas científicas.

Espíndola *et al.* (2018) visualizam os dados como insumos que trazem valor para as investigações científicas. Neste sentido, para os autores, dados de qualidade são fundamentais para o sucesso de investigações científicas, por isso a necessidade de adotar boas práticas de coleta, armazenamento e recuperação.

Conforme Sayão e Sales (2016), lidar com o ciclo de vida dos dados da pesquisa é uma atividade complexa, pois envolve a captura de dados em distintas fases de seu processamento, além de considerar linhagens e versões que mudam com o passar o tempo. Desta forma, Anjos (2019) mostra que para dar suporte à condução dos ciclos de vida dos dados, é preciso haver profissionais capacitados. A este respeito, Costa (2017) enfatiza a necessidade do bibliotecário se posicionar como um profissional apto a atuar nesse contexto.

Ao se referirem ao profissional bibliotecário, Goben e Raszewski (2015) revelam que o entendimento do processo referente ao CVD o beneficia em seu próprio trabalho, por ele mesmo produzir diariamente conjuntos de dados valiosos. A habilidade em lidar com os dados leva à compreensão sobre os dados gerados pela própria biblioteca, assim, com base nas evidências desse entendimento, esses profissionais podem aplicar a biblioteconomia e obter um trabalho administrativamente eficaz. Além do mais, podem contribuir com os pesquisadores, ao possibilitar a redução de barreiras na coleta e análise de dados, assim como na curadoria e na

descoberta de conjunto externos importantes. Logo, esse profissional pode cooperar com o usuário e com a sua biblioteca.

No que tange ao acesso a dados científicos, segundo Silva (2016), os bibliotecários são capazes de auxiliar os pesquisadores durante as várias fases do CVD, quando eles geram e utilizam os dados em seu plano de trabalho. Nesse contexto, serviços de apoio, tais como ajudar na escolha do repositório, oferta de oficinas sobre o plano de gestão de dados, desenvolvimento de estratégias de metadados, dentre outras ações devem ser desenvolvidas em aproximação com os pesquisadores e podem envolver a própria realização de plano de gestão para a devida documentação e organização dos dados, por meio de ferramentas seguras de armazenamentos de dados.

Sales (2020) coaduna com a ideia de o bibliotecário contribuir com o pesquisador na elaboração do plano de gestão de dados e acrescenta que o profissional pode corroborar na coleta/geração dos dados, na decisão sobre a escolha dos metadados, bem como na contextualização dos dados, fornecendo informações sobre questões éticas e legais e ainda ajudando a identificar o melhor repositório para depositar os dados científicos.

Em estudo recente, realizado por Anjos (2019), foi investigado como os profissionais da informação, especificamente os bibliotecários, estão capacitados para atuar na curadoria de dados científicos durante as etapas do CVD-CI. A partir dos achados, verificou-se a ausência de componentes curriculares obrigatórios a nível de graduação no Brasil que tratem do quarto paradigma da ciência, no tocante ao processo de gestão dos dados ou ao CVD. Apesar de os cursos de graduação em Biblioteconomia no Brasil possuírem competências voltadas à realidade dos dados, assim como ao CVD-CI, a autora sugere reformar as matrizes curriculares dos cursos. Esta realidade pode oportunizar mais um campo de atuação para o profissional dessa área.

Para atuar na curadoria dos dados científicos, é preciso reformular os perfis dos bibliotecários, já que agora esses profissionais precisam se preocupar com a origem dos dados e com toda documentação pertinente à interpretação e contextualização dos dados em longo prazo (SAYÃO; SALES, 2016).

É possível constatar a importância do profissional da informação no contexto do ciclo de vida dos dados, seja no desenvolvimento do seu trabalho, seja ao oferecer um suporte aos pesquisadores, o que pode resultar em dados de qualidade. Porém, é importante enfatizar a necessidade de capacitação desses profissionais, tanto por meio das estruturas curriculares dos cursos de graduação e pós-graduação e/ou por cursos de treinamentos que os habilitem a essa

realidade. Além do mais, é necessário um profissional suficiente para não tornar o trabalho exaustivo com uma sobrecarga e, conseqüentemente, gerar desmotivação.

3.1 INICIATIVAS DE CICLOS DE VIDA DOS DADOS

A importância dos modelos de ciclos de vida dos dados está relacionada a proporcionar uma estrutura que envolva as diversas atividades que precisam ser realizadas em um registro de dados no decorrer do seu uso (SILVA, 2019). Na literatura, é possível encontrar diversos modelos voltados a descrever o ciclo de vida dos dados, conforme ilustrados no Quadro 2.

A tradução do termo ciclo de vida dos dados para o inglês é *data life cycle*. Em estudo recente realizado por Araújo, Dias e Araújo (2019) além do termo já mencionado, foram identificados na literatura três formas terminológicas distintas, a saber: *data life-cycle*, *data lifecycle* e *big data life cycle*, fato esse que deve ser levado em consideração quando da realização de investigações acerca da referida temática. Para esta pesquisa, adotamos a versão em português ciclo de vida dos dados.

Segundo o DataONE (2020), entre os diferentes domínios ou comunidades existem diversas versões de um ciclo de vida de dados com suas distintas atribuições. Araújo *et al.* (2019) apresentam um quadro baseado em 18 modelos identificados por Sant’Ana (2013). O referido pesquisador, com base nas 18 iniciativas, desenvolveu um modelo específico para a área da Ciência da Informação, por entender que esta área do conhecimento pode contribuir com o acesso e uso intenso dos dados (SANT’ANA, 2013).

O Quadro 2 também possui dois modelos apresentados por Rice e Southall (2016), um modelo encontrado na obra “*Linked Open Data - Creating Knowledge Out of Interlinked Data*” (Dados abertos vinculados, criando conhecimento a partir de dados conectados) e outro da *The University of California, Santa Cruz*, com um total de 23 iniciativas com suas respectivas características e usos.

Quadro 2 - Iniciativas de modelos de Ciclos de Vida dos Dados

INICIATIVAS	CARACTERÍSTICAS	FONTES
Entropia de Dados	Denuncia desordem e degradação dos dados e metadados em função do tempo e da obsolescência.	Michener <i>et al.</i> (1997)
CVD – Componentes e recursos estratégicos	Mostra insumos e resultados do processo como produtos bem delineados e definidos; característica de aplicabilidade.	Ferderer (2001)
Ciclo de Vida de Metadados - CVMD	Sinaliza padrões adequados para assegurar a descrição e o controle ao longo do tempo.	Chen <i>et al.</i> (2003)
Ciclo de Vida dos Dados – DDI	Composto por oito fases identificadas em sequência; base para outros modelos.	Data Documentation Initiative – DDI (2004)

CVD e a Pesquisa Translacional	Caracterizado como “ <i>Translational Research</i> ”, o diagrama é composto por quatro ciclos que apresentam a percepção do acesso aos dados.	VUMC (2005)
CVD de Conhecimento de Pesquisa	Ilustra as principais considerações para o arquivamento em cada etapa do processo de criação de dados.	Humphrey (2006)
CVD – Pennock	Identifica fase de acesso que o registro é utilizado; fase de acesso e reuso que pode gerar a criação de novos registros.	Pennock (2007)
CVD e Publicações	Citado na CI: descreve fases envolvidas no processo de publicação de produções científicas.	Gold (2007)
Modelo de Ciclo de Vida de Dados Digitais	Extrapolando o fluxo informacional em si e incorpora elementos contextuais como a própria organização, perfis profissionais necessários e políticas de informação.	IWGDD (2009)
Ciclo de Vida de Dados Abertos Vinculados	Colabora no processo de compreensão das diferenças nos ciclos de vida de dados ordinários e de dados ligados.	Villazón-Terrazas et al. (2009)
Tarefas e atores envolvidos no gerenciamento da base de dados da SHARE	Mostra sequência circular e linear de fases, acrescentando ainda as tarefas e atores envolvidos em cada uma.	Hunkler <i>et al.</i> (2011)
CVD de Pesquisa – UCL*	Inclui as principais tarefas envolvidas em cada fase e enfoque em dados coletados por meio de entrevistas.	UCL (2012)
CVD – DAMA	Figura da operação de descarte dos dados: “ <i>retention and destruction</i> ”.	DAMA (2012)
O compartilhamento de dados para pesquisa demográfica	Cita fases que congregam mais de uma tarefa e apresentam no diagrama estes objetivos, facilitando o entendimento.	DSDR (2012)
CVD - comparado com o projeto que o cria	Proporciona um cenário de análise para o acompanhamento de cada uma das fases.	UKDA (2013)
Ciclo de Vida dos dados de Materiais	Identifica práticas de gestão, análise e uso de informações sobre materiais, vinculando suas iniciativas à aplicação.	MDMC (2013)
Ciclo de Vida dos Dados	Atua com estruturas distribuídas e sustentáveis que atendam às necessidades da ciência e da sociedade de acesso aberto, destacando a pesquisa na área ambiental.	DataONE (2013)
Ciclo de Vida de Curadoria	Destaca gestão e ações administrativas planejadas para promover a curadoria e preservação ao longo do ciclo de vida da curadoria.	DCC (2004)
CVD – CI	Propõe um modelo CVD que seja compartilhado pela CI, fortalecendo a compreensão dos esforços e resultados das pesquisas realizadas.	Sant’Ana (2013)
Ciclo de Vida de Dados Conectados	Pode contribuir com a ciência, economia, cultura e a sociedade em geral.	Auer (2014)
Ciclo de Vida de Dados de Pesquisa	Mostra que os serviços podem ser mapeados em um fluxo de trabalho centrado no pesquisador, com diagrama circular sinalizando que os dados continuam a viver após a pesquisa.	Jisc (2016)
Fluxo de trabalho do RDM da Biblioteca do Imperial College London	Fornecer um exemplo de serviços que podem ser mapeados em modelo de ciclo de vida de dados do <i>Imperial College London</i> .	Barnes (2016)
Ciclo de Vida do Gerenciamento de Dados de Pesquisa	Auxilia os professores, funcionários e alunos da UC Santa Cruz na organização, preservação e gerenciamento de dados de pesquisa.	UCSC

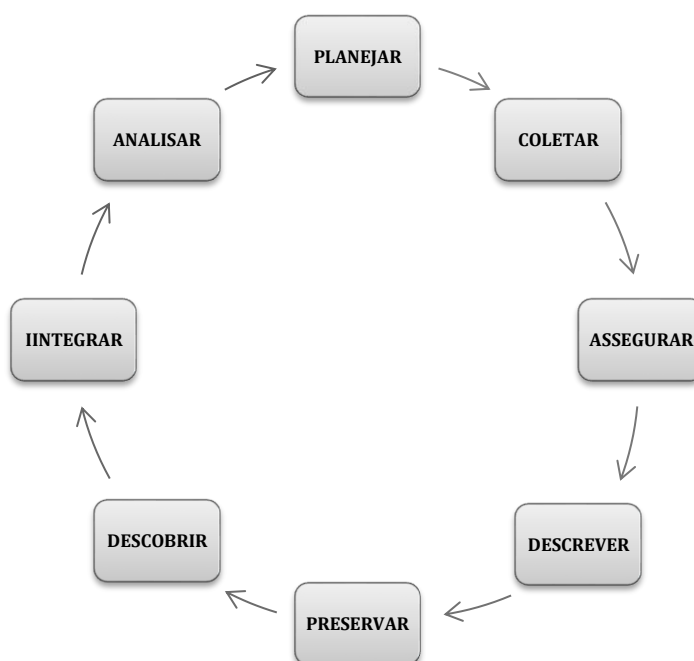
Fonte: Adaptado de Araújo *et al.* (2019)

Entre as iniciativas de ciclo de vida dos dados mais populares, Goben e Raszewski (2015) exemplificam o *Digital Curation Center* e o DataONE. Estes modelos e a iniciativa do DDI serão utilizados no presente estudo e estão detalhados a seguir. Por esse motivo, eles foram destacados em negrito no Quadro 2.

3.1.1 Ciclo de Vida dos Dados – DataONE

O DataONE (2020) é um projeto que tem a colaboração de diversas organizações parceiras e é financiado pela *National Science Foundation* (FNS) dos Estados Unidos, mediante um acordo cooperativo. Trabalha com a gestão dos dados da terra e ambientais. Neste contexto, utiliza um ciclo de vida de dados para apresentar os estágios inerentes à gestão bem-sucedida e à preservação para que os dados possam ser usados e reutilizados, conforme ilustrado na Figura 3:

Figura 3 - Ciclo de vida dos dados - DataONE



Fonte: Adaptado e traduzido do DataONE (2020)

Strasser *et al.* (2012) apresentam de forma resumida e detalhada os 8 (oito) componentes do ciclo de vida de dados do DataONE, conforme elencados no Quadro 3.

Quadro 3 - Ciclo de vida dos dados – DataONE

Componentes	Descrição
Planejar	É o momento da descrição dos dados que serão compilados e como será realizada a sua gestão e disponibilização no decorrer da sua vida útil.
Coletar	As observações são realizadas de forma manual, por meio de sensores ou outros instrumentos e os dados ficam no formato digital.
Assegurar	É o momento em que é garantida a qualidade dos dados através de verificações e inspeções.
Descrever	Os dados são descritos de modo preciso e detalhado por meio de padrões de metadados pertinentes.
Preservar	Os dados são colocados em um arquivo de longo prazo adequado (exemplo, <i>data centers</i>)
Descobrir	Os dados potencialmente úteis são encontrados e obtidos com as informações relevantes sobre eles (metadados).
Integrar	Ocorre a combinação dos dados de fontes distintas para formar um conjunto homogêneo de dados para que sejam analisados.
Analisar	Possibilita que os dados sejam analisados.

Fonte: Adaptado de Strasser *et al.* (2012)

De forma mais aprofundada, serão apresentados detalhes que devem ser explorados em cada um desses estágios para proporcionar que os dados sejam submetidos a um processo de gestão com qualidade.

No primeiro estágio, que é o **planejar**, é preciso decidir sobre vários aspectos. No que se refere à coleta de dados, Strasser *et al.* (2012) mostram algumas questões que precisam ser respondidas, tais como: quais dados serão gerados? De que forma eles serão coletados e analisados? Quais os instrumentos serão utilizados e a infraestrutura pertinente para dar suporte à pesquisa?

Quanto ao armazenamento, é preciso decidir a respeito de um repositório de dados que mais se adeque aos dados que serão gerados e para a comunidade que os utilizará, além de levantar as informações do repositório sobre os requisitos de envio, como documentação de dados, padrões de metadados e qualquer restrição de uso que possa existir (a exemplo de direitos de propriedade intelectual). Com base nessas especificações do repositório, é possível organizar os formatos de arquivo e os tipos de produtos de dados que serão gerados (STRASSER *et al.*, 2012).

No que tange a gestão de dados, Strasser *et al.* (2012) evidenciam algumas questões que precisam ser esclarecidas sobre: quem será o responsável pela gestão dos dados? Como deve ocorrer o controle de versão? De que forma é realizado o *backup* dos dados e com que frequência?

A questão de descrever os dados envolve a necessidade de deixar claro: como será produzido um registro de metadados e com base em qual padrão? Ao compartilhar os dados, é preciso desenvolver um plano com a equipe do projeto, com outros colaboradores e com a comunidade científica. Algumas decisões serão requeridas nesse momento, a saber: o período de embargo e de que forma os dados serão liberados (STRASSER *et al.*, 2012).

A preservação dos dados ocorre por meio da execução de um plano de preservação de dados a curto prazo. Quanto ao orçamento disponível, é importante considerar o pessoal necessário para executar o plano de gerenciamento de dados, quais os tipos de recursos computacionais serão utilizados (por exemplo, tipos de *hardware* e *software*), que outras despesas poderão surgir (STRASSER *et al.*, 2012).

Oliveira e Silva (2016) mostram que o planejar é um momento de descrever o mapeamento dos processos e a ciberinfraestrutura (metodológica e tecnológica) que será necessária para atingir os estágios do ciclo de vida. Neste momento, os autores orientam o uso da ferramenta *Data Management Plan Tool (DMPTool)*, por apresentar uma boa prática de gestão de dados e por ser de livre acesso.

O segundo estágio é o **coletar**, os dados devem ser coletados com a garantia de que poderão ser usados posteriormente. Assim, é importante considerar métodos e documentação antes de iniciar a coleta, é preciso a criação de um modelo para usar no decorrer da coleta de dados. Recomenda-se a descrição do conteúdo dos arquivos de dados, a exemplo do formato e as unidades utilizadas. É preciso que os dados sejam organizados de forma consistente. Fazer uso do mesmo formato em todo arquivo. Utilizar um texto sem formatação complexa, o que possibilitará que o arquivo de dados seja legível pelo número maior de programas de *software*. É pertinente fazer uso de *software* e *hardware* estáveis e não proprietários. Fornecer nomes de arquivos descritivos que representem de forma ideal o projeto, o conteúdo do arquivo, o local e a data. Manter os dados brutos com todas as suas imperfeições e documentar todas as etapas de limpeza dos dados. Desenvolver uma tabela de parâmetros com a finalidade de descrever o código e as abreviações, as unidades, valores máximos e mínimos, o tipo de dados (por exemplo, texto, numérico) e apresentar uma tabela de sites que mostre os sites que os dados foram coletados (STRASSER *et al.*, 2012).

A coleta dos dados ocorre segundo as metodologias de cada pesquisa. Neste contexto, Oliveira e Silva (2016) alertam que os pesquisadores devem, neste momento, também focar a sua atenção na curadoria dos dados científicos. Algumas ferramentas são apresentadas pelas autoras para apoiar esse processo: a exemplo do *dropbox*, *google drive*, *software* específico,

equipamentos laboratoriais que condicionem e garantam que os dados primários sejam protegidos.

O terceiro estágio é o **assegurar**, que busca a garantia e o controle da qualidade dos dados durante a coleta de dados e sua análise. Para tanto, é recomendado descrever qualquer situação ocorrida no decorrer da coleta de dados que possa refletir na qualidade dos dados. É preciso analisar os valores estimados ao analisar os dados inseridos de forma manual e fazer uso de sinalizadores de nível de qualidade para levantar possíveis problemas que possam existir. É recomendada a identificação do formato dos dados para possibilitar a consistência em todo o seu conjunto. Por meio da codificação no interior do conjunto dos dados, é necessário comunicar a sua qualidade e analisar se existem valores que não estão presentes com a finalidade de não afetar a interpretação dos dados (STRASSER *et al.*, 2012).

Para complementar, Oliveira e Silva (2016) revelam a importância da adoção de recomendações legais antes que os dados passem para o próximo estágio, assim eles receberão também um *status* (aberto, fechado ou embargado), o que influenciará a questão da sua disponibilidade no futuro. Este também é um momento de decidir a plataforma ou o repositório que possibilitarão que os dados sejam compartilhados no futuro.

O quarto estágio é o **descrever**, que consiste na questão fundamental para que os dados sejam compreendidos no futuro, ao apresentar uma documentação complexa do contexto do arquivo de dados (intitulada de metadados). Nesse momento, é preciso levar em consideração a descrição do contexto digital com a nomenclatura do conjunto de dados; o nome do(s) arquivo(s) de dados e a data da última modificação do conjunto de dados; o *software* (com a informação sobre o número da versão) utilizado para trabalhar com o conjunto de dados, o processamento de dados usado. Além disso, é preciso ter uma descrição do pessoal e das partes interessadas, quem coletou os dados, quem deve ser considerado como fomentador; apresentar o contexto científico justificando a razão da coleta dos dados; apresentar os dados que foram coletados, quais os instrumentos foram utilizados e as condições ambientais no momento da coleta (STRASSER *et al.*, 2012).

Este é um momento de escolher um padrão de metadados mais adequado para natureza da pesquisa. Na literatura, é possível encontrar diversos padrões de metadados específicos para distintos tipos de dados científicos. Para os metadados serem descritos com qualidade, recomendam-se a utilização de vocabulários controlados, tesouros e padrões (OLIVEIRA; SILVA, 2016).

O quinto estágio é o **preservar**. Recomenda-se utilizar um *data center*, que pode apresentar algumas orientações como preparar os metadados, preservar os dados, o formato de

arquivo e conter serviços extras para usuários que, no futuro, tenham interesse em utilizar os dados. Nesse momento, é pertinente considerar os dados com valor em longo prazo, ou seja, arquivar os dados que serão mais úteis para possíveis usuários futuro dos dados (geralmente os dados brutos); armazenar dados com precisão adequada; e fazer uso de terminologia padrão – uma solução seria por meio da consulta a ontologias da área. É preciso levar em consideração, ainda, as políticas e questões legais que podem refletir da reutilização dos dados, por isso o acesso deve ser o mais aberto possível com a quantidade mínima de restrições. Portanto, é preciso focar em questões de privacidade e confidencialidade, direitos autorais e nos dados a serem licenciados, bem como é importante o pessoal responsável durante a vida útil do conjunto de dados e como citá-lo de forma adequada (STRASSER *et al.*, 2012).

O sexto estágio é o **descobrir**, o sétimo é o **integrar** e o oitavo é **analisar**. Eles, facilitados por um conjunto de ferramenta, contribuem para que os dados sejam descobertos, integrados e analisados. Nesse contexto, quando um conjunto de dados compreende o ponto de partida para novos conjuntos de dados, é fundamental documentar e identificar o conjunto de dados derivado. Este fato possibilitará o entendimento da proveniência do conjunto de dados e os seus elementos, a citação os autores do conjunto de dados original e o levantamento de possíveis erros que possam existir no conjunto original (STRASSER *et al.*, 2012).

No estágio **descobrir**, é o momento que os dados são encontrados por outros. Para que este processo ocorra, se faz necessária a publicação, uma vez que, se não estiverem armazenados em repositórios adequados, voltados ao compartilhamento, é muito provável que os dados sejam esquecidos, engavetados ou até mesmo descartados. Outra questão fundamental é deixar claro se os dados serão disponibilizados apenas para serem visualizados ou para serem reutilizados (SILVA SEGUNDO; ARAÚJO, 2019). Neste contexto, Sayão e Sales (2018) coadunam com a ideia e mostram a necessidade de os pesquisadores identificarem repositórios apropriados para a publicação dos seus dados, o que cada vez mais é uma exigência das agências fomentadoras da pesquisa, dos editores da ciência e da comunidade que faz parte. Além de ser fundamental para a descoberta de coleções de dados de qualidade que possam ser reutilizados para o avanço da ciência.

Conforme vimos anteriormente, a descoberta dos dados por meio da publicação não ocorre apenas mediante os repositórios de dados científicos, Curty e Aventurier (2017) acrescentam duas abordagens, as publicações ampliadas e os artigos de dados.

Silva Segundo e Araújo (2019) revelam que para o efetivo funcionamento do estágio **integrar**, momento em que ocorre a reutilização, é preciso que tenha sido feita uma adequada gestão das etapas antecessoras e que os dados estejam acessíveis para serem integrados em

outros projetos. Oliveira e Silva (2016) acrescentam a importância de aderir a protocolos de interoperabilidade para integrar os metadados cujo propósito é proporcionar visibilidade, novas análises e pesquisas. Desta forma, o caminho pode ser a adoção de ferramentas disponíveis pelo *Open Archival Information System* (OAIS), com a finalidade de preservar e disponibilizar, através de seus protocolos de interoperabilidade, dados, informações e documentos ligados ao projeto de pesquisa.

Por fim, há o estágio **analisar**, que para Silva Segundo e Araújo (2019) é o momento de verificar se estão sendo alcançados os objetivos propostos. Para tanto, se faz necessário que o curador analise com regularidade se o plano de gestão irá assegurar a segurança e preservação dos dados. Para o DataONE (2020), é o momento de considerar a compatibilidade dos dados que estão sendo integrados, é a hora de descrever as etapas de processamento para produtos de dados derivados, as quais devem ser devidamente documentadas para possibilitar que outras pessoas avaliem e interpretem os resultados. Também é o momento de utilizar métodos estatísticos para levantar possíveis erros em dados potenciais. Além de ser um período de identificar valores estimados, quando as medições reais não puderem ser alcançadas.

Lenhardt *et al.* (2014), ao referenciar a iniciativa do DataONE, revelam que as diversas etapas inerentes ao seu ciclo de vida dos dados são autoexplicativas. Porém, faz parte do processo básico a geração, controle de qualidade, armazenamento e análise dos dados necessários para oferecer suporte à ciência. Ultrapassando esse nível basilar, o ciclo de vida está focado em apresentar as informações que forem precisas para possibilitar a gestão efetiva dos dados em um contexto de arquivamento de longo prazo, além de oferecer apoio à reutilização. No decorrer do processo, metadados e documentações são adicionadas.

Anjos e Dias (2019) explicam que os estágios do ciclo de vida dos dados do DataONE podem ser adaptados à necessidade de cada pesquisador, pois eles são independentes entre si. As etapas não necessitam ser obrigatoriamente desempenhadas. Desta forma, elas não são exaustivas, mas são norteadoras para um adequado gerenciamento de dados, de acordo com a realidade de cada investigação científica.

As etapas do ciclo de vida do DataONE podem ser remetidas a uma bússola que orienta o pesquisador, de forma que a gestão empreendida nesse caminho possa gerar conhecimento. Cada etapa corresponde a um momento da pesquisa necessário para lidar com os dados. Assim sendo, por meio desse processo, é possível assegurar que o valor dos dados seja duradouro.

3.1.2 Ciclo de Vida dos Dados – *Digital Curation Centre*

Os pesquisadores, de modo geral, sejam dos campos das ciências, das artes ou das humanidades, comumente produzem dados em formatos digitais, por meio da digitalização das informações analógicas, tais quais: mapas históricos, gravações de áudio, livros, documentos, fotografias – como exemplificam Siebra, Borba e Miranda (2016) – ou da criação de novos dados nascidos digitais. Na visão de Abbott (2008), além de confiar neste processo de criação de dados, os pesquisadores passaram a dar credibilidade ao conteúdo digital originado no trabalho de outros. Neste contexto, os dados podem estar vulneráveis à obsolescência tecnológica e podem ficar em risco pelas limitações da mídia.

Para Sanchez, Vidotti e Vechiato (2017), com o crescente volume de dados gerados no ambiente da *Word Wide Web* (WWW), sejam eles criados digitalmente ou migrados para os formatos digitais, existem algumas inquietações sobre assegurar com qualidade a guarda, organização, gestão, tratamento, representação, recuperação e preservação dessas informações. Além do mais, em decorrência das transformações nas Tecnologias da Informação e Comunicação (TIC), preocupações mais atenuadas surgem a respeito da obsolescência de *hardware* e *software* utilizados para criação de sistemas informacionais e dos formatos dos objetos digitais gerados e dos dados científicos.

Para trazer solução a essa realidade, a curadoria digital compreende o gerenciamento e a preservação de dados digitais a longo prazo. No que concerne à questão do gerenciamento, engloba as diversas atividades relacionadas aos dados, desde o planejamento da sua criação, as práticas de digitalização, documentação, além de assegurar que eles estejam disponíveis de forma adequada para serem descobertos e utilizados no futuro. Outrossim, envolve o gerenciamento de vastos conjuntos de dados para o uso cotidiano, de forma que possam ser pesquisados e se mantenham legíveis (ABBOTT, 2008).

Segundo o *Oxford Dictionary*, o curador é “um detentor ou guardião de um museu ou outra coleção.”⁹ É importante notar que o conceito vem da área de museologia, logo, ao adaptar a definição para fins deste trabalho, existe uma ligação com os atos do pesquisador de cuidar, zelar, manter, organizar ou tratar de forma adequada os seus dados científicos de forma que eles não pereçam, mas estejam disponíveis para serem usados em novos estudos.

Para o DCC (2020), a definição da curadoria está atrelada a manter, preservar e proporcionar valor aos dados das investigações no percurso de todo seu ciclo de vida. O *Digital*

⁹ *A keeper or custodian of a museum or other collection.*

Curation Institute (2020, tradução nossa), da Universidade de Toronto, apresenta um conceito mais abrangente: “A curadoria digital é o envolvimento ativo no gerenciamento, incluindo a preservação, de recursos digitais para uso futuro.”¹⁰ Esta definição vai além da curadoria dos dados científicos, inclui outros ativos como jogos de computador, arte digital e registros.

Siebra, Borba e Miranda (2016, p.27) revelam que a curadoria é uma prática interdisciplinar, porém, ao refletirem curadoria digital na Ciência da Informação, mostram que nesta área ela vem sendo utilizada com o foco em três abordagens: “na filtragem e seleção de dados da *Web* (Curadoria de Conteúdo ou de Informação), na agregação de valor, gestão ativa e preservação de dados digitais (Curadoria Digital) e na curadoria de dados de pesquisa (*e-science*).” A terceira abordagem apresentada pelos autores constitui o foco deste trabalho.

O conceito de curadoria digital não se detém na preservação, ele é mais amplo, ao incluir o gerenciamento em sua prática curatorial possibilita a sustentabilidade e assegura o valor dos dados no longo prazo. Araújo *et al.* (2019) mostram que os processos de curadoria, além de envolverem as etapas de gestão e preservação, incluem também a etapa de arquivamento.

Sales (2020) acrescenta que a curadoria está relacionada à agregação de valor e que neste contexto o trabalho do bibliotecário é primordial para tratar os dados científicos e colocá-los em um repositório, de modo que estes dados possam ser disponíveis, recuperáveis e compreensíveis.

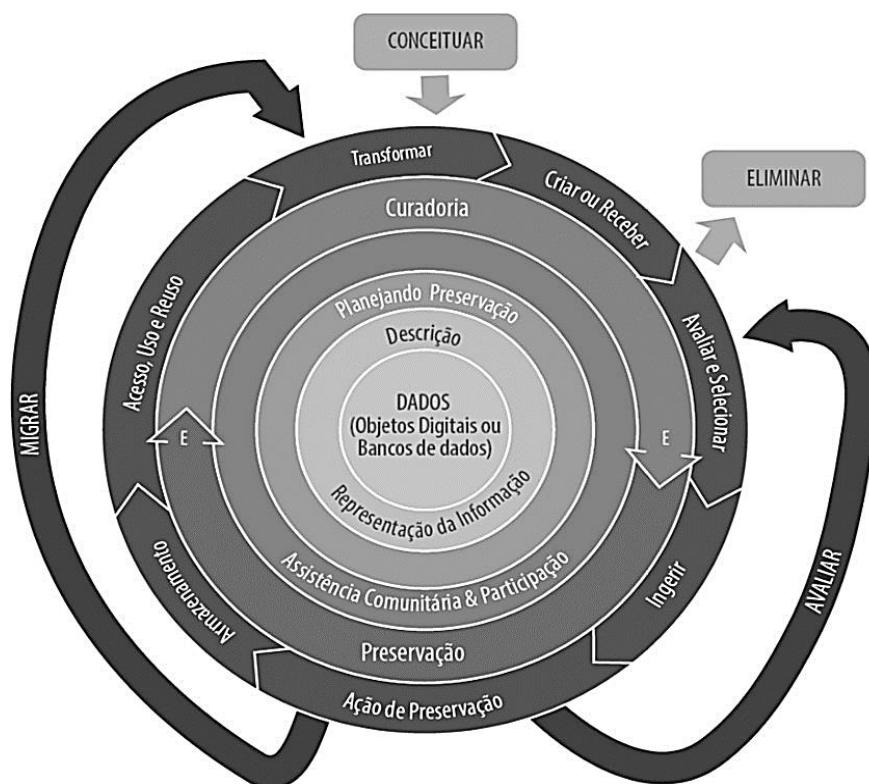
O arcabouço teórico aponta que a curadoria digital é um caminho a ser trilhado para assegurar a qualidade dos dados. Desta forma, o pesquisador precisa ter em mente que, como consequência de uma adequada curadoria, os dados poderão ser reutilizáveis e entendidos, o que garantirá a otimização dos dados.

O centro líder mundial de curadoria digital, que teve início em 2004, no Reino Unido, *O Digital Curation Centre* (DCC, 2020), apresenta o *Curation Lifecycle Model* ao buscar uma curadoria bem-sucedida e a preservação dos dados a partir da conceituação inicial ou recebimento até o ciclo de curadoria interativa. Siebra *et al.* (2013) mostram que o dado é a unidade principal de gestão do ciclo e configura-se como qualquer informação digital em código binário. Na parte central do ciclo estão os objetos digitais simples, complexos e as bases de dados. Ao redor do dado digital estão as ações que são indicadas durante o processo de curadoria. O modelo de curadoria apresenta três tipos de ações com os seus componentes. Para

¹⁰ *Digital curation is the active involvement in the management, including the preservation, of digital resources for future use.*

um melhor entendimento, serão elencadas as etapas das ações para todo o ciclo de vida dos dados de acordo com a Figura 4.

Figura 4 - Ciclo de vida dos dados do *Digital Curation Centre*



Fonte: Anjos (2019)

As ações para todo o ciclo de vida são intituladas dessa forma por corresponderem a atividades que permeiam todo o CVD. Essas ações podem ser visualizadas como anéis concêntricos rodeando os objetos de dados que se encontram no centro do modelo (SAYÃO, SALES, 2012). As ações são as seguintes:

- **Descrição e Representação Informação** – Definir os metadados administrativos, descritivos, técnicos, estruturais e de preservação, por meio de padrões apropriados, para assegurar descrição e controle a longo prazo. Definir a forma de atribuição de informações de representação apropriada para entender e contextualizar o material digital;
- **Planejamento da Preservação** - Planejar as estratégias de preservação a serem utilizadas durante todo o ciclo de vida da curadoria de material digital.
- **Acompanhamento e participação da comunidade** – motivar e envolver a comunidade interna (produtores do conteúdo/dados, desenvolvedores, pessoas da equipe) e externa (comunidade em geral, usuários da informação) no ciclo de curadoria para cooperar na

representação da informação, definir a melhor maneira de acessar aos dados, colaborar na definição de padrões, ferramentas e produtos de *software* pertinentes, entre outros;

- **Curadoria e Preservação** - Estar consciente e realizar ações administrativas e gerenciais planejadas e intencionais para contribuir com a curadoria e a preservação durante todo o ciclo de vida da curadoria. Levar em consideração aspectos de direitos autorais e permissões de acesso.

As ações sequenciais compreendem etapas que precisam ser realizadas de forma repetida para assegurar que o dado se mantenha continuamente em processo de curadoria. Tal sequência não deve ser cumprida apenas uma vez do começo ao fim, mas continuamente durante todo tempo em que o dado estiver sob curadoria, de forma cíclica, não necessariamente na ordem apresentada (SAYÃO; SALES, 2012). Destaca-se que há a possibilidade do ciclo ser personalizado às necessidades da organização/instituição e, inclusive, algumas ações podem ser omitidas. Segundo o DCC (2020), as etapas são:

- **Conceitualização** - Conceber e planejar a criação de dados com o método de captura e opções de armazenamento;
- **Criação ou recebimento** - Criar dados, acoplando a eles metadados administrativos, descritivos, estruturais e técnicos. Metadados de preservação também podem ser utilizados no momento da criação. Receber dados, segundo políticas de coleta documentadas, de criadores de dados, outros arquivos, repositórios ou centros de dados e, se necessário, atribuir metadados pertinentes;
- **Avaliação e Seleção** - Avaliar os dados e selecionar para curadoria de longo prazo e preservação, partindo de orientações documentadas, políticas ou requisitos legais;
- **Arquivamento** - Transferir dados de um arquivo, repositório, centro de dados ou outra entidade custodiante, atrelada a orientações documentadas, políticas ou requisitos legais;
- **Ação de preservação** - Realizar ações para garantir a preservação a longo prazo e a retenção da natureza autorizativa dos dados. As ações de preservação necessitam assegurar que os dados permaneçam autênticos, confiáveis e usáveis, conservando sua integridade;
- **Armazenamento** - Armazenar os dados com segurança, sob a orientação de padrões relevantes. Além de seguir rotinas de *backup*;

- **Acesso, uso e reuso** - Garantir que os dados estejam acessíveis aos usuários designados no cotidiano, adequadamente ao público-alvo;
- **Transformação** - Criar dados com base no original, por exemplo: por meio da migração para um formato diferente.

Por fim, Sayão e Sales (2012) evidenciam que as ações ocasionais, que são executadas eventualmente, podem interromper ou reordenar as ações sequenciais, mediante uma decisão. De acordo com o DCC (2020), essas ações são:

- **Descarte** - Eliminar os dados que não foram designados para curadoria e preservação de longo prazo, de acordo com as políticas documentadas, orientação ou requisitos legais ou aqueles que já atingiram seu período máximo de curadoria, de acordo com tabelas de temporalidades adotadas.
- **Reavaliação** - Retornar os dados que passaram por procedimentos de validação falhos, para uma avaliação e seleção adicional. Reavaliar todas as atividades executadas no ciclo para verificar a necessidade de ajustes/melhorias;
- **Migração** - Migrar dados para um formato diferente. O que pode ser realizado de acordo com o ambiente de armazenamento ou para assegurar a preservação dos dados, evitando a obsolescência de *hardware* ou *software*.

Silva Segundo e Araújo (2019) alertam sobre a necessidade do ciclo de vida da curadoria digital desenvolver estruturas tecnológicas que sejam capazes de suportar o volume de dados gerados, assim como recrutar colaboradores que tenham a formação específica na utilização de ferramentas de *software* que processam, analisam e compartilham dados e seus metadados. É preciso também que haja políticas públicas para lidar com a questão legal dos dados, além de recursos financeiros para financiar as estruturas, tecnologias e pessoas.

Neste contexto, Siebra, Borba e Miranda (2016) complementam que os desafios enfrentados não se limitam aos problemas técnicos, mas especialmente de ordem humana, por isso é preciso que o profissional da informação se atualize de forma adequada para lidar com o processo curatorial, além de desenvolver habilidades para trabalhar em equipes interdisciplinares e buscar apoio financeiro que contemple a necessidade da equipe e da infraestrutura, assim como possuir políticas institucionais que sejam capazes de regulamentar o trabalho que precisa ser realizado.

O processo de curadoria digital por meio do ciclo de vida dos dados indica um bom gerenciamento de dados para o pesquisador, constituindo uma ferramenta que pode ser adaptada de acordo com o seu projeto, de modo a possibilitar que os dados possam ser encontrados no futuro, dentro de um contexto que ofereça subsídio para a reutilização.

3.1.3 Ciclo de Vida dos Dados – *Data Documentation Initiative*

A *Data Documentation Initiative* (DDI) compreende um padrão internacional que busca descrever os dados oriundos de pesquisas e outros métodos de observação nas ciências sociais, comportamentais, econômicas e da saúde. O DDI consiste em um padrão gratuito que pode documentar e gerenciar distintos estágios do ciclo de vida dos dados científicos, tais como: conceituação, coleta, processamento, distribuição, descoberta e arquivamento. Neste contexto, documentar dados por meio do DDI contribui com a compreensão, interpretação e uso, tanto por pessoas como por sistemas de *software* e redes de computadores (DDI, 2020).

O DDI teve início em 1995, na cidade de Quebec, com um pequeno grupo internacional focado em metadados de ciências sociais. Posteriormente, o DDI passou a ser formado por uma Aliança, compreendendo uma nova estrutura de membros e suporte financeiro para possibilitar que a iniciativa do DDI pudesse continuar (VARDIGAN, 2013).

Desde os seus primórdios, a Aliança do DDI desenvolveu produtos de trabalhos voltados às necessidades de usuários que têm como foco a documentação dos dados e os processos que envolvem a sua coleta. O DDI – *Lifecycle* compreende um dos produtos existentes, sendo capaz de descrever além dos resultados da coleta de dados, os metadados em todo esse processo, desde a conceituação até o arquivamento dos dados resultantes. O DDI *Lifecycle* foi publicado pela primeira vez como DDI versão 3.0, em abril de 2008, e representou um grande avanço para o DDI ao incorporar totalmente os esquemas *Extensible Markup Language* – XML, passando a adotar uma abordagem de ciclo de vida dos dados. A sua versão mais recente é o DDI *Lifecycle* 3.3, publicado em 15 de abril de 2020, sendo projetado para documentar e gerenciar dados em todo o ciclo de vida, a partir da conceituação até a publicação de dados, análise e muito mais (DDI, 2020).

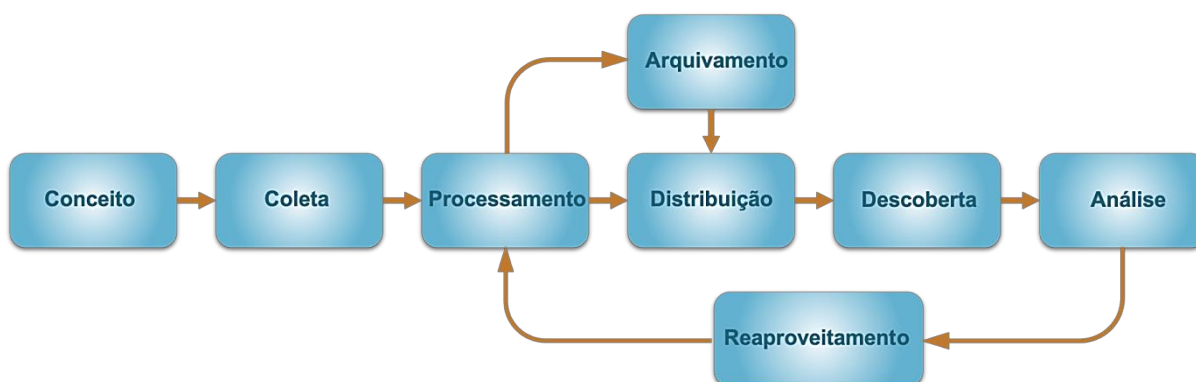
O DDI 3.0 teve como princípio basilar a reutilização de metadados para abolir custos redundantes e para apoiar comparações explícitas dentro e entre os estudos (VARDIGAN; HEUS; THOMAS, 2008). Segundo Ionescu (2007), entre 2004 e 2006 ocorreu o planejamento e o desenvolvimento da versão 3.0 do DDI, que foi projetada para cobrir todas as fases do ciclo de vida dos dados. O modelo foi fruto de uma combinação de dois ciclos de vida, o modelo de

Ann Green e Jean-Pierre Kent e o modelo de I-lin Kuo que, mesmo sendo paralelos, possuem distintas áreas e diferentes ênfases (DDI, 2004).

Vardigan, Heus e Thomas (2008) corroboram que o ciclo de vida de dados faça parte da iniciativa do DDI 3.0 e revelam que no decorrer de uma pesquisa na área de ciências sociais uma quantidade expressiva de metadados é gerada, a partir do estudo de uma articulação de conceitos associados com uma proposta para financiar a pesquisa. Envolve os esforços para a coleta de dados, a criação de um arquivo de dados e sua documentação, publicação, arquivamento e disseminação, descobertas de dados e, por fim, análise de dados que provoca novas descobertas e conhecimentos que contribuem com a literatura das ciências sociais.

Segundo Ball (2012), o modelo do ciclo de vida dos dados do DDI apresenta uma estrutura predominantemente linear, oferecendo um caminho alternativo e um *loop de feedback*, conforme pode ser representado na Figura 5.

Figura 5 - Modelo de ciclo de vida dos dados do DDI



Fonte: Adaptado e traduzido do DDI (2020)

Ball (2012) apresenta de forma detalhada as características das fases do Modelo de Ciclo de Vida Combinado DDI versão 3.0, conforme o Quadro 4.

Quadro 4 - Descrição das fases do ciclo de vida de dados do DDI versão 3.0

Fases	Descrição
Conceito de estudo	O estágio introdutório do modelo é o momento em que uma pesquisa está sendo planejada. O modelo, além de envolver a escolha da questão de pesquisa e a metodologia para coletar os dados necessários, inclui também planos para a forma como os dados serão processados, analisados e usados para responder à pergunta, e a forma que essa resposta terá. Os pesquisadores também devem definir nesta fase as relações que existirão entre os produtos de dados da pesquisa.

Coleção de dados	Exemplos de métodos e fontes de coleta de dados incluem pesquisas, censos, registros de votação ou de saúde, estatísticas de comércio ou coleções baseadas na web. As fontes de dados primárias e secundárias necessitam ser diferenciadas de forma clara.
Processamento de dados	Uma vez que os dados de entrada são reunidos, eles são processados e analisados gerando dados de saída (por exemplo, uma estatística ou conjunto dela), que respondem à problemática da pesquisa. Esses dados de saída podem ser registrados em uma forma legível por máquina ou de orientação humana, como um relatório técnico.
Distribuição de dados	Os dados são distribuídos aos usuários diretamente ou por meio de uma biblioteca ou arquivo de dados.
Arquivamento de dados	Com a finalidade de garantir o acesso a longo prazo aos dados, eles devem ser transferidos para um arquivo, em vez de serem apenas mantidos por pesquisadores. O arquivo, além de preservar os dados (e metadados), agrega valor a eles ao longo do tempo.
Descoberta de dados	Os dados podem ser divulgados através de livros, publicações em periódicos, páginas de <i>Web</i> ou outros serviços <i>online</i> .
Análise de dados	Os dados podem ser usados por outros dentro dos limites da conceituação original; por exemplo, ao optar por estatísticas importantes para um relatório de pesquisa.
Reaproveitamento	Os dados também podem ser usados dentro de uma estrutura conceitual diferente; exemplos envolvem amostragem ou reestruturação dos dados, combinando os dados com outros conjuntos semelhantes, ou produção de materiais pedagógicos.

Fonte: Ball (2012)

A fase conceitual envolve objetivo / esboço do estudo, conceitos, população de estudo, autor(es) e fontes de financiamento. Momento de planejar o estudo. As fases de coleta e processamento de dados estão relacionadas à metodologia de coleta de dados: amostragem, tempo, etc. Envolvem ainda as características do instrumento, questionário, questões referentes à limpeza de dados, ponderação, codificação, entre outros. Este é um momento de realizar o que foi proposto no planejamento. A fase de distribuição está relacionada com a publicação dos dados, refere-se ao conteúdo intelectual: variáveis, categorias, códigos e representação física como formato de dados, estrutura de registro e estatísticas, consiste ainda em um momento de fazer checagens para a próxima fase, que é o arquivamento. Esta fase envolve participações, disponibilidade e condições de acesso. Para proporcionar visibilidade, temos as fases de descobrir e analisar, em que o DDI reúne todas as fases do ciclo de vida, obtém sua própria identidade como objeto e torna-se uma ferramenta para descoberta e análise de dados; e a fase de reuso, que está atrelada ao uso secundário de dados, com a formação de uma nova estrutura

conceitual. Esta nova instância do DDI promove um novo propósito, novo produto lógico e nova descrição física de dados (IONESCU, 2007).

Kanjala *et al.* (2017) esclarecem que DDI - *Lifecycle* foi desenvolvido a partir da ideia de que um conjunto de dados representa o processo que o resultou. Assim, utiliza-se o ciclo de vida de dados como um modelo conceitual. Este compreende módulos que são pacotes de metadados, cada um correspondendo, aproximadamente, a um estágio no ciclo de vida dos dados.

De acordo com Thomas (2013), o DDI - *Lifecycle* (DDI-L) foi projetado para apoiar processos e comparabilidade no decorrer de todo o ciclo de vida dos dados, o que culminou em um padrão que tem como base a reutilização planejada de metadados, que descreve as metodologias adotadas, os processos de captura, limpeza e modificação de dados, além de envolver os diversos formatos de armazenamento dos próprios dados. Esta realidade contribui com o controle de qualidade da investigação ao suportar um processo de pesquisa orientado por metadados.

Stent (2018), ao enfatizar uma estrutura de governança de dados desenvolvida na Nova Zelândia, mostra que os padrões e as declarações de qualidade do DDI são capazes de registrar aspectos decisivos durante o ciclo de vida dos dados.

Segundo Sant'Ana (2013), o modelo do ciclo de vida dos dados proposto pelo DDI foi tomado como base para criação de outros modelos, como a proposta da *University College London* - UCL (2012). Este modelo já inclui na sua figura as principais tarefas inerentes a cada fase, constituindo um diferencial ao comparar com a proposta do DDI e apresenta um enfoque em dados coletados através de entrevistas. Ainda segundo o autor, o modelo do DDI faz uso da gestão de dados de bibliotecas, a exemplo da *Massachusetts Institute of Technology* –MIT, sendo considerada uma temática de interesse da CI.

De acordo com Gupta e Müller-Birn (2018), o modelo de ciclo de vida dos dados do DDI inclui de forma efetiva diversos aspectos da pesquisa, porém apresenta algumas limitações, tais como: o modelo tem o foco na coleta de dados primários, o alerta é que o pesquisador pode querer usar dados que já existem e que estejam armazenados no arquivo. Outra deficiência apresentada é que não existe ênfase na publicação de dados que foram analisados, o que poderia ser arquivado para ser utilizado em futuras pesquisas. E, por fim, a agilidade desse modelo na perspectiva dos pesquisadores é limitada, uma vez que o pesquisador necessita passar pelas etapas mesmo que não seja exigido passar por alguma etapa em particular. Isso demanda um maior tempo para realizar a mesma tarefa e, portanto, compromete a eficiência do modelo.

O modelo apresenta limitações que foram abordadas no presente estudo, como a identificada por Sant’Ana (2013), de não ter as tarefas primordiais expressas já no diagrama. Também possui as mencionadas anteriormente por Gupta e Müller-Birn (2018), uma vez que que os pesquisadores podem ter interesse em fazer o reuso de dados, os dados de saída também podem ser utilizados por outros e o *design* do modelo pode causar um atraso na agilidade do processo. Além dessas, não foi possível identificar no próprio *site* do DDI a descrição das fases do ciclo de vida dos dados, o que prejudica o entendimento do significado de cada uma delas de acordo com a própria iniciativa. Mesmo diante desses apontamentos, é evidente a importância do modelo do DDI no cenário da pesquisa científica ao apresentar um foco voltado para o uso eficiente dos metadados, especialmente na gestão de dados das ciências sociais, sendo também considerado um parâmetro para a criação de outros modelos.

De modo geral, o capítulo 3 (três) explorou a ferramenta CVD com a identificação de diversas versões encontradas na literatura. Em seguida, houve o detalhamento individual e ilustrativo das iniciativas do DataONE, DCC e da DDI para ser estabelecida uma conexão com a construção do próximo capítulo que busca entender o caminho percorrido pelos dados nos percursos metodológicos das investigações das ciências sociais.



4 A METODOLOGIA NO CONTEXTO DOS TRABALHOS CIENTÍFICOS NA ÁREA DAS CIÊNCIAS SOCIAIS

A pesquisa em ciências sociais está relacionada ao estudo do comportamento humano, referente ao passado, ao presente ou ao futuro – o que não é algo fácil por se tratar da complexidade das pessoas e das instituições sociais. Esta realidade remete à necessidade de adotar um procedimento metodológico, sendo fundamental a experiência acadêmica para projetar o estudo, para escolher os métodos de população e de amostragem pertinentes, e para limpar, analisar e interpretar os dados resultantes (BORGMAN, 2015).

É possível observar que a fase metodológica do trabalho científico envolve ações relacionadas aos dados para entender melhor a ligação existente nessa realidade. Para tanto, são necessárias algumas reflexões que esclareçam o que é a metodologia no contexto da pesquisa e a relação dos dados nesse processo.

A metodologia de um trabalho acadêmico exige do pesquisador empenho e atenção. Além de ser uma descrição formal de métodos e técnicas a serem empregados, aponta as conexões e a leitura operacional realizada pelo pesquisador na bibliografia consultada, assim como os seus objetivos (DESLANDES, 2009).

Para Minayo (2009), a metodologia envolve concepções teóricas da abordagem, relacionando-se com a teoria, com a realidade empírica e com os pensamentos acerca da realidade, ou seja, vai além das técnicas. Desta forma, é possível observar que a teoria e a metodologia caminham de mãos dadas.

Gondim e Lima (2006) aprofundam o diálogo e revelam que a metodologia aponta as questões norteadoras e as estratégias que serão adotadas para a abordagem empírica do objeto. Tais questões, que de forma implícita ou explícita já são encontradas na definição do objeto, devem ser redirecionadas em termos de estratégias metodológicas para serem seguidas, relacionando-se com o referencial teórico. De modo geral, a metodologia está presente desde o início do projeto de pesquisa, uma vez que não é fácil separar o que fazer de como fazer. Assim:

Pode-se entender metodologia como um caminho traçado para se atingir um objetivo. É a forma, o modo de se resolver problemas e buscar respostas para necessidades e dúvidas. A metodologia científica é a busca da verdade num processo de pesquisa ou aquisição de conhecimento; um caminho que utiliza procedimentos científicos, racionais e critérios normalizados e aceitos pela ciência. É o corpo de regras e diligências confiáveis estabelecidas para se realizar uma pesquisa. (MICHEL, 2015, p.34).

Neste contexto, Richardson (2017) alerta que há diferença entre o método e metodologia, uma vez que a palavra método é originada do grego *méthodos* (meta = além de, após de + ódos = caminho). Desta forma, partindo da sua procedência, método é o caminho ou a maneira para chegar a uma determinada finalidade, diferenciando-se do conceito de

metodologia, que deriva do grego *méthodos* (caminho para chegar a um objetivo) + *logos* (conhecimento). Logo, para o autor, o método científico é o percurso da ciência para atingir um objetivo, já a metodologia vai além, compreende o estudo do método, envolve os procedimentos e regras usados por determinado método. Por exemplo: a necessidade de observar, a necessidade de formular hipóteses, a elaboração de instrumentos.

Para Minayo (2009), a metodologia inclui, de forma concomitante, a teoria da abordagem (o método), os instrumentos para operacionalizar o conhecimento (técnicas) e a criatividade do pesquisador (suas vivências, sua capacidade e sensibilidade). A este respeito, Michel (2015) acrescenta que atitude, método e disciplina são fundamentais para atingir os objetivos da pesquisa.

Portanto, os autores supracitados são unânimes em considerar que os procedimentos metodológicos, por meio de métodos e técnicas aprovados pela ciência e através das habilidades do pesquisador, buscam trazer respostas às problemáticas das investigações científicas, cujo processo vai além de apontar o caminho, envolve a aquisição de conhecimentos e está ancorado no aporte bibliográfico e na definição de objetivos da pesquisa. Nesse momento, também é possível visualizar que existe uma relação intrínseca entre a metodologia e o caminho que os dados percorrem para trazer resultados à pesquisa.

4.1 A INVESTIGAÇÃO CIENTÍFICA E A PRESENÇA DOS DADOS NOS PERCURSOS METODOLÓGICOS

A pesquisa científica é a busca da realidade, sua atividade está atrelada a descobrir a verdade e a essência das coisas, é um processo inacabável, intrinsecamente processual. A Pesquisa compreende um fenômeno de busca de conhecimento com aproximações sucessivas e não uma situação definitiva, pois diante dela pode sempre existir algo a descobrir (MICHEL, 2015).

Gil (2019) enfatiza a pesquisa social como um conjunto de procedimentos que, por meio da utilização de métodos científicos, tem a finalidade de obter novos conhecimentos no campo da realidade social, a qual inclui todos os aspectos inerentes às pessoas em seus diversos relacionamentos com outras pessoas e com organizações sociais.

Uma pesquisa científica pode ocorrer, por exemplo, na construção de uma tese acadêmica com vista à obtenção do título de doutorado. Nesse cenário, Gondim e Lima (2006) mostram que a tese acadêmica requer o aprofundamento de procedimentos teóricos-metodológicos, além do levantamento de proposições originais a serem pesquisadas, cuja

originalidade está relacionada à utilização de novas abordagens ao analisar os problemas, a recomendação de questões inéditas e levar em consideração elementos que não foram considerados em outros estudos. Desta forma, a originalidade não se limita a estudar algo absolutamente novo ou desconhecido.

É evidente o caráter formal e sistemático da pesquisa, por isso a investigação científica necessita de planejamento e de um documento escrito que aponte de forma descritiva os elementos fundamentais da pesquisa, a escolha do assunto, a tematização, a determinação dos objetivos, a coleta de dados, sua análise e interpretação, até as possíveis contribuições do trabalho. Esse plano deverá responder a perguntas, tais como: O que será pesquisado? (Assunto ou tema); Por que será pesquisado? (Justificativa ou introdução); Para que será pesquisado? (Objetivos); Como será pesquisado (Metodologia); Quando será pesquisado? (Metodologia) (RICHARDSON, 2017).

É possível observar que, nesse processo de planejamento, dois questionamentos fazem parte do procedimento metodológico do trabalho científico, na busca por apontar o caminho de como será realizado o trabalho e o seu período. De forma mais detalhada, veremos as fases desse momento mais adiante.

Michel (2015) acrescenta que é importante a utilização de uma metodologia adequada à pesquisa, que leve em consideração alguns aspectos como: o planejamento, a divisão da pesquisa em etapas, as atividades específicas de cada etapa, os recursos que serão usados, um cronograma de controle e os objetivos da investigação científica.

Deslandes (2009), ao apresentar um foco em pesquisa qualitativa, revela que faz parte da seção dos procedimentos metodológicos a descrição da fase de exploração de campo (decisão sobre o espaço da pesquisa, critérios e estratégias para a escolha de grupo/sujeitos da investigação, a definição de métodos, técnicas e instrumentos para produção dos dados e os mecanismos para adentrar em campo), as etapas do trabalho de campo e os procedimentos para análise. Com base nessas fases apontadas pela autora, às próximas seções estão adaptadas para um melhor entendimento das fases metodológicas que envolvem a obtenção, a execução e os procedimentos para analisar os dados oriundos das investigações científicas.

4.1.1 Descrição da Fase de Planejamento dos Dados

Este é um momento de planejar os procedimentos para a coleta e análise dos dados, momento de decisão sobre o tipo da pesquisa, o campo que será estudado (sujeitos, locais, amostragem), a seleção de métodos e técnicas para posteriormente executar a coleta dos dados.

Michel (2015), ao tratar do êxito da pesquisa, recomenda o planejamento da coleta dos dados, de forma organizada, que siga um cronograma, que envolva uma fase de registro e análise dos dados, levando em consideração aspectos como local, hora, responsáveis, ou seja, o desenvolvimento de um plano tático de ação para a realização da pesquisa.

No processo de preparação da coleta de dados, é importante decidir sobre a identificação e escolha das pessoas que trarão respostas às perguntas levantadas no projeto do relatório de pesquisa ou que trarão respostas por meio da observação dos seus comportamentos. A partir deste estágio, é possível obter uma série de informações (números, respostas, opiniões, citações). Esta etapa faz parte da seção dos procedimentos metodológicos. Neste contexto, o pesquisador precisa levantar alguns questionamentos para a obtenção de respostas, assim precisa escolher onde vai ser realizado o estudo, quem fará parte dele, de que forma vai obter permissão dos participantes e quais os dados serão escolhidos e como reunirá os dados. Neste momento, o pesquisador diz o período da coleta de informações e fala sobre a colaboração dos entrevistados (RICHARDSON, 2017).

Segundo Borgman (2015), geralmente devem ser feitas escolhas difíceis a respeito dos dados que podem ou devem ser coletados, de que forma serão obtidos e o porquê, assim como a maneira que eles podem ser relatados e divulgados. Principalmente por se tratar de um campo que pode envolver humanos na pesquisa, é necessário levar em consideração as questões éticas, tendo em mente que os dados sobre pessoas estão submetidos a diversas regras e regulamentos, a respeito de quais dados podem ser coletados e em que circunstâncias.

Michel (2015) sugere a organização, que envolve uma descrição precisa correspondente às variáveis definidas na população da pesquisa: o objeto de pesquisa (pessoas, instituições, variáveis elegidas), as técnicas de coletas de dados, e a definição da população e da amostra.

De acordo com Gil (2019), comumente as pesquisas sociais podem envolver um imenso universo de elementos, sendo difícil analisar a sua totalidade. Por este motivo, é comum neste tipo de pesquisa trabalhar com uma parcela dos elementos que fazem parte do universo, ou seja, a amostra. Nesse caso, a amostragem compreende um processo de escolha de um grupo de pessoas, eventos, comportamentos ou demais elementos de um universo a ser pesquisado. Na literatura existem diferentes tipos de amostragem, que podem ser selecionados pelo pesquisador de modo mais adequado a sua investigação.

Richardson (2017) ratifica a ideia ao mostrar que nas ciências sociais, por se estudar grupos humanos, é possível observar a diversidade de seus membros, o que requer a adoção de um procedimento que possibilite um controle mais efetivo. Assim sendo, o autor enfatiza que

as técnicas de amostragem direcionam a seleção das amostras pertinentes para os propósitos de investigação.

Para Sampieri, Collado e Lúcio (2006), o projeto, os dados coletados, a forma de obtê-los, a amostragem e outros elementos do processo de pesquisa são classificados nos tipos de estudos: exploratórios, descritivos, correlacionais e explicativos, tanto no que tange a uma pesquisa qualitativa como a uma quantitativa ou mista. Uma mesma pesquisa pode envolver componentes de mais de um desses 4 (quatro) tipos.

Richardson (2017) mostra que a pesquisa exploratória busca uma visão abrangente do objeto em investigação; a descritiva tenta descrever de forma sistemática uma situação, problema, fenômeno ou programa na busca por mostrar a estrutura do comportamento de um fenômeno; a explicativa procura deixar claro o motivo e de que forma é produzida uma relação envolvendo dois aspectos de uma situação ou fenômeno; e, por fim, a pesquisa correlacional tem a finalidade de medir o nível de relação existente entre dois ou mais conceitos ou variáveis de uma determinada situação ou fenômeno.

A seleção dos métodos e técnicas a serem utilizados na investigação científica pode ocorrer com base na proposição do problema, na formulação das hipóteses e na limitação do universo e da amostra, o que denota uma ligação com o problema a ser pesquisado (MARCONI; LAKATOS, 2019).

As técnicas compreendem ferramentas fundamentais para a pesquisa e requerem atenção na sua elaboração para assegurar que os resultados sejam alcançados com fidelidade, qualidade e completude. Sua elaboração e aplicação devem estar ancoradas em critérios técnicos rigorosos para não prejudicar os dados e, conseqüentemente, o propósito da pesquisa. As técnicas normalmente utilizadas na coleta de dados na pesquisa social são: a análise de documentos da empresa ou grupo social, a observação pessoal, as diversas modalidades de entrevista e o questionário. Cada uma dessas técnicas tem o seu objetivo, um público específico e, cada uma, a sua maneira, agrega valor ao processo de análise (MICHEL, 2015).

De acordo com Deslandes (2009), o uso das técnicas pode se voltar para a construção primária dos dados, ou seja, quando o pesquisador cria o dado ao interagir diretamente com os sujeitos por meio de entrevistas, observações, aplicação de questionário. Assim como na procura de dados secundários, isto é, com base em acervos já existentes, como: documento, banco de dados, revistas, jornais, entre outros.

Michel (2015) ainda acrescenta os dados terciários, que são oriundos de terceiros, citados ou fornecidos por outra pessoa. Neste caso, não se teve acesso à fonte original para sua devida comprovação (ex.: notas de aula, citações de conferências, palestras).

As ciências sociais trabalham com uma variedade de métodos. Segundo Gil (2019, p.9), “os métodos são classificados em dois grandes grupos: o dos que proporcionam as bases lógicas da investigação científica e o dos que esclarecem acerca dos procedimentos técnicos que poderão ser utilizados.” Para o autor, no primeiro grupo, os métodos são desenvolvidos com um grande nível de abstração, que permitem ao pesquisador tomar decisões em relação ao alcance da sua pesquisa, das regras de esclarecimento dos dados e da validade de suas generalizações. Os métodos seguintes fazem parte deste grupo: dedutivo, indutivo, hipotético-dedutivo, dialético e fenomenológico.

Gil (2019), ao falar do segundo grupo, revela que a sua finalidade é disponibilizar uma direção para a realização da pesquisa social, principalmente no que se refere à obtenção, processamento e validação dos dados ligados à problemática da pesquisa. Dentre os métodos pertencentes a este grupo, Gil (2019, p.15) mostra que “os mais adotados nas ciências sociais são: o experimental, o observacional, o comparativo, o estatístico, o clínico e o monográfico.” Para fins pretendidos neste trabalho, os métodos apresentados nos dois grupos não serão descritos de forma detalhada, apenas apresentada a sua classificação a título de informação.

De modo geral, utiliza-se mais de um método e até uma combinação de dois ou mais dele. A organização desse instrumental da pesquisa comumente necessita de tempo, mas é um momento importante a ser planejado (MARCONI; LAKATOS, 2019).

No que se refere à articulação das abordagens quantitativas e qualitativas, Michel (2015) mostra que a pesquisa qualitativa focaliza na realidade, busca trazer significados ou interpretação aos fenômenos segundo o ponto de vista do que o sujeito do contexto estudado enxerga. Busca também explorar as diferentes opiniões e as distintas representações sobre a temática pesquisada. Nesta perspectiva, o pesquisador participa, compreende e interpreta. Já a pesquisa quantitativa trabalha com o princípio de que tudo pode ser traduzido em números, sendo aplicado tanto na coleta como no tratamento dos dados, por meio de técnicas estatísticas simples e complexas. Neste aspecto, o pesquisador descreve, explica e prediz os resultados.

É possível observar que a fase de descrição da construção dos dados envolve o planejamento. Este direciona a pesquisa para alcançar o seu objetivo, com base em um rumo previamente traçado, que pode ser flexível, não ocorrendo de forma aleatória, e compreende um momento de decisão com base nos recursos disponíveis como tempo, assim como as pessoas que irão realizar o trabalho.

4.1.2 Descrição da Fase de Coleta dos Dados

No processo de levantamento dos dados, pode se fazer uso de três procedimentos: pesquisa bibliográfica, pesquisa documental e contatos diretos ou de campo. Para Gil (2019), quando a pesquisa é de natureza bibliográfica, os dados são levantados através da leitura de livros, artigos de periódicos, anais de eventos e impressos diversos. Para o autor, a segunda natureza pode acontecer mediante a exploração de arquivos, análise de documentos ou análise de artefatos físicos.

No que se refere à pesquisa de campo, os contatos diretos são feitos com pessoas que podem disponibilizar dados ou apontar possíveis fontes de informações relevantes, em que a pesquisa bibliográfica pode ser realizada juntamente com a pesquisa de campo (MARCONI; LAKATOS, 2019). Nesse contexto, Gil (2019) revela que a obtenção dos dados acontece no local em que os fenômenos ocorrem de forma espontânea, por meio de observação, aplicação de questionários e entrevistas.

Diante dos apontamentos dos autores supracitados (GIL, 2019; MARCONI; LAKATOS, 2019), os dados podem ser alcançados através de naturezas ou procedimentos diferentes, ou seja, a partir de fontes distintas que podem até ser combinadas, cuja escolha fica a critério do pesquisador.

A fase de coleta está atrelada à execução da pesquisa. Segundo Michel (2015), ela deve cumprir e acompanhar o cronograma pré-estabelecido, alinhada com o plano de ação. Neste momento, é primordial que os dados levantados sejam armazenados à medida que forem obtidos, mesmo que provisoriamente e sem formato deliberado. Assim sendo, não se deve adiar o registro dos dados, o que denota que eles necessitam ser preservados. Uma questão relevante a destacar é que para serem reutilizados em novos estudos, a qualidade dos dados precisa ser assegurada.

Segundo Marconi e Lakatos (2019), a coleta constitui o momento da pesquisa em que se introduz a aplicação dos instrumentos elaborados e das técnicas selecionadas com a finalidade de obter os dados previstos. Esta situação requer tempo do pesquisador, que precisa ser atencioso com o registro dos dados e ter sido preparado anteriormente. Portanto, é preciso realizar o trabalho em tempo hábil, não ultrapassar o orçamento previsto e preparar o pessoal, assim, quanto mais planejamento, melhor. Esta realidade demonstra que a fase de obtenção dos dados está intrinsicamente ligada ao planejamento.

Segundo Michel (2015), para que os dados brutos coletados se transformem em resultados da pesquisa relevantes, é preciso adotar procedimentos que permitam ao pesquisador

sistematizar, categorizar e analisar os dados por meio das técnicas, o que será enfatizado no próximo tópico.

É possível visualizar que os dados comumente são originais da própria pesquisa, porém uma realidade que deve ser enfatizada é que eles podem ser encontrados em outros estudos e, conseqüentemente, integrados à pesquisa, caso estejam disponíveis de forma contextualizada e atendam às necessidades do pesquisador.

4.1.3 Procedimentos de Análise dos Dados

Segundo Gil (2019), para a obtenção do significado dos dados requeridos pela pesquisa, se faz necessário passar pelo processo de análise e interpretação. Desta forma, é pertinente que sejam tabulados, resumidos, organizados e apresentados em tabelas, gráficos ou diagramas. Richardson (2017) acrescenta que esta etapa de analisar os dados levantados fornece as respostas para a problemática da pesquisa, gerando, assim, os resultados da investigação.

Uma questão destacada por Gil (2019) é que uma forma de prosseguir com a análise dos dados está relacionada ao confronto dos dados oriundos da pesquisa com outros dados, provenientes de arquivos ou de pesquisas realizadas. Diante desta recomendação, é possível observar a necessidade da disponibilização dos dados de uma pesquisa para servir como parâmetro para outros estudos, assim como para atestar a sua veracidade.

É preciso ainda que os dados levantados sejam interpretados e analisados à luz de alguma teoria. Este fato possibilita uma interpretação mais robusta, pois uma das mais importantes funções da teoria na pesquisa é a de certificar que os dados disponham de uma maior significância (GIL, 2019).

Richardson (2017) alerta que, se for realizada uma análise quantitativa, é necessário apontar como os dados serão tratados: tabelas, gráficos e testes estatísticos. Caso seja por meio da análise qualitativa, explicitar as técnicas adotadas: tipo de análise (documentário, de conteúdo ou histórico).

Dietrich, Loison e Roupnel (2015) mostram que para se analisar os dados que foram coletados por meio do método quantitativo, os conhecimentos qualitativos oferecem esclarecimentos que possibilitam interpretar e complementar os dados estatísticos. Por outro lado, os processos sociais e as relações que as pesquisas qualitativas estão limitadas a descobrir podem ser evidenciados pela calibragem por dados cifrados. Desta forma, o interesse do método “de dupla cabeça” é, sem sombra de dúvida, o de dispor de uma “dupla prova”, ou seja, o de ratificar com um método os resultados oriundos do outro (“triangulação”). Isso contribui com

a qualidade do trabalho científico, possibilitando ao pesquisador um melhor entendimento do seu objeto de estudo.

Michel (2015) corrobora ao enfatizar que um dado estatístico pode possibilitar que a análise qualitativa se torne mais robusta e, simultaneamente, a confiabilidade dos dados estatísticos pode ser auxiliada por uma análise crítica de um dado.

Existem vários métodos de pesquisa usados na literatura para a análise e interpretação de textos. A análise qualitativa do conteúdo, a etnografia, a teoria fundamentada, a fenomenologia e a pesquisa histórica exemplificam esta realidade. A análise do discurso compreende uma ferramenta de interpretação fundamental nas Ciências Sociais (RICHARDSON, 2017).

Neste contexto, o procedimento metodológico utilizado incentiva o pesquisador a refletir sobre o trabalho, levantando questionamentos no que diz respeito a sua prática de campo e sobre os resultados obtidos, buscando equilibrar posicionamento que muitas vezes são considerados opostos (DIETRICH; LOISON; ROUPNEL, 2015).

Ao fazer um paralelo dos elementos dos ciclos de vida dos dados das iniciativas do DataONE, DCC e DDI, enfatizados no capítulo anterior, com as subseções prévias do presente capítulo: “Descrição da fase de planejamento dos dados”, “Descrição da fase de coleta dos dados” e “Procedimentos para a análise dos dados”, é possível identificar uma relação do procedimento metodológico com as fases dos CVDs. Essa realidade está expressa no Quadro 5, que também se caracteriza como uma contribuição da pesquisa ao ilustrar as fases encontradas de forma mais evidente na literatura que trata das ciências sociais.

Os trabalhos de Gil (2019), Marconi e Lakatos (2019), Michael (2015) e Richardson (2017) foram tomados como base para a construção do Quadro 5. Esses foram selecionados por discutirem os procedimentos metodológicos aplicados a grande área das ciências sociais, a qual a CI está inserida. Além disso, após a conclusão do presente capítulo teórico, foi feita uma leitura do material e agrupadas as partes que apresentaram alguma similaridade com os CVDs. O que possibilitou identificar a relação dos conteúdos dos autores com os elementos das iniciativas do DataONE, DCC e DDI.

Quadro 5 - Fases metodológicas dos dados e os ciclos de vida do DataONE, DCC e DDI

Literatura e modelos	Fases evidentes			Fases pouco evidentes	
	Planejamento dos dados	Coleta dos dados	Análise dos dados	Preservação dos dados	Integração dos dados
Michael (2015)	Decidir sobre: local, hora e	Acompanhar o cronograma e o plano de ação.	-	Dados armazenados	

	responsáveis pela coleta.				
Richardson (2017)	Decidir sobre as pessoas que serão questionadas ou observadas, local do estudo e responsáveis; Levar em consideração a permissão dos participantes e de que maneira os dados serão obtidos e a definição do período de coleta.	-	Tabelas, gráficos e testes estatísticos; Aplicação de análise de conteúdo, etnografia, teoria fundamentada, fenomenologia, pesquisa histórica e análise do discurso.	-	-
Gil (2019)	Seleção de amostras e Seleção dos métodos	Observação; Aplicação de questionários e entrevistas.	Dados em tabelas, gráficos ou diagramas; Dados interpretados à luz da teoria.	-	Confrontar dados com dados de outras pesquisas
Marconi e Lakatos (2019)	Seleção dos métodos e técnicas	Realizar pesquisa bibliográfica, documental, contatos diretos ou de campo; Aplicação dos instrumentos e das técnicas escolhidas e Alinhar a fase com o planejamento.	-	Registro de dados	-
DataONE (2020)	Planejar	Coletar	Analisar	Preservar	Integrar
DDI (2020)	Conceito de estudo	Coleção de dados	Análise de dados	Arquivamento de dados	Reaproveitamento
DCC (2020)	Descrição e Representação da informação; Planejamento da preservação; Curar e preservar e Acompanhamento e participação da comunidade	Conceitualização; Criação ou recebimento e Transformação	Reavaliar	Curadoria e preservação; Arquivamento; Ação de preservação e Armazenamento	Criação e recebimento

Fonte: Dados da pesquisa (2021)

De forma parcial, no mesmo Quadro 5 também estão elencados alguns momentos correspondentes às fases de preservação e integração dos dados que foram pouco evidenciadas e que, pela limitação de conteúdo na literatura das ciências sociais, não foram criadas subseções com a argumentação sobre elas. Desta forma, é possível observar que algumas ações ligadas ao planejamento, à coleta e à análise foram mais facilmente destacadas no discurso dos autores das ciências sociais, ao passo que outras aparecem em menor proporção, como a integração e a

preservação e algumas que não são notórias, a exemplo de assegurar a qualidade dos dados e descrever com a utilização de metadados adequados e descobrir, de forma que os dados possam ser encontrados. Logo, isso é algo necessário a ser explorado, por fazer parte do caminho trilhado pelos dados e eles estarem diretamente integrados às metodologias científicas.

No presente capítulo, observou-se a relação dos dados com o percurso metodológico de uma pesquisa científica, especificamente no contexto das ciências sociais. Nesse momento, diversas ações adotadas definem o caminho dos dados, principalmente, ao enfatizar o planejamento de como os dados serão obtidos, a coleta propriamente dita e as decisões sobre como esses dados serão analisados.



5 PROCEDIMIENTOS METODOLÓGICOS

Os procedimentos metodológicos adotados no presente estudo foram fundamentais para alcançar os objetivos propostos referentes ao problema de pesquisa. Por meio do caminho que foi percorrido e desvendado, buscou-se adquirir novos conhecimentos que possibilitaram o aprofundamento e reflexões acerca da temática investigada.

Neste sentido, a criação da própria pesquisa está atrelada à busca de dados adequados a questões específicas, com vistas à construção do objeto de estudo e suas hipóteses com os dados ajustados sob medida (DIETRICH; LOISON; ROUPNEL, 2015). Nesse processo de construção da própria pesquisa, almeja-se a criação de uma nova base de dados que seja útil para este trabalho e que sirva como insumo para novos estudos que tenham interesse de investigar sobre o assunto.

Para a realização da pesquisa, foram levantadas as teses brasileiras da Ciência da Informação que estavam disponíveis nos *sites* dos seus respectivos programas, nos repositórios institucionais e em uma biblioteca digital. Posteriormente, os procedimentos metodológicos foram identificados para analisar, em seus percursos, os elementos dos ciclos de vida dos dados.

O planejamento da pesquisa passou pela elaboração de um plano de trabalho e um cronograma com etapas e datas definidas para o desenvolvimento da tese. Já o planejamento da obtenção dos dados ocorreu por meio de um plano de gestão de dados- PGD, baseado na iniciativa *DMP online* do DCC, conforme apresentado no Apêndice A, contendo informações relevantes como: a coleta de dados, documentação e metadados, questões éticas, de armazenamento, questões de compartilhamento e sobre responsabilidades e recursos investidos.

A princípio, o trabalho passou por uma avaliação da banca de qualificação do doutorado. A partir das recomendações sugeridas na reunião e de discussões com o orientador, foram realizados ajustes na tese, como a inclusão do ciclo de vida dos dados da iniciativa do DDI, por ser voltado para a realidade das ciências sociais. Desse modo, novos debates passaram a fazer parte do trabalho, assim como ocorreu a eliminação de alguns pontos, a fim de contribuir com a qualidade do estudo em questão.

5.1 TIPO E NATUREZA DA PESQUISA

Na busca por identificar os elementos componentes dos ciclos de vida dos dados, pertencentes às iniciativas do DataONE, DCC e DDI, nas teses brasileiras da CI, tendo como base a teoria sobre a temática e os dados que se buscam alcançar diante da problemática em questão, a presente pesquisa caracteriza-se como descritiva, bibliográfica, documental e qualitativa.

Sampieri, Collado e Lúcio (2006) revelam que a pesquisa descritiva mede, avalia e coleta dados sobre vários aspectos, dimensões ou componentes do fenômeno a ser estudado. Logo, a pesquisa em questão busca descrever os elementos dos ciclos de vida dos dados a serem encontrados no contexto dos procedimentos metodológicos das teses brasileiras da Ciência da Informação.

A pesquisa bibliográfica foi realizada para aprofundar as definições conceituais sobre os dados científicos e discutir assuntos basilares nessa tese, como os ciclos de vida dos dados e as metodologias no contexto das ciências sociais, para tanto, foi realizada uma investigação na literatura pertinente (MARCONI; LAKATOS, 2019). Nesse contexto, foram consultados artigos de periódicos, livros, relatórios, teses, dissertações, *sites* institucionais, anais de eventos, dentre outros materiais já publicados acerca da temática. As fontes utilizadas para a construção do estado da arte corresponderam às seguintes bases de dados do campo da CI: Base de Teses e Dissertações da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (BDTD/Capes); Base dos Encontros Nacionais de Pesquisa e Pós-Graduação em Ciência da Informação (BENANCIB); Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI); Library, Information Science & Technology Abstracts (LISTA); Library and Information Science Abstracts (LISA); Web of Science (WoS) e Scopus.

A pesquisa bibliográfica também foi utilizada para a obtenção dos dados, pois foram investigadas as teses da CI dos anos de 2012-2020, conforme será especificado adiante. Assim como também é considerada uma pesquisa documental, uma vez que os percursos metodológicos das teses da CI compreendem documentos que deram suporte à pesquisa (MARCONI; LAKATOS, 2019).

Os métodos quantitativos e qualitativos foram articulados com o propósito de obter um olhar duplo para o objeto de estudo, visto que quando os dois métodos são usados de forma complementar, eles se nutrem de forma recíproca (DIETRICH; LOISON; ROUPNEL, 2015).

Para assegurar a precisão dos resultados, a pesquisa configura-se como quantitativa, pois técnicas estatísticas foram utilizadas para analisar e classificar as informações levantadas, sendo expressas em números (KAUARK; MANHÃES; MEDEIROS, 2010). O estudo também é qualitativo por ir além das representatividades numéricas, buscando esclarecer os aspectos reais, que não podem ser quantificáveis. (SILVEIRA; CÓRDOVA, 2009). Esta realidade possibilitará uma visão mais abrangente sobre o estudo. Para fins do presente trabalho, a pesquisa assume de maneira predominante um enfoque qualitativo, sendo considerada quali-quantitativa, já que quantificações foram utilizadas para auxiliar os dados descritos de forma mais detalhada e de forma interpretativa.

5.2 CORPUS DA PESQUISA

O *corpus* trabalhado na pesquisa engloba as teses dos programas de pós-graduação brasileiros da Ciência da Informação que foram desenvolvidas quando a área começou a debater a questão dos dados científicos no país. Este fato remete às publicações: “*Compartilhamento de dados e e-Science: explorando um novo conceito para a comunicação científica*”, dos autores Medeiros e Caregnato (2012); e “*Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa*”, dos autores Sayão e Sales (2012). Dessa forma, o recorte temporal da pesquisa compreende as teses da CI, a partir de 2012 a 2020.

Os programas de pós-graduação da CI no Brasil, com curso de doutorado, foram encontrados por meio de uma busca preliminar no *site* da Plataforma Sucupira, ao clicar no ícone “cursos avaliados e reconhecidos” e depois na opção “área de avaliação” – o que levou ao campo cursos avaliados e oferecidos pela Plataforma. Na sequência, foi selecionada a opção de “Comunicação e Informação” e, finalmente, o curso “Ciência da Informação”.

A partir da busca, foram identificados 27 programas de pós-graduação, porém, para fins desta pesquisa, eliminaram-se os programas que oferecem apenas os cursos de Mestrado Acadêmico e Mestrado Profissional. O recorte teve o foco nos cursos de Doutorado. Desta forma, foram identificados 13 programas de pós-graduação na área de Ciência da Informação das seguintes instituições: Universidade de Brasília (UnB), Universidade de São Paulo (USP), Universidade Estadual de Londrina (UEL), Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP-MAR), Universidade Federal da Bahia (UFBA), Universidade Federal da Paraíba (UFPB-JP), Universidade Federal de Minas Gerais (UFMG), Universidade Federal de Pernambuco (UFPE), Universidade Federal de Santa Catarina (UFSC), Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)/Universidade Federal do Rio de Janeiro (UFRJ), Universidade Federal Fluminense (UFF) e da Universidade Fumec (FUMEC). A lista dos programas selecionados pode ser visualizada no Quadro 6 a seguir:

Quadro 6 - Programas de Pós-Graduação brasileiros com curso de doutorado na área da Ciência da Informação

Programas de Pós-Graduação na área da CI	Doutorado	Doutorado Profissional	Mestrado/Doutorado
PPGCINF-UNB	-	-	1
PPGCI-USP	-	-	1
PPGCI-UEL	-	-	1
PPGCI – UNESP	-	-	1
PPGCI-UFBA	-	-	1
PPGCI –UFPB	-	-	1

PPGCI –UFMG	-	-	1
PPGGOC – UFMG	-	-	1
PPGCI-UFPE	-	-	1
PPGCIN-UFSC	-	-	1
PPGCI/IBICT-UFRJ	-	-	1
PPGCI-UFF	-	-	1
PPGSIGC- FUMEC	-	-	1

Fonte: Dados da Pesquisa (2021)

Através da pesquisa, ficou constatado que os programas que oferecem o curso de doutorado são classificados como Mestrado/Doutorado, logo, não foram encontrados programas apenas de Doutorado e/ou de Doutorado Profissional. Foi identificado que 11 programas são nomeados de Programas de Pós-Graduação em Ciência da Informação - PPGCI e 2 programas possuem nomenclaturas diferentes, sendo um intitulado de Programa de Pós-Graduação Gestão e Organização do Conhecimento – PPGOC, da UFMG, e o outro Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento – PPGSIGC, da FUMEC, e ainda foi visualizado que a instituição da UFMG possui 2 (dois) Programas.

A opção pelo *corpus* da pesquisa (as teses dos programas de pós-graduação em Ciência da Informação) se deu por compreender que a CI estuda, na contemporaneidade, determinados assuntos conforme aponta Araújo (2018), como Produção e Comunicação Científica com ênfase no acesso livre do conhecimento e a curadoria digital; discute a Representação e Organização da Informação, no tocante, por exemplo, à exploração de assuntos como ontologias, *web* semântica e recuperação da informação; inclui ainda os estudos sobre os Sujeitos da Informação, dentre outros temas, já que a área explora essas temáticas que estão diretamente relacionadas com os aspectos que envolvem os elementos de ciclos de vida dos dados. É importante identificar se esses pontos estão de fato presentes na forma como os pesquisadores lidam com os seus dados científicos, pois com base no que a área estuda, ela tem potencial de contribuir com os pesquisadores de outros domínios do conhecimento ao ser uma fonte de informação de como gerir com boas práticas os dados provenientes das pesquisas científicas.

A consulta à Plataforma Sucupira permitiu a obtenção de informações relevantes que facilitaram o levantamento das teses, tais como: o ano de início dos cursos de doutorado e os *sites* dos programas. A maioria das teses estava listada nas páginas dos programas na *internet*, com exceção das teses da FUMEC, conforme será apresentado posteriormente. Em diversos casos, as teses estavam listadas com um *link* de acesso ao repositório da instituição, o que facilitou o levantamento de grande parte delas, não sendo necessário o deslocamento físico para obtê-las. Outras informações foram adquiridas como o período de estudo, a quantidades de teses

listadas e o número de teses disponíveis para *download*. A realidade obtida está ilustrada no Quadro 7.

Quadro 7 - Teses brasileiras da Ciência da Informação

Programa	Ano de início dos cursos de Doutorado no Brasil	Período de estudo -	Quantidade de Teses Listadas	Quantidade de Teses Restritas ou não disponíveis	Quantidade de Teses Disponíveis em acesso aberto
PPGCINF-UNB	1992	2012-2020	109	-	109
PPGCI-USP	2006	2013-2019	40	-	40
PPGCI-UEL	2018	-	-	-	-
PPGCI – UNESP	2005	2012-2020	119	-	119
PPGCI-UFBA	2011	2013-2019	27	4	23
PPGCI – UFPB	2012	2017-2019	17	1	16
PPGCI – UFMG	1997	2012-2020	99	4	95
PPGGOC – UFMG	2016	2017-2020	36	7	29
PPGCI-UFPE	2017	-	-	-	-
PPGCIN-UFSC	2013	2017-2020	32	-	32
PPGCI/IBICT-UFRJ	1994	2012-2020	90	2	88
PPGCI-UFF	2014/2015	2018-2019	12	-	12
PPGSIGC- FUMEC	2016	2018-2020	2	-	2
TOTAL DAS TESES			583	18	565

Fonte: Dados da pesquisa (2021)

Os dados do Quadro 7, no tocante ao item “Ano de início dos cursos de Doutorado no Brasil”, foram levantados do *site* da Plataforma Sucupira e ratificados com o conteúdo sobre a história dos programas nos seus respectivos *sites*. Nas páginas da *internet* do PPGCI - UEL e do PPGGOC-UFMG, não foi possível identificar a informação sobre quando o curso de doutorado foi iniciado. Desse modo, as datas encontradas na Plataforma Sucupira foram mantidas. Houve divergência entre a Plataforma Sucupira e o *site* do PPGCI-UFF, uma vez que a primeira mostra que o curso de doutorado iniciou em 2014 e o *site* do programa apresenta o ano de 2015, porém, não é uma diferença discrepante ao comparar com a informação levantada do PPGCI/IBICT-UFRJ, pois a Plataforma Sucupira apresenta a data de início do curso de doutorado em 2009 e o *site* do respectivo programa mostra que foi em 1994. Optou-se por esta última informação, pois algumas teses com data anterior a 2009 foram listadas na página da *internet* do referido programa. Essas realidades divergentes mostram que é preciso haver um alinhamento das informações para que o usuário possa recuperá-las com credibilidade.

O PPGCI-UEL é recente, assim como o PPGCI-UFPE, logo, no momento da coleta, não foi possível identificar teses nesses Programas. Algumas teses foram encontradas em acesso restrito e outras não foram descobertas, não sendo possível o acesso. O fato das teses não

estarem disponíveis nos *sites* dos programas de pós-graduação, nos repositórios institucionais e/ou nas bibliotecas digitais limita a visibilidade da pesquisa e, conseqüentemente, prejudica a notoriedade dos seus dados. Apesar de alguns textos das teses de doutorado em CI, defendidas no período de 2012-2020, não estarem disponíveis, uma maioria correspondente a 97% foi baixada para ser utilizada nesta investigação.

A população do estudo foi levantada no período de dezembro de 2020 a janeiro 2021, correspondendo a 565 teses de doutorado brasileiras em Ciência da Informação, defendidas em quaisquer anos durante o período de 2012 a 2020, e que estão disponíveis em acesso aberto.

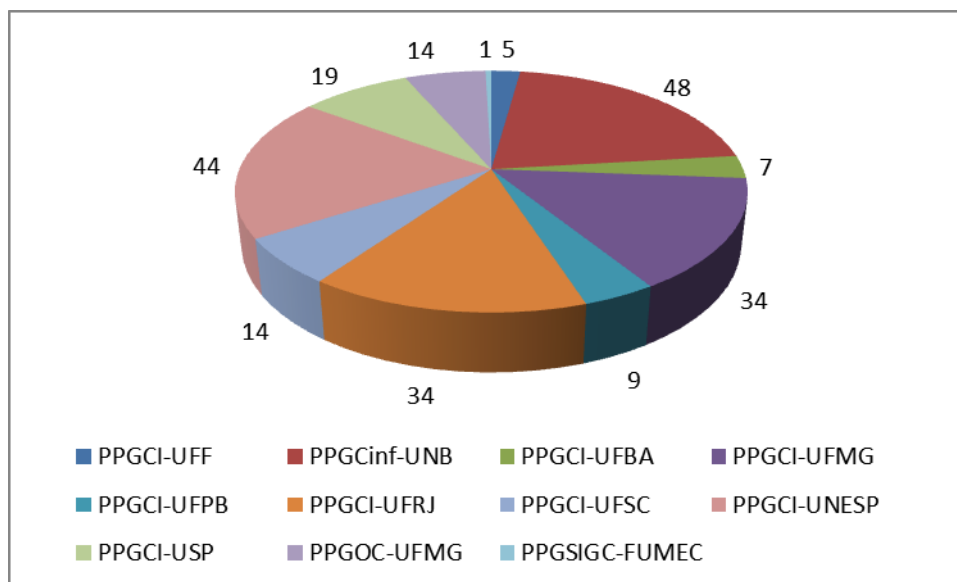
Foi realizada uma amostragem aleatória simples sem repetição. Para obter o tamanho amostral (n), utilizou-se o cálculo amostral para populações finitas (LOHR, 2009), considerando uma proporção de $p = 0,5$, erro amostral de $e = 0,05$ sob o nível de confiança de 95%. A seguir, listamos a fórmula utilizada para o cálculo do tamanho da amostra:

$$Tamanho\ da\ amostra = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \left(\frac{z^2 \times p(1-p)}{e^2 N}\right)}$$

O N na fórmula corresponde ao tamanho populacional e z é o valor tabelado da distribuição normal padrão que corresponde ao nível de 95% de confiança. Assim, obteve-se o tamanho amostral de 229 teses.

As teses disponíveis entre 2012 e 2020 foram baixadas e enumeradas de 1 (um) a 565 para a obtenção da amostragem aleatória simples, que se deu por meio de um sorteio realizado no *software* Excel da empresa *Microsoft*. Em seguida, foi utilizada a função =ALEATÓRIOENTRE(\$A\$1;\$A\$565)+ALEATÓRIO() e, posteriormente, o sorteio foi ordenado através da função =ORDEM.EQ(B1;\$B\$1:\$B\$565). Portanto, foram selecionadas as primeiras 229 teses sorteadas para a composição da amostra analisada. No Gráfico 1, está ilustrado o quantitativo das teses dos programas que foram contempladas para o estudo em questão.

Gráfico 1 – Quantitativo da amostra das teses dos programas de Pós-Graduação brasileiros da Ciência da Informação, do período de 2012 a 2020



Fonte: Dados da pesquisa (2021)

Apesar de a amostra ter sido tomada de forma aleatória, percebe-se que esta contemplou todos os programas brasileiros de pós-graduação na área da Ciência da Informação com teses disponíveis em acesso aberto, no interstício de tempo estudado.

5.3 TÉCNICAS E INSTRUMENTOS PARA A COLETA DE DADOS

A princípio, foi realizada uma visita aos *websites* dos programas que possuem o curso de doutorado, em seguida houve a identificação das teses listadas na maioria dos respectivos *sites*, com exceção da FUMEC, cujas teses foram encontradas apenas no repositório institucional. Posteriormente, foram levantadas as teses que estavam disponíveis nos sites dos programas de pós-graduação em CI, nos repositórios institucionais e/ou biblioteca digital, realizando-se assim o *download* de cada uma delas. Neste contexto, o primeiro objetivo específico foi atingido, o de “Levantar as teses disponíveis dos Programas de Pós-Graduação em Ciência da Informação”.

Ao acessar a página do PPGCINF-UnB, na guia “Pesquisa”, foi selecionada a alternativa “Teses e Dissertações” e houve um direcionamento para o repositório institucional da UnB, representado pela Figura 6. Posteriormente, foi escolhida a opção FCI-Doutorado em Ciência da Informação (Teses) e, para facilitar a recuperação, optou-se pela escolha do filtro “Data de Defesa”, sendo possível obter a lista de 109 teses correspondentes aos anos de 2012-2020.

Através da consulta, foi possível constatar que todas essas teses estavam disponíveis para serem acessadas.

Figura 6 - Repositório Institucional da UnB/Ciência da Informação/Teses

The screenshot shows the UnB Institutional Repository interface. At the top, there's a navigation bar with 'UnB' logo, home icon, and links for 'Comunidades e Coleções', 'Navegar', 'Documentos', and a search bar. Below this, a blue button labeled 'Voltar' and a header 'REPOSITÓRIO INSTITUCIONAL DA UNB' are visible. The main section is titled 'Navegando "FCI - Doutorado em Ciência da Informação (Teses)" por Data de defesa'. It features a search filter for 'Data de defesa' with dropdowns for year and month, and a 'Ir' button. Below this, there's a 'Classificar por' section with dropdowns for 'Data de defesa', 'Em ordem: Ascendente', 'Resultados/Página: 20', and 'Registro(s): Todos', followed by an 'Atualizar' button. A table displays search results, showing columns for 'Data de publicação', 'Data de defesa', 'Título', 'Autor(es)', 'Orientador(es)', and 'Coorientador(es)'. The table lists three theses from 1996 and 1997.

Data de publicação	Data de defesa	Título	Autor(es)	Orientador(es)	Coorientador(es)
12-Abr-2019	1996	A informação encantada : um estudo sobre a busca de informações para o processo de criação publicitária	Suaides, Samir	Boteiro, Tânia Mara	-
15-Abr-2019	1997	A pesquisa em sementes no Brasil : produção científica e tendências de desenvolvimento	Oliveira, Maria José de	Miranda, Antonio	Wetzel, Clovis Terra
16-Abr-2019	15-Out-1997	Hipermedia : novo canal para difusão de tecnologia agropecuária	Salvati, Maria	Robredo, Jaime	-

Fonte: Repositório Institucional da UnB (2020)

O levantamento das teses do PPGCI-USP foi possível através da página na *web* do programa, ilustrada pela Figura 7. Mediante a opção “Teses e Dissertações” da guia “Publicações”, ao escolher o filtro “Nível” e digitar a palavra “doutorado”, e no filtro “Ano” pesquisar por cada ano de 2012 a 2020, foram listadas 40 teses. Ao clicar nos títulos das teses, houve um direcionamento para a Biblioteca Digital USP, sendo possível fazer o *download* de todas as teses. Não foram encontradas teses referentes aos anos de 2012 e 2020.

Figura 7 - ECA USP/Ciência da Informação/Teses PPGCI

The screenshot shows the ECA USP website. The header includes the ECA USP logo and the USP logo. Below the header, there's a navigation bar with links for 'Institucional', 'Ensino', 'Pesquisa', 'Cultura e Extensão', 'Departamentos', 'Acesso rápido', 'contato', 'localização', 'entrar', and a search bar. The main section is titled 'Teses PPGCI'. It features a search filter for 'Ano' and 'Nível' with dropdowns for 'Doutorado' and 'Título', and an 'Aplicar' button. Below this, there's a table displaying search results, showing columns for 'Autor', 'Título', and 'Orientador'. The table lists several theses from 2012 to 2020.

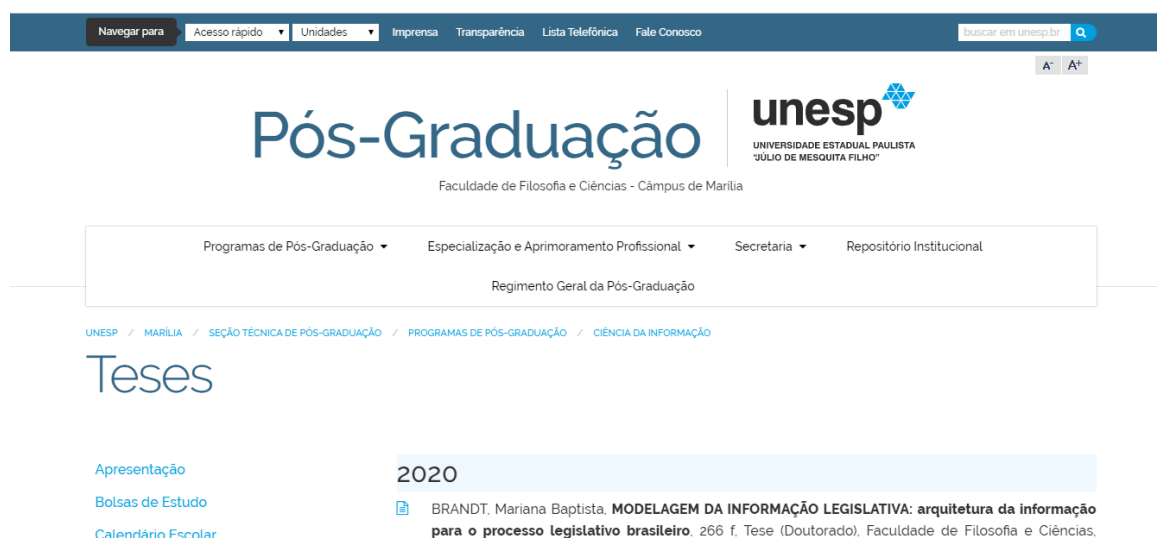
Autor	Título	Orientador
Adaci Aparecida Oliveira Rosa da Silva	Formação e trabalho do profissional da informação na contemporaneidade: subsídios para o ensino de graduação	Asa Fujino
Amanda Leal de Oliveira	A negociação cultural: um novo paradigma para a mediação e a apropriação da cultura escrita	Edmir Perrotti
Amanda Pinto da Fonseca Tojal	Políticas públicas culturais de inclusão de públicos especiais em museus	Maria Helena Pires Martins
Ana Carolina Silva Biscachin	Apropriação social da informação, cultura e tecnologia: software livre e políticas culturais no Brasil	Marco Antonio de Almeida
Ana Claudia Silverio Nascimento	Mapeamento temático das teses defendidas nos programas de pós-graduação em educação física no Brasil (1994-2008)	Dinah Aparecida de Mello Aguiar Poblacion

Fonte: Escola de Comunicação e Artes Universidade de São Paulo (2021)

Por meio de uma visita ao *site* do PPGCI-UEL, buscou-se levantar as teses do referido programa, porém não havia teses disponíveis, tendo em vista que o curso de doutorado teve início no ano de 2018 e o tempo padrão para a realização de um o curso de doutorado é de 4 (quatro) anos.

Através de uma pesquisa na guia “Teses e Dissertações” e ao escolher a opção “Teses” no *site* do PPGCI-UNESP, conforme ilustrado pela Figura 8, foram listadas 119 teses referentes ao período de 2012-2020, porém 4 (quatro) teses, correspondentes ao ano de 2016, estavam apenas listadas, mas não estavam disponíveis, foi possível encontrá-las e recuperá-las no repositório institucional da UNESP. Quanto às demais, foram realizados os *downloads* diretamente do próprio *site*, com a recuperação das 119 teses.

Figura 8 - UNESP/Programa de Pós- Graduação em Ciência da Informação/Teses



Fonte: Programa de Pós-Graduação em Ciência da Informação UNESP (2021)

No *site* do PPGCI-UFBA, conforme ilustrado na Figura 9, ao acessar a guia “Trabalhos de Conclusão” constatou-se a listagem de 27 teses referentes ao período de 2013 a 2019, tendo em vista que o curso de doutorado iniciou em 2011, por isso entende-se que não foi possível encontrar teses no ano de 2012. Dentre elas, estavam disponíveis 23 teses que possuem um *link* de acesso ao Repositório Institucional da UFBA. Assim sendo, foi realizado o *download* das teses disponíveis para serem usadas no estudo em questão. Dos 4 (quatro) trabalhos de conclusão de curso de doutorado que não foram encontrados, 1(um) é referente a 2018 e 3 (três) a 2019. Um fato que chama atenção é a existência de 2 (dois) trabalhos disponibilizados nomeados de teses, 1 (um) do ano de 2013 e o outro cadastrado no ano de 2018, mas ao clicar

nos *links* de acesso, verificou-se que se referem a dissertações, logo foram descritos de forma equivocada, o que dificultou a pesquisa.

Figura 9 - Programa de Pós-Graduação em Ciência da Informação/UFBA/Trabalhos de Conclusão



Fonte: Programa de Pós-Graduação em Ciência da Informação da UFBA (2021)

Ao acessar a página na *internet* do PPGCI-UFPB e clicar na guia “Teses e Dissertações”, houve um encaminhamento para o repositório institucional da UFPB, representado pela Figura 10. Foram usados os filtros Tipo de documento: “Tese”, Área do conhecimento: Ciências Sociais aplicadas e, em seguida, foi escolhida a opção Ciências Sociais Aplicadas: Ciência da Informação. A consulta possibilitou a listagem de 17 teses correspondentes aos anos de 2017 a 2020. Foi possível realizar o *download* de 16 delas por data, que se encontram disponibilizadas em acesso aberto, porém 1 (uma) delas está com o acesso embargado.

Figura 10 - Repositório Institucional da Universidade Federal da Paraíba



Fonte: Repositório Institucional da UFPB (2021)

Através da página na *internet* do PPGCI – UFMG, na opção “Teses defendidas”, conforme ilustrado na Figura 11, e do Repositório Institucional da UFMG, foi possível visualizar 99 teses listadas referentes aos anos de 2012 a 2020, porém 4 (quatro) delas não estavam disponíveis para *download*. Observou-se que no *site* do programa não foram cadastradas as teses de 2019 e 2020 e algumas teses que estavam cadastradas no repositório não estavam listadas no referido *site*, além de não possuir *link* de acesso às teses, o que dificultou a recuperação delas. Um fato a destacar é que, como os trabalhos são disponibilizados separados por ano, em 2012 foi possível identificar um trabalho referente a 2015. No ano de 2013, 4 teses foram listadas de forma repetida. Desta forma, 95 teses foram recuperadas e incluídas no estudo em questão.

Figura 11 - PPGCI/UFMG/Teses defendidas



Fonte: Programa de Pós-Graduação em Ciência da Informação UFMG (2021)

A obtenção das teses do PPGGOC – UFMG ocorreu mediante uma consulta ao *site* do respectivo Programa, a seguir representado pela Figura 12. Foi acessada a guia “Defesa” e a opção “Teses defendidas”, assim como no repositório institucional da UFMG, em que foram listadas 36 teses referentes ao período de 2017 a 2020. É importante notar que, de acordo com o Quadro 3, apresentado anteriormente, o curso de doutorado teve início no ano de 2016 e no ano seguinte já foi possível constatar defesas de teses. Para visualizar as teses, foi recomendado clicar no título (cor roxa) para ser direcionado ao repositório institucional da UFMG, onde as teses estão localizadas, porém algumas delas só foram encontradas no repositório, não sendo possível o acesso por meio do *site*. Não foi factível o *download* de 7 (sete) teses referentes ao ano 2020, 6 (seis) delas não foram encontradas e 1 (uma) se encontra embargada em acesso

restrito. Logo, foi realizado o *download* de um total de 29 teses que estavam armazenadas e disponíveis para serem pesquisadas.

Figura 12 - PPGGOC/UFMG/Teses defendidas



Fonte: Programa de Pós-Graduação em Gestão & Organização do Conhecimento UFMG (2021)

Na busca pelas Teses do PPGCI-UFPE, foi realizada uma consulta ao *site* do respectivo programa. Na guia “Navegação” e na opção “Produção Científica”, houve um direcionamento para o repositório institucional da UFPE, porém na guia “Coleções desta Comunidade” na opção “Teses de Doutorado” não foram encontradas teses, diante da realidade que o curso é recente, o seu início ocorreu em 2017.

Na consulta à página na *internet* do PPGCin-UFSC, conforme ilustra a Figura 13, mais especificamente na guia “Informações do curso” e na opção “Teses e Dissertações”, foram listadas e levantadas, no próprio *site* e no repositório institucional da UFSC, 32 teses correspondentes aos anos de 2017 a 2020.

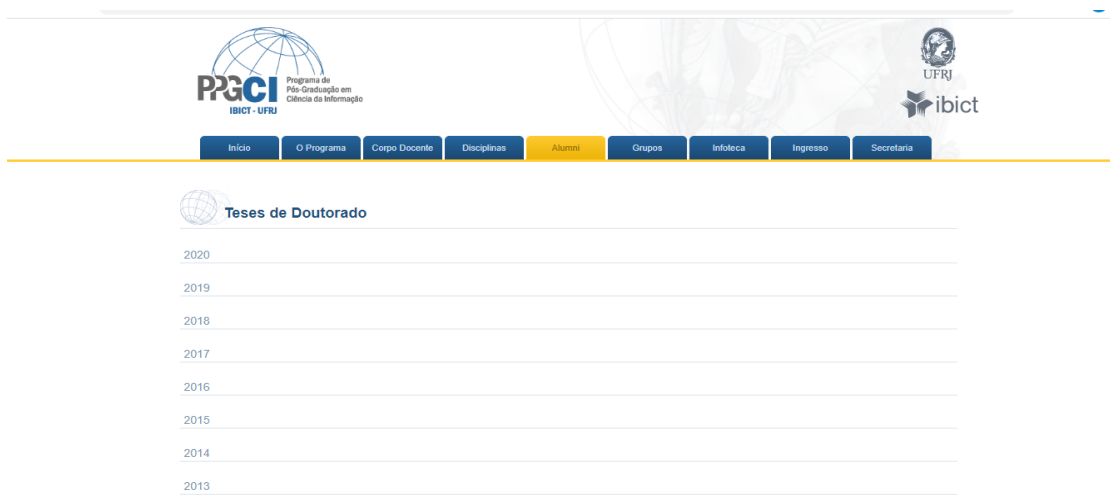
Figura 13 - PPGCin/UFSC/Teses e Dissertações

TIPO	ANO	TÍTULO	AUTOR
Dissertações	2020	Sistema para gerenciamento das informações relacionadas ao treinamento continuado de agentes de segurança pública.	COELHO JÚNIOR, João Daniel
	2019	As dimensões da competência em informação no projeto Lucie para Aprender da organização de estados ibero-americanos para a educação, a ciência e a cultura (ODE) na perspectiva da vulnerabilidade social.	BRITO, Tânia Regina de
		As contribuições da Escola de Chicago	VIEIRA, Keltty Rodrigues

Fonte: Programa de Pós-Graduação em Ciência da Informação da UFSC (2021)

O levantamento das teses do PPGCI/IBICT- UFRJ ocorreu através de uma visita ao *site* do programa. Na guia “Alumni” e na opção “Teses e Dissertações”, foram listadas 90 teses pertencentes ao período de 2012 a 2020, conforme ilustra a Figura 14. Ao clicar nos *links* de acesso, as teses estavam armazenadas no Repositório Institucional do Instituto Brasileiro de Informação em Ciência e Tecnologia (RIDI). Dois (2) arquivos estavam protegidos por direitos autorais, 1 (um) referente a 2018 e o outro a 2014, não sendo possível o acesso. Estavam disponíveis para *download* 88 teses que foram utilizadas no estudo em questão.

Figura 14 - PPGCI/IBICT/UFRJ/Teses de Doutorado



Fonte: Programa de Pós-Graduação em Ciência da Informação IBICT-UFRJ (2021)

No *site* do PPGCI-UFF, conforme representado na Figura 15, na guia “Produção Acadêmica” e na opção “Teses”, foram localizadas 12 teses disponíveis, referentes aos anos de 2018 e 2019. As teses estavam listadas com um *link* de acesso ao repositório institucional da UFF, o que facilitou a recuperação dos documentos.

Figura 15 - PPGCI/ Teses defendidas



Fonte: Programa de Pós- Graduação em Ciência da Informação da UFF (2021)

Em uma busca realizada no *site* do PPGSIGC- FUMEC, não foram encontradas teses pertencentes ao programa, porém na busca ao repositório institucional da Universidade FUMEC, ao escolher a opção comunidade “FUMEC”, posteriormente ao clicar na subcomunidade “Faculdade de Ciências Empresariais (FACE)”, subsequente na opção Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento e, por fim, nas coleções “Teses”, foram listados 2 (dois) trabalhos, correspondentes aos anos de 2018 e 2020, eles estavam disponíveis para *download*, sendo possível acessá-los, o que está representado na Figura 16.

Figura 16 - Universidade/FUMEC



Fonte: Repositório Institucional - FUMEC (2021)

As teses recuperadas foram agrupadas em pastas por programa e separadas em subpastas por ano para facilitar a organização dos documentos. Na sequência, as teses foram enumeradas com a finalidade de selecionar a amostra. Após esse momento, realizou-se a identificação e o isolamento das metodologias dos trabalhos sorteados em uma nova pasta denominada de “Metodologia-Teses selecionadas”. Neste contexto, os documentos foram intitulados de “Amostra_ (nº)”, cujos arquivos foram salvos no formato *Portable Document Format* (PDF), por já terem sido disponibilizados nesse formato e por ele ser importado pelo *software* utilizado na análise de dados.

Para atingir o segundo objetivo específico de “Identificar etapas que podem emergir da intersecção entre os ciclos de vida de dados do DataONE, do DCC e o DDI”, foi realizado um estudo com base na literatura que trata sobre cada ciclo de vida de dados mencionados.

A princípio, os estágios do DataONE foram colocados em evidência e comparados com as ações dos DCC e as fases do DDI. Na sequência, o DDI foi tomado como referência para analisar o há em comum com o DCC. A partir da correlação entre os elementos das três iniciativas, foi possível identificar a interseção entre os componentes dos CVDs, a necessidade de não reunir atividades essenciais em uma única fase e os elementos que, de modo geral, podem ser detalhados de uma forma mais prática.

No que se refere ao terceiro objetivo específico, de “Propor um modelo de referência para descrição dos dados científicos dentro do contexto de metodologias das teses da Ciência da Informação, tendo como parâmetro a análise baseada nesses três modelos”, as etapas sugeridas foram tomadas como referência e foi mostrado o que contém de cada ciclo, DataONE, DCC e DDI. A partir de então, foi criado um modelo de referência, levando em consideração o que precisava ser acrescentado e/ou aperfeiçoado. Portanto, foi proposto um ciclo com características alinhadas com o referencial teórico que trata da metodologia científica no contexto das ciências sociais e com os ciclos de vida de dados correlacionados.

O quarto objetivo específico de “Classificar a tipologia dos dados usados nas teses brasileiras, na área da Ciência da Informação, de acordo com a sua natureza e origem” foi levantado com base no referencial teórico. Desse modo, foram quantificadas as teses através das suas respectivas metodologias quanto a sua natureza, tais como: números, imagens, textos, vídeos, filmes, áudios, *software*, algoritmos, equações, animações, modelos e simulações. A obtenção desta informação é relevante para identificação dos tipos de dados que predominam na CI. A classificação dos dados usados nas respectivas teses incluiu também uma análise quantidade segundo as suas origens, seja observacional, computacional, experimental e registro ou a partir de uma correlação entre estas origens. A obtenção da informação foi fundamental para que os pesquisadores da área possam adotar medidas de preservação dos seus dados.

5.4 MÉTODOS E TÉCNICAS DE ANÁLISE DE DADOS

Na busca por analisar e interpretar as metodologias das teses brasileiras na área da CI, à luz dos elementos do ciclo de vida de dados elaborado no presente estudo, foi adotada a técnica de pesquisa de análise de conteúdo. Esta técnica possibilita analisar as comunicações (MOZZATO; GRZYBOVSKI, 2011) por intermédio de procedimentos sistemáticos e objetivos que descrevem o que contém nas mensagens (BARDIN, 2011).

A análise de conteúdo foi direcionada pelos três polos cronológicos enfatizados por Bardin (2011), os quais são: (1) a pré-análise, (2) a exploração do material e o (3) tratamento dos resultados, a inferência e a interpretação.

Na fase da pré-análise, foi realizada a organização dos documentos com a finalidade de operacionalizar e sistematizar as ideias, ao estruturar um esquema que permita desenvolver o trabalho (RICHARDSON, 2017). Neste momento foram levantadas as metodologias das teses de doutorado brasileiras correspondentes à Ciência da Informação para analisar os elementos do ciclo de vida dos dados de referência.

Na fase de análise do material, Richardson (2017) mostra ser o momento de realizar a codificação, delimitar as categorias de análises dos dados coletados e a quantificação da informação estabelecida. Assim sendo, as categorias corresponderam aos elementos do ciclo de vida dos dados elaborado no estudo em questão, conforme estão elencadas com os seus respectivos códigos, no Quadro 8.

Quadro 8 - Categorias e códigos referentes às teses brasileiras da Ciência da Informação

Categorias	Códigos
Planejar	Plano de gestão de dados, plano de gerenciamento do projeto, campo de estudo, período de coleta, escolha de métodos e técnicas, recursos, questões éticas, equipe de trabalho, planejamento da descrição, planejamento da preservação, natureza dos dados e origem dos dados;
Coletar	Pesquisa bibliográfica, pesquisa de campo e pesquisa documental;
Organizar	Agrupamento e codificação e categorização;
Descrever	Metadados;
Assegurar a qualidade	Garantir a qualidade;
Preservar	Preservação e curadoria;
Encontrar	Descobrir;
Reutilizar	Integrar;
Analisar	Análise de dados e comparação de dados.

Fonte: Elaboração própria (2021)

Para a organização das mensagens em cada categoria, foram verificados os aspectos em comum dos conteúdos das metodologias com as características de cada etapa, o que possibilitou o agrupamento dos trechos das metodologias nos códigos que apresentou similaridade.

Por fim, essa categorização permitiu o tratamento dos resultados, inferência e interpretação, possibilitando trazer à tona a realidade sobre como os pesquisadores fizeram a gestão dos seus dados científicos nos percursos metodológicos de suas teses de doutorado.

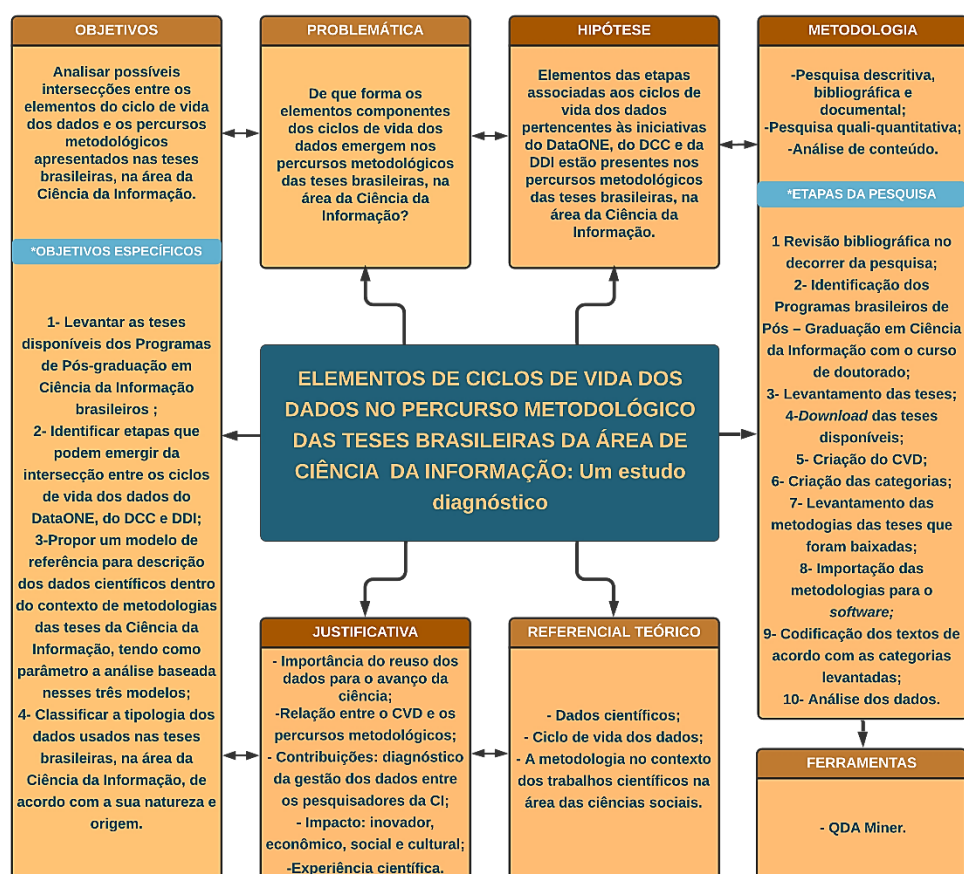
Para a realização deste processo de categorização e codificação, foi utilizado o QDA Miner¹¹, que é um *software* qualitativo para análise de dados. Por meio dele, foram geradas codificações de textos para análise das metodologias.

A quantificação da informação foi gerada no respectivo *software*, que possibilitou a utilização de técnicas estatísticas básicas e a obtenção de dados numéricos, que foram representados em quadros e gráficos.

Na seção da análise e discussão dos resultados, no que se refere à classificação e tipologia dos dados e o momento da intersecção entre o ciclo vida de referência e os percursos metodológicos, foram utilizados segmentos dos textos recuperados com a finalidade de melhor exemplificar os códigos identificados no estudo em questão. Essas citações foram referenciadas no trabalho como Amostra_ e a numeração dos respectivos procedimentos metodológicos. As referências desses trechos podem ser encontradas no Apêndice B deste trabalho.

Para ilustrar o caminho que foi percorrido antes de atingir os resultados da pesquisa. A Figura 17 resume os principais pontos do presente trabalho.

Figura 17 - Resumo dos principais pontos para atingir os resultados da pesquisa



Fonte: Elaboração própria (2021)

¹¹ Disponível em: <https://provalisresearch.com/products/qualitative-data-analysis-software/>

Através do diagrama representado pela Figura 17, é possível apreender o título da tese, a problemática da pesquisa, a hipótese, objetivo geral e objetivos específicos, a justificativa do estudo, o referencial teórico, a metodologia e a ferramenta utilizada no trabalho para o tratamento dos dados.

5.5 CICLO DE VIDA DOS DADOS CIENTÍFICOS NO PRESENTE ESTUDO

O caminho percorrido pelos dados para impulsionar o andamento da presente pesquisa passou pela adoção de algumas ações direcionadas aos dados. Esse processo teve início com a etapa planejar. Através da elaboração do plano de gestão de dados, buscou-se contemplar o processo de curadoria dos dados em todo o seu percurso com as decisões sobre os tipos de dados, técnicas de coleta e análise dos dados, planejamento da descrição e da preservação.

No momento da coleta, foram utilizadas a pesquisa documental e a pesquisa bibliográfica para a obtenção dos dados. Conforme mencionado anteriormente, as teses da CI foram recuperadas nos repositórios institucionais, nos sites dos programas e em uma biblioteca digital.

Na sequência, buscou-se organizar o material obtido em pastas. Após a seleção da amostra, foi criada uma planilha eletrônica no *software* Excel da empresa *Microsoft*, com informações relevantes sobre o ano das teses, o vínculo institucional, os locais dos programas, o idioma do material coletado e as seções ou subseções que as metodologias foram encontradas nas teses. Os dados levantados foram então reunidos de acordo com esses descritores. Na sequência, as metodologias dos trabalhos selecionados foram agrupadas em uma pasta, cujos arquivos em pdf foram importados para o *software* QDA Miner.

Posteriormente, com base na revisão da literatura, ocorreu a criação das categorias com os seus respectivos códigos, a partir do desenvolvimento do CVD de referência, elaborado com base nos ciclos de vida dos dados do DataONE, DCC e DDI. Buscou-se então identificar, nesse momento, se os conteúdos das metodologias possuíam alguma correlação com os elementos do CVD. Desse modo, os segmentos dos textos receberam a codificação de acordo com a similaridade do conteúdo dos trechos com os códigos.

Para que os dados obtidos tenham a possibilidade de serem reutilizados e compartilhados no futuro, utilizou-se o padrão de metadados *Dublin Core* como referência para descrição dos dados do presente trabalho (Quadro 9). Foi levado em consideração o conjunto dos seus quinze elementos apresentados pelo *Dublin Core™ Metadata Initiative* (2021) -

criador, colaborador, editor, título, data, idioma, formato, assunto, descrição, identificador, relação, fonte, tipo, cobertura e direitos.

Quadro 9 - Descrição dos dados

Elementos	Descrição
Criador	Débora Gomes de Araújo Rodrigues
Colaborador	Guilherme Ataíde Dias Universidade Federal da Paraíba
Editor	Figshare
Título	Elementos de ciclos de vida dos dados no percurso metodológico das teses brasileiras da área de ciência da informação: um diagnóstico de gestão e curadoria.
Data	29-04-2021
Idioma	Português (pt) e Inglês (en).
Formato	CSV - “ <i>comma-separated-values</i> ” (valores separados por vírgulas);
Extensão	KB
Assunto	Dados Científicos. Ciclo de Vida dos Dados. Ciência da Informação. Metodologia Científica.
Descrição	O conjunto de dados consiste em uma planilha eletrônica com as informações sobre: a identificação da metodologia da tese selecionada (Amostra_nº); o ano das teses; o programa de vinculação; a cidade do programa; a seção ou subseção das metodologias nas teses; o idioma das metodologias; os segmentos de textos recuperados correspondentes aos códigos das categorias: planejar, coletar, organizar, descrever, assegurar a qualidade, preservar, encontrar, reutilizar e analisar; a codificadora que é a própria autora ou administradora do conjunto de dados; a data da codificação; a quantidade e percentual das palavras dos segmentos encontrados; e a variável documento que caracteriza o material pesquisado.
Fonte	Teses brasileiras da Ciência da Informação referentes ao período de 2012 a 2020.
Tipo de dados	Natureza: Textuais, números e visualizações. Origem: Registro
Direitos	Os dados serão licenciados com uma Licença <i>Creative Commons</i> - Atribuição 4.0 Internacional (CC BY 4.0).

Fonte: Elaboração própria (2021)

As ações voltadas a assegurar a qualidade passou pela releitura dos 229 textos e dos segmentos que foram codificados, com a finalidade de identificar se havia alguma inconsistência no processo de codificação do material. Neste contexto, houve a eliminação de alguns trechos considerados como duplicidade de dados. Também foi visto que alguns materiais apresentaram problemas de interoperabilidade com o *software* QDAMiner, pois partes de alguns textos não foram lidas pelo programa, porém para evitar que uma determinada sentença importante não fosse incluída na análise, o material original foi frequentemente consultado.

A princípio, a preservação dos dados está limitada ao armazenamento dos dados no *laptop* pessoal e no *google drive*. O conjunto de dados está organizado em 2 (duas) planilhas eletrônicas do *software* Excel: a primeira contém os dados das teses, conforme citado anteriormente, e a segunda planilha foi gerada no *software* QDA Miner, com os dados sobre a codificação e categorização do material analisado. Após a aprovação da banca examinadora de Doutorado, os dois conjuntos de dados associados à pesquisa serão unificados e preservados no formato CSV, no repositório de dados Figshare¹², com a descrição dos dados enfatizada na etapa anterior. Após o período de embargo de 6 (seis) meses - tempo necessário para a produção e publicação de artigos resultantes da pesquisa - os dados poderão ser localizados, acessados e, consequentemente, favoráveis a novos estudos.

Para que os dados da pesquisa em questão possam ser encontrados por outros pesquisadores, além de armazená-lo no repositório supracitado, se buscará a publicação dos dados em periódico de dados voltado para os usuários que investigam sobre a temática e para os pesquisadores da Ciência da Informação.

No presente trabalho não houve o reúso de dados de terceiros, pois a pesquisa gerou os seus próprios dados. No que se refere à etapa analisar, a partir dos processos de categorização e codificação por meio da ferramenta de análise de conteúdo, os resultados quali-quantitativos obtidos foram disponibilizados em gráficos e quadros e discutidos na próxima seção.

¹² Disponível em: <https://figshare.com/>



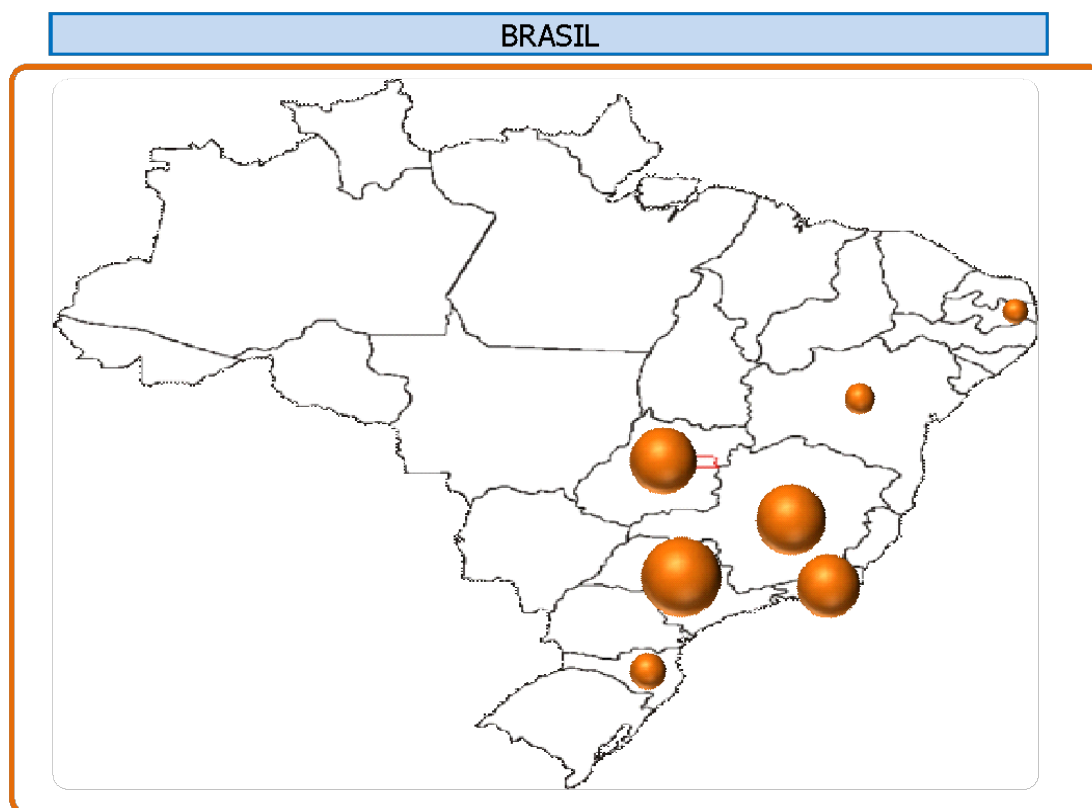
6 ANÁLISE E DISCUSSÃO DOS RESULTADOS

A presente seção trata da análise e discussão dos resultados. Ela foi dividida em 4 (quatro) subseções. Na primeira parte, foi abordada a descoberta das teses brasileiras na área da CI, com ênfase no acesso desses trabalhos e na localização dos procedimentos metodológicos. Na segunda parte, discutiu-se a intersecção entre os modelos de ciclos de vida dos dados do DataONE, DCC e DDI. A partir de então, foi elaborado um modelo de referência que pode ser aplicado pela área. A terceira parte abordou a tipologia dos dados dessas teses, no tocante à natureza e origem dos dados e, finalmente, na quarta parte, foi realizada a intersecção entre o ciclo de vida de dados referência e as metodologias das teses brasileiras da Ciência da Informação.

6.1 A DESCOBERTA DAS TESES DA CIÊNCIA DA INFORMAÇÃO NO BRASIL

Ao ser realizado o levantamento das teses brasileiras da Ciência da Informação referente aos anos de 2012 a 2020, foi possível identificar os estados brasileiros com teses defendidas na área, disponíveis para serem consultadas, o que pode ser ilustrado pela Figura 18:

Figura 18 - Teses disponíveis em acesso aberto do período de 2012 a 2020 por estado da federação



Fonte: Dados da pesquisa (2021)

Do mapa do Brasil representado pela Figura 18, o estado de São Paulo apresentou uma quantidade maior de teses, sendo um total de 159 trabalhos disponibilizados. O estado possui 2 (dois) programas de pós-graduação na área com curso de doutorado, sendo um na cidade de São Paulo e outro em Marília. Na sequência, uma quantidade expressiva foi encontrada no estado de Minas Gerais, em que se identificaram 126 teses disponíveis referente ao tempo estudado, provenientes de 3 (três) programas de pós-graduação de Belo Horizonte, sendo 2 (dois) correspondentes à instituição UFMG e 1 (um) da FUMEC.

O Distrito Federal apresentou uma quantidade de 109 teses no PPGCINF-UnB em Brasília. Este programa brasileiro é o mais antigo na área da CI. O estado do Rio de Janeiro possui 2 (dois) programas de pós-graduação com doutorado na CI, sendo um da UFRJ, na cidade do Rio de Janeiro, e o outro na UFF, em Niterói. Neste contexto, foram encontradas 100 teses para o estudo em questão. Uma realidade a destacar é que o PPGCI/IBICT/ UFRJ, apesar de ser o segundo mais antigo do Brasil, foi o programa que defendeu a primeira tese brasileira, intitulada de “*Um modelo cognitivo de usuário baseado na percepção do valor da informação*”, do autor Luiz Fernando Sayão. Conforme enfatizado anteriormente, este mesmo autor faz parte do marco inicial da temática de gestão de dados no Brasil.

No estado de Santa Catarina foram encontradas 32 teses em Florianópolis do PPGCin-UFSC. Na Bahia, foram obtidas 23 teses da cidade de Salvador, pertencentes ao PPGCI-UFBA e, por fim, no estado da Paraíba, na cidade de João Pessoa, foi possível levantar 16 teses correspondentes ao período estudado.

Apesar do estado brasileiro de Pernambuco possuir programa de pós-graduação com curso de doutorado na área pesquisada, não faz parte da Figura 18, pois no período estudado não foram identificadas teses, haja vista que o curso é novo no referido estado. Esta realidade também foi identificada com o curso de doutorado do programa da UEL, que é o mais recente do Brasil.

A partir do cenário obtido, pressupõe-se que os cursos de doutorado na área da CI, no Brasil, podem se expandir para outros estados, uma vez que existem 14 programas de pós-graduação na área, apenas com os cursos de mestrado (acadêmico ou profissional), distribuídos em outros estados brasileiros. De acordo com a Plataforma Sucupira (2021), esses programas estão vinculados à Universidade Federal de Alagoas (UFAL), à Universidade Federal do Ceará (UFC), à Universidade Federal do Espírito Santo (UFES), à Universidade Federal do Pará (UFPA) e à Universidade Federal do Rio Grande do Norte (UFRN), dentre outras instituições.

Diante da possível expansão dos cursos de doutorado na área da CI pelo país, a temática dos dados científicos tem potencialidade para ser mais divulgada e explorada no contexto

científico, por já se tratar de um assunto discutido entre os pesquisadores e pelos profissionais da Ciência da Informação serem capazes de oferecer suporte para que o acesso e o uso dos dados ocorram da melhor forma (SANTA'ANA, 2016).

6.1.1 O Acesso às Teses Brasileiras da Ciência da Informação

Através do levantamento das teses, foi possível identificar que a maioria delas está listada nos *sites* dos programas com um *link* de acesso ao repositório institucional, com exceção da USP, que disponibiliza as teses em sua biblioteca virtual, e a UNESP e UFSC, que as disponibilizam diretamente nos respectivos *sites*, além de algumas serem encontradas nos repositórios das referidas instituições, conforme ilustra o Quadro 10.

Quadro 10 - Acesso às teses da Ciência da Informação no Brasil

Programas	Local de Armazenamento das teses		
	Repositório Institucional	Biblioteca Virtual	Sites dos programas brasileiros, na área da CI
PPGCINF-UNB	X	-	-
PPGCI-USP	-	X	-
PPGCI-UNESP	X	-	X
PPGCI-UFBA	X	-	-
PPGCI-UFPB	X	-	-
PPGCI-UFMG	X	-	-
PPGGOC-UFMG	X	-	-
PPGCIN-UFSC	X	-	X
PPGCI/IBICT-UFRJ	X	-	-
PPGCI-UFF	X	-	-
PPGSIGC -FUMEC	X	-	-

Fonte: Dados da pesquisa (2021)

O estudo apontou que existe alinhamento entre o que é listado no *site* do PPGCINF-UnB e o seu repositório institucional. O mesmo ocorreu com as teses identificadas no *site* do PPGCI-USP, com a biblioteca virtual da instituição. Apesar de não terem sido encontradas as teses de 2020 da USP, supõe-se que os pesquisadores podem ter um período para depositar os seus trabalhos após a defesa, por isso não foram descobertas.

Na busca pelas teses do PPGCI-UEL e do PPGCI-UFPE, não foi possível encontrá-las. Infere-se que elas não estão disponíveis porque os cursos de doutorado são recentes, o da UEL iniciou em 2018 e da UFPE em 2017. Normalmente, o doutorado demanda um período de 4 (quatro) anos, assim é possível que existam apenas teses em andamento ou que tenham sido

defendidas, mas não foram disponibilizadas pelo período concedido pelos programas após a defesa.

Através do Quadro 10, é possível observar que as teses do PPGCI-UNESP são disponibilizadas no próprio *site*, assim como no repositório da instituição. Mesmo que não exista um alinhamento total em termos de acesso, pois algumas teses estão listadas no *site* do programa e disponíveis apenas no repositório, existe conformidade sobre os títulos das teses encontrados em ambos.

As teses do PPGCI-UFBA apresentaram alguns problemas na recuperação da informação, pois alguns trabalhos que estavam listados não foram encontrados, além de que alguns deles estão categorizados de forma errada, se referindo a teses, quando na verdade são dissertações. Ainda é importante salientar que a forma como as teses estão disponibilizadas prejudica o acesso, pois elas estão listadas juntamente com as dissertações, o que demandou um maior espaço de tempo para realizar o *download* dos trabalhos. A realidade em questão prejudica a visibilidade da pesquisa, consequentemente, a notoriedade dos dados científicos.

As teses do PPGCI-UFPB estão disponibilizadas no repositório institucional, porém, diante da realidade de que o curso de doutorado no referido programa teve início no ano de 2012 e pela quantidade de teses que foi descoberta, infere-se que alguns trabalhos não estão armazenados no repositório – o que dificulta o acesso aos resultados dos estudos científicos que foram realizados e não estão disponíveis.

Na busca pelos trabalhos de conclusão de curso de doutorado do PPGCI-UFMG, ficou constatado que é preciso organizar tanto o *site* do programa como o repositório da instituição, pois as informações estão parcialmente desencontradas entre eles, além disso, o *site* está desatualizado. Em relação ao PPGGOC-UFMG, o *site* também não está totalmente de acordo com o repositório. Algumas teses de 2020 não estão em acesso livre, o que pode ocorrer pelo fato de estarem dentro do prazo após a defesa para posteriormente serem disponibilizadas.

Através do PPGCin-UFSC e o repositório da instituição, foi possível recuperar as teses que estavam listadas, o que facilitou a pesquisa. Apesar do PPGCI/IBICT-UFRJ possuir 2 (dois) trabalhos que não estão acessíveis por estarem protegidos por direitos autorais, existe um alinhamento com o que está listado no *site* e o repositório da instituição. Este cenário também contribuiu com o levantamento das teses.

Os títulos das teses encontrados no *site* do PPGCI-UFF estão de acordo com o repositório institucional. Apesar desse alinhamento, não foram identificadas teses no ano de 2020, pois se compreende que algumas teses, que possam ter sido defendidas neste ano, estão

dentro do prazo para serem disponibilizadas ou por uma falta de atualização tanto do *site* do programa como do repositório da instituição.

Em relação ao PPGSIGC- FUMEC, não existe alinhamento entre o acesso às teses por meio do *site* do programa e do repositório da instituição. É possível que exista uma quantidade maior de trabalho que não foi disponibilizada.

Diante da discussão proposta sobre o acesso às teses da CI no Brasil, foi possível identificar uma falta de alinhamento em alguns casos, de forma mais notória com as teses que estão listadas em *sites* como PPGCI-UFMG, PPGGOC-UFMG e PPGSIGC-FUMEC e as que estão disponibilizados nos seus respectivos repositórios institucionais, o que denota que os *sites* dos referidos programas estão desatualizados. Outra questão preocupante é a organização das teses do PPGCI- UFBA, que prejudica a recuperação dos documentos. Essa realidade pode afetar o acesso aos dados científicos, pois já existe uma dificuldade no acesso às próprias pesquisas.

De modo geral, não houve complicações para encontrar as teses do PPGCI-UNESP, PPGCin-UFSC, PPGCI/IBICT-UFRJ, PPGCI-UFF e, principalmente, no PPGCINF-UnB e PPGCI-USP. Isso pode ser considerado um cenário propício para a disponibilização dos dados científicos, pois já existe uma cultura mais voltada a tornar as pesquisas acessíveis, tornando-as visíveis.

6.1.2 A Descoberta das Metodologias nas Teses Brasileiras da Ciência da Informação

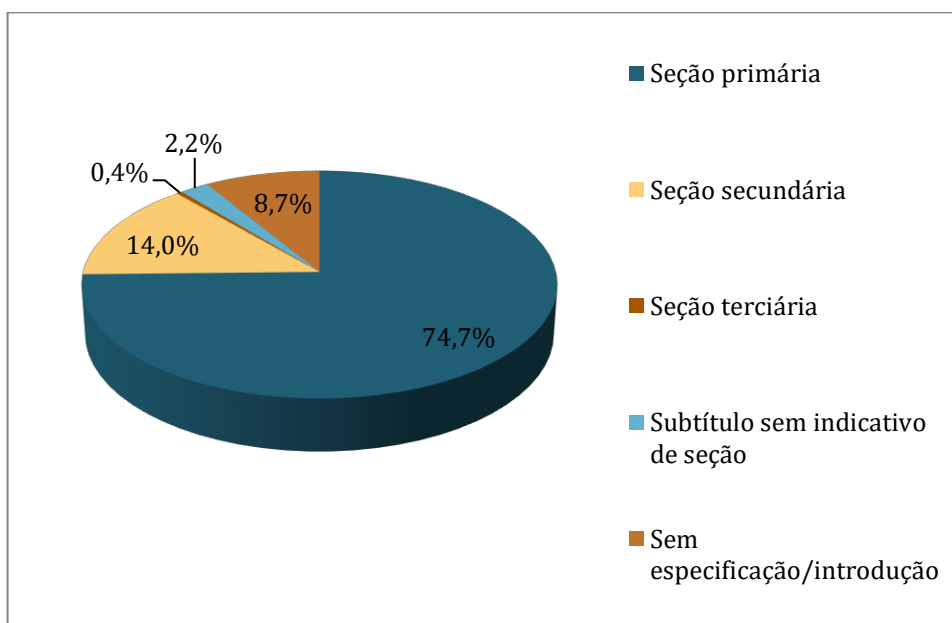
Neste momento do estudo, buscou-se localizar os procedimentos metodológicos nas teses brasileiras da CI obtidas através do cálculo amostral, correspondentes a um total de 229 trabalhos. A partir da pesquisa, ficou constatado que 171 procedimentos metodológicos foram encontrados em uma seção primária das teses (Gráfico 2). Isso significa a existência de um capítulo específico voltado para a discussão da metodologia, o que facilitou a recuperação da matéria procurada.

É importante destacar que a seção primária corresponde à principal divisão de um texto, conforme a NBR 6024 (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2012), logo, infere-se que uma parte representativa das teses posicionou as suas metodologias em um local de destaque, o que pode ser um espaço favorável para a discussão da gestão dos dados científicos.

No entanto, apesar de 3 (três) trabalhos possuírem um capítulo exclusivo para os procedimentos metodológicos, foram intitulados com termos que comumente não são usados

para se referir a esta fase do trabalho científico. A partir de então, foi necessário fazer uma leitura dos capítulos dessas teses para identificar os procedimentos metodológicos. Assim, o fato de não receberem um título com termos que geralmente são usados como percurso metodológico, metodologia, métodos, caminhos da pesquisa, dentre outros, prejudicou a localização da informação.

Gráfico 2 - Localização das metodologias nas teses brasileiras da CI



Fonte: Dados da pesquisa (2021)

O Gráfico 2 mostra que 14% das metodologias, que correspondem a 32 teses, foram encontradas em seções secundárias dos trabalhos, que correspondem, segundo à NBR 6024 (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, p. 1, 2012), a “subdivisão do texto a partir de uma seção primária.” Logo, foram localizadas em uma subseção de um capítulo. Desse total, foi verificado que 27 metodologias pertencem à introdução e 5 (cinco) a outros capítulos das teses. A recuperação da informação procurada foi facilitada por existir um subtítulo com numeração. A partir de então, não foi necessário ler os capítulos na íntegra para identificar as metodologias, uma vez que a visualização estava clara.

Um fato importante a destacar no que concerne à estrutura do trabalho acadêmico apresentado pela NBR 14724, no tocante à divisão dos elementos textuais em introdução, desenvolvimento e conclusão, é que, segundo a norma, na parte introdutória, são apresentados os objetivos dos trabalhos e os motivos de sua elaboração, ao passo que, no desenvolvimento, é o momento de detalhar a pesquisa e deve existir uma parte conclusiva (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2011). Diante dessa realidade, infere-se que as

metodologias não deveriam ser limitadas a serem encontradas apenas como subdivisão da introdução, mas poderiam ser posicionadas em um momento subsequente, ou seja, no desenvolvimento do trabalho com um espaço voltado a especificar, com clareza, a forma como o estudo foi realizado.

As 5 (cinco) metodologias encontradas em capítulos aleatórios foram recuperadas com um certo nível de dificuldade por não se encontrarem em um capítulo próprio, porém os sumários dos trabalhos facilitaram a descoberta e, assim, foram localizadas.

Através da busca, foi possível localizar 1 (uma) metodologia em uma seção terciária (Gráfico 2), pertencente ao capítulo da introdução. De acordo com a NBR 6024 (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, p. 2, 2012), a seção terciária corresponde à “subdivisão do texto a partir de uma seção secundária”, o que posiciona a metodologia em um espaço ainda menos representativo, podendo limitar o debate sobre o tratamento que foi empregado aos dados científicos.

Na parte introdutória, foram encontradas 2,2% das metodologias das teses, que correspondem a 5 (cinco) metodologias com subtítulos, porém sem indicativo numérico, o que prejudicou a busca, por não terem a possibilidade de serem listadas no sumário dos respectivos trabalhos.

Uma realidade crítica foi encontrada em 20 trabalhos, em que 8,7% das metodologias foram localizadas na introdução, sem título, subtítulo e indicativo numérico. Neste contexto, a recuperação da informação foi limitada ao entendimento da pesquisadora sobre a seleção das partes que compreendeu ser referente à metodologia.

Depreende-se das informações levantadas que, apesar da maioria dos trabalhos de conclusão de curso de doutorado possuir um capítulo específico que discute a metodologia e que este pode ser um cenário promissor para debater com mais ênfase o caminho que foi percorrido para a obtenção e tratamento dos dados científicos, uma parte significativa das teses não organizou a informação de forma a permitir uma recuperação da informação com facilidade.

6.2 INTERSECÇÕES ENTRE OS MODELOS DE CICLOS DE VIDA DOS DADOS DO DATAONE, DCC E DDI

Para atingir o objetivo específico de “Identificar etapas que podem emergir da intersecção entre os ciclos de vida dos dados do DataONE, do DCC e do DDI”, procurou-se, nessa subseção do trabalho, analisar em quais aspectos as etapas dos CVDs se assemelham. Os

elementos do Quadro 11 mostram que o modelo DataONE se correlaciona com as iniciativas do DCC e do DDI.

Quadro 11 - Correlação do ciclo de vida de dados do DataONE com o DCC e o DDI

DataONE	DCC	DDI
Planejar	Descrição e Representação da informação; Planejamento da preservação; Curadoria e preservação Acompanhamento e participação da comunidade Conceitualização	Conceito de estudo
Coletar	Conceitualização; Criação e recebimento e Transformação	Coleção de dados
Assegurar a qualidade	Ação de preservação; Acompanhamento e participação da comunidade e Reavaliação	Distribuição de dados
Descrever	Criação e recebimento	Processamento de dados
Preservar	Avaliação e seleção; Arquivamento; Armazenamento; Descarte e Ação de preservação	Arquivamento de dados
Descobrir	Acesso, uso e reúso.	Descoberta de dados
Integrar	Acesso, uso e reúso.	Reaproveitamento
Analisar	Acesso, uso e reúso Reavaliação	Análise de dados

Fonte: Dados da pesquisa (2021)

Em estudo realizado por Araújo *et al.* (2019), foi feita a intersecção dos elementos do CVDs do DataONE com o DCC. Os resultados apontaram que a etapa planejar do DataONE está alinhada com as ações de **descrição e representação da informação**, porque é o momento de fazer o planejamento de como os dados serão descritos, com o objetivo de possibilitar que, posteriormente, eles possam ser encontrados e que os seus conteúdos possam ser entendidos. Isso é algo enfatizado por Strasser *et al.* (2012), ao revelarem que é um momento de esclarecer a produção do registro de metadados, baseado em um padrão de metadados. Envolve também as ações de **planejamento da preservação, curadoria e preservação**, pois estratégias devem ser levantadas para direcionar a preservação dos dados no decorrer do seu ciclo de vida. E ainda inclui a ação de **acompanhamento e participação da comunidade**, porém, conforme destacam Araújo *et al.* (2019) de forma parcial, no DCC se debate o envolvimento do usuário da informação e no DataONE o foco está voltado à definição da equipe e suas responsabilidades no contexto da gestão de dados. Adicionalmente, compreende-se que a etapa de

conceitualização também faz parte dessa reflexão, pois envolve o planejamento da coleta dos dados.

A etapa planejar do DataONE se relaciona com a fase de **conceito de estudo** do DDI, uma vez que nesta fase decisões devem ser tomadas levando em consideração a coleta, processamento e análise dados, constituindo uma fase de planejar o estudo (BALL, 2012).

Segundo Araújo *et al.* (2019), a etapa coletar do DataONE (Quadro 11) está ligada com algumas ações sequenciais dos DCC, a saber: a ação de **conceitualização**, pois de acordo com o que foi planejado, esta ação também está voltada para a realização da coleta de dados. De forma parcial, inclui a ação de **criação e recebimento**, pois neste contexto é possível a aquisição/criação de dados. E ainda inclui a ação de **transformação**, por ser um momento de criar dados com base no conjunto original. A etapa coletar do DataONE se correlaciona com a etapa **coleção de dados** do DDI, porque de acordo com o que foi proposto nas metodologias adotadas em cada pesquisa, a coleta deve ser realizada levando em consideração os métodos e as fontes.

A etapa de assegurar a qualidade do DataONE (Quadro 11) relaciona-se com a **ação de preservação** do DCC, cujas práticas devem proporcionar a qualidade dos dados, incluindo aspectos como a autenticidade, confiabilidade e a usabilidade. É possível existir uma relação com a ação de **acompanhamento e participação da comunidade**, uma vez que em alguns contextos a garantia da qualidade dos dados está atrelada àqueles que possuem conhecimento acerca do processo que envolve a gestão de dados ou das necessidades e usos que podem ser realizados com bases nesses dados. Há ainda uma relação com a etapa **reavaliação**, pois os dados que não foram considerados de qualidades, neste momento serão avaliados novamente (ARAÚJO *et al.*, 2019). No que se refere à relação do DataONE com o DDI, no que tange a etapa assegurar a qualidade, existe uma ligação com a etapa **distribuição de dados**, pois segundo Ionescu (2007) é um momento de realizar checagens para um posterior arquivamento, ou seja, fazer verificação que consequentemente poderá implicar na qualidade dos dados.

Para Araújo *et al.* (2019), a etapa descrever do DataONE (Quadro 11) está alinhada com a etapa **criação e recebimento** do DCC, visto que ao serem criados/recebidos os dados, os metadados são preenchidos de acordo com o que foi planejado na etapa de descrição e representação da informação, o que de modo similar é decidido na etapa planejar do DataONE. A etapa descrever desta iniciativa se relaciona com a etapa de **processamento de dados** do DDI, uma vez que os dados precisam ser registrados de uma forma que seja legível por máquinas e por humanos. Para tanto, necessitam ser descritos com metadados apropriados, levando em

consideração, por exemplo, os instrumentos que foram utilizados, questões inerentes a limpezas de dados, dentre outros (BALL, 2012; IONESCU, 2007).

A etapa de preservar do DataONE (Quadro 11) está relacionada com encaminhar os dados para um arquivo apropriado que possibilite o seu uso a longo prazo (STRASSER *et al.*, 2012). Esta realidade está alinhada com algumas ações do DCC e tem ligação com a ação de **avaliação e seleção**, visto que os dados são avaliados e selecionados para a curadoria de longo prazo. Envolve as ações de **arquivamento** e **armazenamento**, uma vez que os dados são enviados para um centro de dados, sob a orientação de padrões relevantes, possibilitando o armazenamento adequado. Parcialmente também tem relação com ação **descarte** do DDC, pois os dados que não foram escolhidos para serem mantidos por meio da preservação poderão ser eliminados. Neste contexto, são necessárias estratégias e ações de preservação, por isso de forma indireta também existe uma ligação com a etapa de **ação de preservação**, de modo a possibilitar que a integridade dos dados seja mantida (ARAÚJO *et al.*, 2019). É possível identificar com clareza a correlação da etapa preservar do DataONE com a etapa de **arquivamento de dados** do DDI pois, segundo Ball (2012), para que o valor dos dados seja mantido no decorrer do tempo, é necessário que eles sejam transferidos para um arquivo, ou seja, é preciso uma preservação adequada.

A etapa descobrir, integrar e analisar do DataONE tem relação com as ações de **acesso, uso e reúso** do DCC. A partir deste momento, torna-se possível fazer o uso e o reúso dos dados no presente e no futuro por estarem acessíveis. A etapa analisar do DataONE também tem ligação com a ação de **reavaliação** do DCC, diante da realidade de que no momento da análise dos dados pode ser identificada a necessidade de ajustes neles, assim como refletido o processo de curadoria como um todo (ARAÚJO *et al.*, 2019). A etapa descobrir do DataONE está diretamente ligada com a etapa de **descoberta de dados** do DDI, pois é um momento de tornar os dados visíveis com uma documentação adequada, para que possam ser úteis em novos estudos. A etapa integrar do DataONE se correlaciona com a etapa **reaproveitamento** do DDI, uma vez que em ambas ocorre a combinação de dados. Segundo Ball (2012), essas combinações de dados ocorrem envolvendo outros conjuntos similares e, segundo Strasser *et al.* (2012), provém de diferentes fontes para a formação de um conjunto homogêneo de dados. Por fim, a etapa analisar do DataONE está ligada com a etapa de análise de dados do DDI; essas etapas são similares, o foco está voltado a analisar os dados, para que atendam às necessidades da própria pesquisa, mas que também possam ser reutilizados em outros estudos.

O Quadro 12 evidencia a ligação do DDI com as demais iniciativas e por meio dele é possível visualizar que a etapa conceito de estudo do DDI tem relação direta com a ação de

conceitualização do DCC, pois é um momento de planejar como os dados serão obtidos e os métodos que serão utilizados. Embora o planejamento dos metadados não esteja detalhado nas descrições das fases do DDI trazidas por Ball (2012) e Ionescu (2007) e por ainda não se encontrar o detalhamento das fases no *site* do DDI, infere-se que nessa fase esse aspecto seja discutido, pois segundo Kanjala *et al.* (2017) o ciclo de vida do DDI envolve módulos que consistem em pacotes de metadados, cujo foco do modelo é o processo de pesquisa orientado por metadados (THOMAS, 2013). Desta forma, é possível que exista uma relação indireta com a ação de **descrição e representação da informação**, que compreende um momento de definição de metadados e com as ações de **curadoria e preservação**, porque se planeja um pouco a gestão do dado, como já especificado na descrição do modelo.

Quadro 12 - Correlação entre os ciclos de vida dos dados do DDI, DataONE e o DCC

DDI	DataONE	DCC
Conceito de estudo	Planejar	Conceitualização; Descrição e representação da informação; Curadoria e preservação
Coleção de dados	Coletar	Conceitualização; Criação ou recebimento
Processamento de dados	Descrever	Criação ou recebimento
Distribuição de dados	Assegurar a qualidade	Avaliação e seleção; Descarte; Acesso, uso e reúso e Acompanhamento e participação da comunidade
Arquivamento de dados	Preservar	Curadoria e preservação; Arquivamento; Ação de preservação e Armazenamento
Descoberta de dados	Descobrir	Acesso, uso e reúso
Análise de dados	Analisar	Reavaliação
Reaproveitamento	Integrar	Acesso, uso e reúso; Transformação Migração

Fonte: Dados da pesquisa (2021)

A etapa coleção de dados do DDI também tem relação com a etapa **conceitualização** do DCC (Quadro 12), pois embora esta planeje o processo de coleta, nela também se inclui a obtenção dos dados propriamente dita. Envolve também a fase de **criação ou recebimento**, pois é um momento de geração ou obtenção de dados em ambas as fases do DCC e do DDI.

A etapa de processamento de dados do DDI tem relação indireta com a **criação ou recebimento** do DCC, pois de acordo com Ball (2012), à medida que os dados são criados ou obtidos, conforme enfatizado anteriormente, eles vão sendo processados de forma a serem

legíveis por máquinas e humanos. Esta realidade denota que neste contexto são atribuídos metadados.

Segundo Ionescu (2007), a fase de distribuição do DDI é um momento de fazer checagens para só assim os dados serem arquivados, o que mostra a relação existente dessa fase com as ações de **seleção e avaliação** do DCC, que resultará na preservação dos dados que foram selecionados. Neste sentido, compreende-se que exista de forma parcial uma correlação com a ação de **descarte** do DCC. É possível também visualizar uma ligação com as ações de **acesso, uso e reúso** e com **acompanhamento e participação da comunidade**, pois a preocupação está em tornar possível o acesso aos dados pelos usuários (BALL, 2012).

A etapa de arquivamento de dados do DDI está relacionada com as ações de **curadoria e preservação, arquivamento, ação de preservação e armazenamento do DCC** (Quadro 12), uma vez que remete à manutenção dos dados de forma apropriada para que tenham a possibilidade de reúso no futuro.

A etapa de descoberta de dados do DDI tem relação direta com **o acesso, uso e reúso** dos dados do DCC, pois busca possibilitar que os dados estejam acessíveis aos possíveis usuários. Para Ionescu (2007), esta etapa compreende um contexto de proporcionar visibilidade aos dados.

A análise de dados do DDI está ligada à fase de **reavaliação** do DCC, pois para uma nova formação de uma estrutura conceitual com os dados, é preciso que seja verificado se existem necessidades de melhorias durante todo o ciclo, conforme aponta o DCC (2020). E, por fim, a fase de Reaproveitamento do DDI está correlacionada com as ações de **acesso, uso e reúso**, de modo que os dados possam estar acessíveis para serem reutilizados e com as ações de **transformação e migração** do DCC, visto que possibilita a criação de novos dados com base no conceito original (DDC, 2020). Neste contexto, também pode existir a combinação de dados com outros conjuntos similares (BALL, 2012), o que é importante para fazer comparações entre estudos e atestar a veracidade das pesquisas.

Diante do exposto, verificou-se que algumas ações do DCC estão presentes em mais de uma etapa do DataONE e do DDI, de modo simultâneo às ações de conceitualização, criação e recebimento e acesso, uso e reúso. Algumas das ações do DCC se repetem apenas no DataONE, tais como: acompanhamento e participação da comunidade, ação de preservação e reavaliação.

Com base no que foi discutido, identificou-se a existência de intersecção entre as fases dos ciclos de vida do DataONE, DCC e DDI em alguns momentos de forma mais acentuada e em outros menos evidente. De modo geral, os ciclos de vida do DataONE, DCC e DDI precisam ser mais detalhados, a fim de que haja uma compreensão mais prática sobre a aplicabilidade

deles. O DDI, por sua vez, embora tenha foco nos metadados, os autores Ball (2012) e Ionescu (2007) não o tratam de forma clara sobre como lidar com eles em cada fase do ciclo.

6.2.1 Modelo de Referência da Intersecção do DCC, DataONE e DDI

A partir do presente estudo, foi possível atingir o segundo objetivo específico de “identificar etapas que podem emergir da intersecção entre os ciclos de vida DCC, DataONE e DDI”. O Quadro 13 ilustra o que as fases propostas, que correspondem a ações, trazem de cada ciclo:

Quadro 13 - Etapas sugeridas da intersecção do DCC, DataONE e DDI

MODELO DE REFERÊNCIA	DCC	DATAONE	DDI
Planejar	Descrição e representação da informação; Planejamento da preservação; Acompanhamento e participação da comunidade; Curadoria e preservação; Conceitualização	Planejar	Conceito de estudo
Coletar	Conceitualização e Criação ou recebimento;	Coletar	Coleção de dados
Organizar	Criação ou recebimento	Coletar	Coleção de dados Processamento de dados
Descrever	Criação ou recebimento	Descrever	Processamento de dados
Assegurar a qualidade	Avaliação e seleção Acompanhamento e participação da comunidade Curadoria e Preservação; Ação de preservação Descarte	Assegurar	Distribuição de dados
Preservar	Arquivamento; Ação de preservação e Armazenamento	Preservar	Arquivamento de dados
Encontrar	Acesso, uso e reúso	Descobrir	Descoberta de dados
Reutilizar	Acesso, uso e reúso; Transformação e Migração	Integrar	Reaproveitamento
Analisar	Reavaliar	Analisar	Análise de dados

Fonte: Dados da pesquisa (2021)

No que se refere à etapa sugerida de “planejar”, ela está relacionada com o DCC nos critérios de decisões sobre os metadados que serão utilizados, sobre as estratégias de preservação, sobre a equipe de trabalho e quem serão os usuários que poderão fazer o uso ou reúso dos dados. Neste contexto, também serão decididas as questões de direitos autorais e

permissões de acesso. A relação com o DataONE ocorre por ser um momento de planejamento do processo de coleta, de armazenamento, do período de embargo e como os dados serão descritos. A correspondência com o DDI é inerente também às decisões sobre a coleta e sobre o processamento e análise de dados.

A etapa proposta “coletar” está relacionada com o DCC com a execução da coleta. Em relação ao DataONE, está alinhada com a forma de obtenção de dados e ligada aos métodos e fontes de dados propostos na etapa coleção de dados do DDI.

É importante enfatizar que na etapa coletar, de modo geral, os três ciclos incluem ações de forma direta ou indireta referentes à organização dos dados, ou seja, a forma de agrupar os dados. Para fins desse novo ciclo de referência, optou-se por criar uma fase, intitulada de “organizar”, para não resumir atividades essenciais, de forma a não trazer dificuldade na hora da aplicação.

A etapa sugerida “descrever” se relaciona com o DCC no tocante à utilização de metadados identificados na ação de criação e recebimento, com o DataONE também segue o mesmo caminho, pois se refere à adoção de um padrão de metadados com a finalidade de descrever os dados. Está atrelada ao DDI, uma vez que essa descrição deve possibilitar que os dados sejam legíveis por humanos e por máquinas.

A sugestão da etapa “assegurar a qualidade” está relacionada com o DCC no quesito de acompanhamento e participação da comunidade, no sentido de que as pessoas ligadas ao projeto devem assegurar que os dados estão sendo tratados de forma adequada, que são autênticos, confiáveis e usuais. Também está ligada às ações de avaliação e seleção, pois é um momento de empreender ações para que os dados sejam preservados, assim alguns dados poderão ser descartados, não passando para a próxima etapa de preservar. Relaciona-se com o DataONE no tocante a assegurar que houve uma coleta de dados adequada e com o DDI, pois é um momento de fazer checagens para verificar a confiabilidade dos dados. Uma questão que merece atenção é que a etapa de assegurar a qualidade, que de forma explícita no DataONE vem antes da etapa descrever, foi realocada posteriormente, pois compreende-se ser o momento de assegurar que o dado foi coletado e a correspondência descritiva dele está adequada.

A etapa proposta “preservar” se relaciona com o DCC, porque é um momento de ter como parâmetro documentos e requisitos legais para depositar os dados em um arquivo de longo prazo, conforme o DataONE orienta, de forma que os dados sejam armazenados com segurança, de acordo com orientação do DDI.

A etapa sugerida “encontrar” está atrelada ao DCC, no sentido de os dados estarem acessíveis aos usuários potenciais; ao DataONE, no quesito desses serem descobertos com os seus respectivos metadados; e ao DDI, quanto à visibilidade dos dados por meio de publicações.

A recomendação da etapa “reutilizar” está relacionada com os três ciclos, com o fato de integrarem a pesquisa dados de outros estudos para incrementar o conjunto de dados já existentes ou para fazer comparações que poderão aperfeiçoar o estudo. Isso será discutido na próxima etapa proposta “analisar”, que é o momento conectado ao DCC de reavaliar e analisar do DataONE no que se refere à verificação dos ajustes que poderão contribuir com os dados, assim como a aplicação de métodos de pesquisa para realizar a análise dos dados.

A partir das ideias identificadas nos ciclos de vida do DataONE, DCC e DDI, foi possível atingir o terceiro objetivo específico de “propor um modelo de referência com etapas mais refinadas e adaptadas para descrição dos dados científicos dentro do contexto de metodologias das teses da Ciência da Informação, tendo como parâmetro a análise baseada nesses três modelos”. Embora este novo ciclo tenha o foco na área da CI, infere-se que pode ser voltado para a grande área das Ciências Sociais Aplicadas, pois as características do modelo têm suas bases na literatura discutida sobre as metodologias das Ciências Sociais, além disso, foi utilizado o CVD do DDI, que é voltado para esse domínio do conhecimento científico. O CVD de referência está ilustrado na Figura 19. O ciclo contém 9 (nove) etapas, cujas práticas precisam ser tomadas em cada uma delas.

Figura 19 - Ciclo de vida dos dados científicos



Fonte: Adaptado de DataONE (2020), DCC (2020) e DDI (2020)

Para uma melhor compreensão do significado das etapas levantadas, a seguir seguem as suas características, expressas no Quadro 14:

Quadro 14 - Detalhamento das etapas do ciclo de vida dos dados científicos

ETAPAS	CARACTERÍSTICAS
Planejar	Etapa de decidir sobre vários aspectos inerentes a cada fase do ciclo de vida dos dados. Trata-se de um momento de elaborar um plano de gestão de dados e um cronograma para que os dados possam ser gerenciados de forma correta. Neste contexto, são necessárias algumas práticas, tais como: ✓ Decidir sobre quais os dados serão obtidos e o que precisa ser realizado para alcançá-los. Neste momento, ocorre a identificação do campo de estudo, a saber: sujeitos que trarão respostas à problemática da pesquisa, local onde se realizará o estudo e a definição da amostragem, assim como a escolha do instrumento de pesquisa. ✓ Definir o período da coleta alinhado com o cronograma, diante do prazo para finalizar o trabalho; ✓ Planejar a descrição dos dados, no que tange aos tipos de metadados e o padrão que será adotado. ✓ Definir estratégias para preservação no que se refere à escolha de um arquivo de dados, de modo que seja garantida a persistência dos dados no longo prazo e deixar clara a frequência de <i>backup</i> dos dados e de que forma ocorrerá. ✓ Definir os colaboradores e suas responsabilidades dentro da equipe de trabalho e quem possivelmente serão os usuários dos dados; ✓ Deixar claro para a equipe de trabalho o que é essencial e não pode ficar de fora, sob pena de comprometer a qualidade dos dados; ✓ Com base na escolha das técnicas, decidir sobre como os dados serão analisados. ✓ Definir metas, ou seja, o que se espera após a execução de cada etapa e ✓ Levantar os recursos financeiros e tecnológicos disponíveis para o funcionamento das etapas do ciclo.
Coletar	É a etapa de ação para geração dos dados com base no planejamento realizado previamente. Assim sendo, por meio de instrumentos ocorre a aquisição, criação ou transformação dos dados que necessitam ser registrados concomitantemente para não serem perdidos.
Organizar	Esta etapa é responsável pelo agrupamento, codificação ou categorização dos dados
Descrever	Etapa de preencher os metadados dos dados que foram obtidos, de acordo com o padrão de metadados que foi anteriormente planejado para ser adotado, para que os dados no futuro possam ser entendidos ao terem a possibilidade de serem legíveis por máquinas e humanos.

Assegurar a qualidade	Compreende a etapa de assegurar que os dados sejam autênticos, confiáveis e usuais, a partir de uma relação com a equipe de trabalho que possui as informações sobre a infraestrutura, os instrumentos utilizados e o contexto de geração de dados. Além de ser um momento de avaliar e selecionar os dados que serão preservados, decidindo os que serão descartados ou serão mantidos, constituindo uma importante fase para a preservação dos dados.
Preservar	Momento de preservar dados em um arquivo, repositório ou bases de dados, com vista à longevidade dos dados. Esta fase deve levar em consideração a questão legal dos dados no tocante a aspectos de direitos autorais e o período de embargo para ficar claro o momento que os dados podem estar acessíveis para os potenciais usuários.
Encontrar	Esta fase possibilita a descoberta de dados acessíveis para serem usados e reutilizados agora e no futuro, o que contribuirá com a visibilidade dos dados, os quais podem ser encontrados em periódicos de dados, publicações ampliadas e repositórios de dados.
Reutilizar	Etapa de agregar um conjunto de dados à pesquisa. Neste momento, é necessário levar em consideração a documentação dos dados, questões de interoperabilidade dos sistemas e é um momento que requer o <i>feedback</i> dos usuários dos dados para a identificação de necessidades para fazer o reuso dos dados da melhor forma.
Analisar	Momento de descrever os métodos e as técnicas utilizadas na análise dos dados, etapa de averiguar se os dados necessitam ser ajustados, também compreende um momento de confrontar dados da pesquisa com dados de outras pesquisas já realizadas para analisar a veracidade dos dados e assim analisar a necessidade de repensar o processo de curadoria como um todo.

Fonte: Dados da pesquisa (2021)

É importante salientar que a etapa “planejar” envolve todo o ciclo, pois nela é decidido o que fazer em cada momento do ciclo, de modo que todas as demais etapas orbitam em torno dela e dos dados científicos (Figura 19).

O cumprimento de cada etapa deve ser adaptado ao projeto de cada pesquisador, se os dados forem provenientes da própria pesquisa, é importante levar em consideração cada etapa do ciclo e a interação entre elas. Nos casos em que o pesquisador estiver fazendo o reuso de dados, algumas etapas podem ser ocasionais. De todo modo, é fundamental dar importância à descrição dos dados, a fim de assegurar a sua qualidade, bem como é preciso que sejam preservados para que, posteriormente, possam ser descobertos e reutilizados em novos estudos.

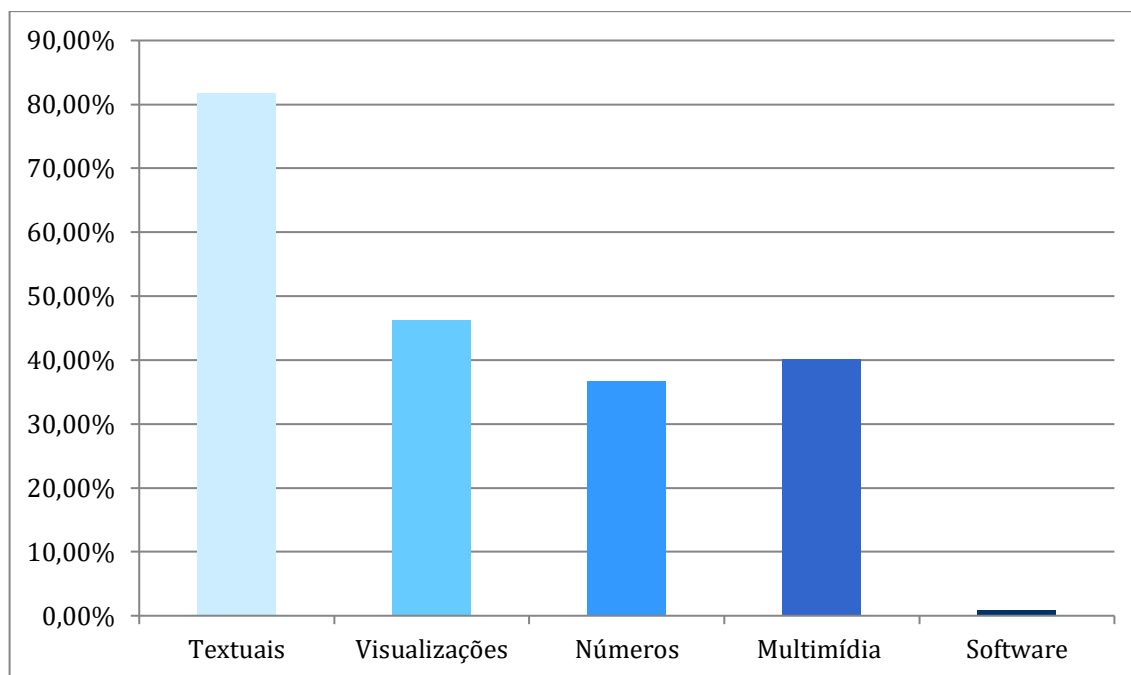
6.3 A TIPOLOGIA DOS DADOS DAS TESES BRASILEIRAS DA CIÊNCIA DA INFORMAÇÃO, DE ACORDO COM A SUA NATUREZA E ORIGEM

De acordo com o relatório NSB (2005), os dados podem se diferenciar por sua natureza. Eles podem ser textos, números, imagens, vídeos, áudios, *software*, algoritmos, equações, animações, modelos, simulações. Esta informação é corroborada pela OECD (2007). O relatório do NSB (2005) mostra que os dados também podem se diferenciar por sua origem: observacional, computacional e experimental e Borgman (2015) acrescenta a quarta categoria de origem dos dados, a de registros.

Para Sayão e Sales (2020), reconhecer as diferenças no tocante à tipologia dos dados é fundamental para a adoção de ações voltadas à gestão dos dados provenientes das investigações científicas e do ciclo de vida da curadoria, haja vista que cada tipo de dado requer processos de gestão diferenciados e de profundidade em relação a medidas de curadoria, a exemplo do arquivamento de longo prazo. Partindo dessa reflexão, a seguir serão analisados os tipos de dados que foram identificados nas metodologias das teses brasileiras da Ciência da Informação, em relação à natureza e à origem dos dados científicos, com a finalidade de alcançar o quarto objetivo específico desta pesquisa.

6.3.1 Natureza dos Dados

No contexto da curadoria, é fundamental compreender com clareza a natureza dos dados, para desenvolver modelos e estratégias, estimar o aparato tecnológico e organizacional necessário à sua gestão, além de mensurar o investimento que será demandado para a preservação dos dados (SAYÃO; SALES, 2020). Em face desse cenário, através do Gráfico 3 é possível visualizar os dados que foram levantados nos textos analisados.

Gráfico 3 - Natureza dos dados da Ciência da Informação no Brasil

Fonte: Dados da pesquisa (2021)

Os dados textuais foram os mais comuns, sendo identificados em 81,7% dos casos, seguidos dos dados de visualizações com uma participação de 46,30% das metodologias. Na sequência, os dados multimídia com 40,2%; em uma quantidade um pouco menor, os dados referentes a números com um percentual de 36,7% e, com uma presença tímida, foram identificados dados como *software*, em 0,90% dos casos. Os resultados apontam que em um mesmo trabalho foram utilizados dados de naturezas diferentes. Assim como é possível verificar que nos dados da CI, mesmo que apresentem uma predominância dos dados textuais, foram encontrados diferentes tipos de dados em suas respectivas metodologias.

A taxonomia adotada em um estudo desenvolvido por Sales e Sayão (2019), a partir de um levantamento da literatura e um estudo empírico envolvendo a área de Ciências Nucleares, no tocante à natureza dos dados, foi tomada como base para reunir em um mesmo grupo diversos tipos de textos, várias formas de visualizações dos dados, distintas representações numéricas, diversos recursos multimídia e diferentes tipos de *software*. Nesse caso, houve uma adaptação para a Ciência da Informação.

Neste sentido, os dados textuais encontrados remetem aos metadados que foram importados de bases de dados com a descrição de documentos. Questionários e formulários que foram preenchidos pelos sujeitos investigados também foram considerados dados textuais, além de transcrição de entrevistas, várias anotações que foram utilizadas como fonte de dados,

padrões, diversos documentos como leis, memorandos, cartas, atas relatórios, entre outros. Também foram incluídos como dados em natureza de textos os que foram provenientes de produções bibliográficas científicas como livros, artigos, material apresentado em evento, etc. Os diários de campo fizeram parte do mesmo contexto, pois neles foram registradas as percepções dos pesquisadores no trabalho de campo e, por fim, foram identificados cadernos de laboratórios. Em face desse cenário, exemplos de segmentos dos textos foram extraídos para representar os dados textuais, conforme Quadro 15.

Quadro 15 - Dados textuais da Ciência da Informação no Brasil

Tipos de dados		Exemplos de segmentos de textos
Textuais	Metadados	Na coleta de dados foram extraídos: título, autor, termo utilizado no escopo do trabalho (ORC; ORI; OC; OI; RC; RI), ano do evento, instituição procedente dos autores, palavras-chave utilizadas (quando apresentadas) e, no caso dos ENANCIB's, nome do GT2 na época da publicação. (Amostra 440, p. 33)
	Questionários	Foram construídos questionários, com questões abertas e fechadas, aplicados à população pesquisada, enviados por correio eletrônico. (Amostra 269, p. 101)
	Transcrições de entrevistas	Para a transcrição das entrevistas, utilizou-se a funcionalidade do Google Docs, que fez a transcrição em texto das gravações. (Amostra 225, p. 106)
	Anotações	No quadro sobre os conteúdos de maior interesse do colecionador, foram levantados, a partir das suas anotações manuscritas, todos os assuntos/conteúdos sinalizados por este na folha de rosto e nas páginas dos exemplares. (Amostra_164, p. 59)
	Padrões	A partir da análise dos padrões de metadados utilizados para descrever objetos de aprendizagem, foi possível identificar que a maioria deles está baseada em dois padrões de metadados: o Dublin Core (DC) e o Learning Object Metadata (LOM), conforme apresentado na seção 2.2. (Amostra 463, p. 35)
	Documentos diversos (leis, planos, relatórios, inventários etc.	A documental teve como base a consulta a fontes produzidas pela SBPC: relatórios, atas, regulamento, regimento, fotografias, programação, resumos,

		cartas, moções. (Amostra_557, p. 395)
	Produções bibliográfica científicas (livros, artigos, trabalhos de eventos, capítulos de livros etc.	Quanto à produção dos membros dos grupos de pesquisa, este estudo tratou apenas dos dados relacionados à produção bibliográfica, conforme Plataforma Lattes do CNPq (artigos completos publicados em periódicos, livros publicados/organizados ou edições, capítulos de livros publicados e trabalhos de eventos), compreendida no período de 1992, início do DGP do CNPq, até 2016 5. (Amostra_375, p. 64)
	Diário de campo	O Diário de Campo, enquanto instrumento de coleta de dados, foi utilizado desde a gênese até a conclusão da pesquisa, em que foram anotadas as percepções do pesquisador e utilizadas na síntese final dos dados. (Amostra_269, p.100)
	Cadernos de laboratório	A pouca ou inexistente sistematização dos dois modos de ciência e seus respectivos novos cadernos de laboratório demandou uma ampla pesquisa documental para delimitá-los, identificar suas origens e principais características em blogs, wikis e serviços <i>online</i> . (Amostra_141, p.31)

Fonte: Dados da pesquisa (2021)

A partir do Quadro 15, verifica-se a existência de uma heterogeneidade dos dados pertencentes ao mesmo grupo, o que demanda cada vez mais práticas específicas de gestão e curadoria dos dados. Outra diversidade de dados identificada remete ao grupo das visualizações, que está ilustrado no Quadro 16.

Quadro 16 - Dados em visualizações da Ciência da Informação no Brasil

Tipos de dados		Exemplos de segmentos de textos
Visualizações	Quadros	O Quadro 16 apresenta os dados referentes ao número de entrevistas realizadas, o tipo de organização e a função de cada entrevistado. (Amostra_200, p. 134)
	Figuras	Os dados extraídos podem ser visualizados na Figura 2. (Amostra_203, p.28)
	Tabelas	Deve-se reforçar o fato de que os dados da Tabela 10 refletem os dados da amostra desta pesquisa (somente as empresas inovadoras) e não a totalidade das bases de

		dados pesquisadas. (Amostra_78, p.126)
	Gráficos	Nesta pesquisa, foram utilizadas a lista de formas ativas (formes actives) que contém a ordem hierárquica da frequência do assunto no corpus, as análises de frequência de palavras através do gráfico Nuvem de Palavras, e análise contextual através do gráfico do Método Reinert. (Amostra 175, p.91)
	Mapas	O mapa conceitual, instrumento selecionado para sistematização e representação dos dados empíricos coletados neste trabalho, é considerado um sistema de conceitos. (Amostra 429, p.28)
	Modelos	A segunda etapa foi a de desenvolvimento do modelo visual do SRDigital tendo à frente do processo agentes computacionais usando técnicas da área de processamento da linguagem natural e linguagem controlada para formular a questão negociada. (Amostra 413, p.18)

Fonte: Dados da pesquisa (2021)

A natureza de visualizações representou os dados que foram escolhidos de forma preliminar para compor os critérios de definição das amostras, como também expressaram os dados validados para posterior análise. De acordo com Gil (2019), a apresentação de dados em tabelas, gráficos e diagramas é essencial para o processo de interpretação e análise dos dados.

Conforme o Quadro 16, foram identificados gráficos de nuvens de palavra e método Reinert, também foram constatados outros exemplos de tipos de dados como gráficos de superfície e do tipo diagrama de caixa (*box plot*), além disso, o mapa conceitual e o temático foram identificados como forma de visualização dos dados.

Os dados referentes a modelos compreenderam pesquisas realizadas em modelos, assim como a própria construção de modelos identificados como SRDigital, modelo-síntese, conceitual, extração de relações semântica, analíticos, dentre outros. O desenvolvimento de modelos pode ser considerado um dado como resultado que, segundo Schöpfel *et al.* (2016), remete aos dados produzidos na própria pesquisa.

Semelhante ao estudo realizado por Sales e Sayão (2019), foram reunidos no mesmo grupo de números segmentos dos textos correspondentes a medidas, resultados de levantamentos, resultados de experimentos, fórmulas, algoritmos e acrescentamos dados de

amostras que foram expressos numericamente nos casos analisados. No Quadro 17, estão apresentados segmentos dos textos para ilustrar os resultados obtidos.

Quadro 17 - Dados numéricos da Ciência da Informação no Brasil

Tipos de dados		Exemplos de segmentos de texto
Números	Medidas	Lago Junior (2005) cita várias medidas utilizadas na metodologia de análise de redes sociais, das quais, neste estudo, serão utilizadas três: de centralidade, de proximidade e de intermediação. (Amostra 351, p.)
	Resultados de levantamentos	Em relação ao levantamento, o banco de dados é composto por informações relativas às respostas e tratamentos estatísticos dos dados de 21 respondentes. Além disso, apresenta as citações referentes às perguntas abertas e os dados tabulados para informações de natureza quantitativa. (Amostra 315, p.78)
	Resultados de experimentos	Na última fase dos experimentos – Tempo e Corretude – são realizadas as Métricas de Corretude e Tempo e são produzidas as Estatísticas de Agrupamento (retângulo 6. da figura 18 anterior). (Amostra 307, p. 119)
	Fórmulas	Para tanto, realizou-se a organização dos dados em uma planilha do Microsoft Office Excel 2016, permitindo efetuar os seguintes cálculos: a porcentagem das respostas em cada nível da escala do item Likert, pela fórmula: $\text{=núm1} \times 100 / \text{total}$; a obtenção da moda, ou seja, a maior frequência de aparecimento das respostas em determinado nível da escala do item Likert, pela fórmula: $\text{=MODO (núm1: núm2,...)}$; e a média aritmética, medida de tendência central, a partir das respostas nos níveis da escala do item Likert, pela fórmula: $\text{=MEDIA (núm1: núm2,...)}$. (Amostra 483, p. 129)
	Algoritmos	Com os corpora de dados limpos, foi desenvolvido o terceiro algoritmo, 32 que contemplam as fases de pré-processamento, transformação, modelagem e processamento e apresentação dos resultados. (Amostra 255, p. 120)
	Resultados de amostras	Os indivíduos entre 18 e 24 anos representaram 18,2% da amostra, e

		aqueles de 50 a 64 anos e de 65 a 70 anos, juntos, 18,2% da amostra, sendo 9,1% em cada uma dessas últimas faixas etárias. (Amostra_404, p.110)
--	--	---

Fonte: Dados da pesquisa (2021)

As medidas referentes a cálculos estatísticos como moda, mediana e frequência também foram incluídas como dados numéricos. Michel (2015) enfatiza a importância do dado estatístico para tornar uma análise qualitativa mais robusta. De modo geral, neste momento foram obtidos os segmentos de textos que expressavam a quantificação dos dados.

Diante da diversidade de dados que foi levantada nas metodologias da CI, no grupo multimídia foram incluídos os dados em natureza de vídeos, áudios, fotografias e uma combinação desses, conforme os exemplos expressos no Quadro 18.

Quadro 18 - Dados de multimídia da Ciência da Informação no Brasil

Tipos de dados		Exemplos de segmentos de texto
Multimídia	Vídeos	Com o consentimento dos entrevistados, foi utilizada uma câmera de vídeo para gravar as entrevistas presenciais e o software Evaer76 para gravar as conversas via Skype. (Amostra_153, p. 89)
	Áudios	As entrevistas foram gravadas no formato de áudio, por meio de um aplicativo de gravação de áudio em telefone móvel. (Amostra_293, p.26)
	Fotografias	Iremos analisar as fotografias que foram doadas pelos fotógrafos profissionais Gilberto Morimitsu (Giba) e Nívia Uchôa. (Amostra_436, p. 69)
	Áudio, vídeo e imagem	O material coletado por registro em áudio, vídeo e imagens foi devidamente tratado e compartilhado com os grupos para auxiliar no processo de construção dos itens exibidos na Mostra Cultural. (Amostra_543, p.82)

Fonte: Dados da pesquisa (2021)

Os vídeos, os áudios e as fotografias identificadas nos textos corresponderam a materiais registrados que foram coletados para análise, assim também como produzidos na própria pesquisa, quando os vídeos e áudios foram gravados no momento da realização de entrevistas presenciais e/ou por *Skype*, bem como as fotografias que foram tiradas do ambiente pesquisado.

Em algumas metodologias, foi possível identificar uma combinação desses recursos multimídia, porém os áudios foram os mais comuns entre eles.

Em uma pequena proporção, com menos de 1% de participação nos casos analisados, apareceram os dados como *software*, nos contextos em que houve o estudo de *software*, assim como a sua criação, conforme apresentado nos exemplos do Quadro 19.

Quadro 19 - Dados em *software* da Ciência da Informação no Brasil

Tipos de dados		Exemplos de segmentos de textos
<i>Software</i>	<i>Software encontrado</i>	Dentre os resultados obtidos por esta pesquisa, está um software escrito em Prolog, que serve de implementação de referência do resultado sobre Ontoprolog. (Amostra 338, p.71)
	Criação de <i>software</i>	Para a fase de prototipação, os resultados obtidos por meio da observação participante bem como os saberes adquiridos a partir da pesquisa bibliográfica subsidiarão o protótipo do curso EaD em um software livre e gratuito. (Amostra 499, p.30)

Fonte: Dados da pesquisa (2021)

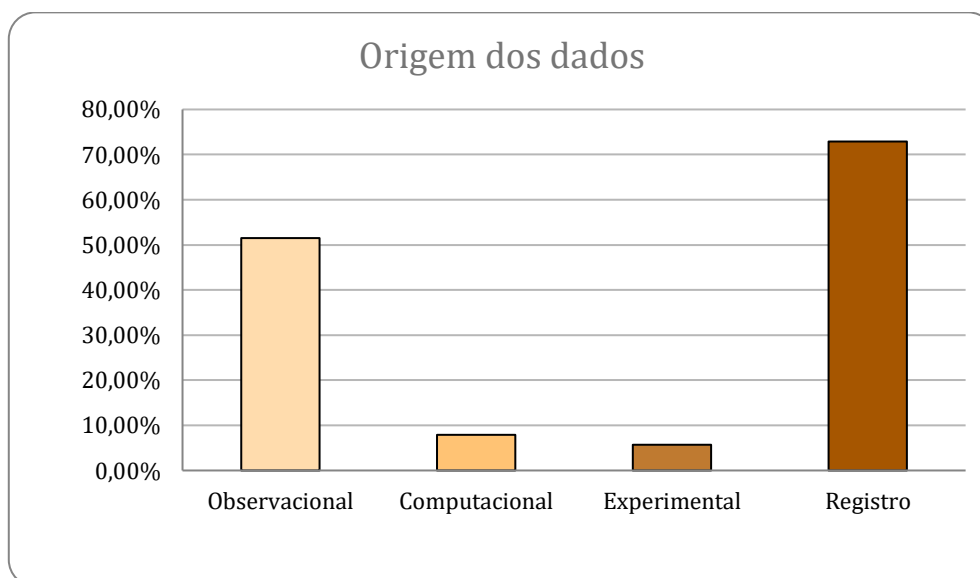
A natureza dos dados da CI faz parte da grande quantidade de trabalhos científicos existentes no Brasil, que produzem pequenas quantidades heterogêneas de dados. Esta realidade remete à reflexão sobre os dados da “cauda longa”, que compreendem, segundo Sayão e Sales (2019), uma forma específica de distribuir estatisticamente uma parte menor da população que tem muita ocorrência (cabeça) e uma parte maior apresenta menos ocorrência (cauda). Assim, os dados da CI se encontram na cauda. Neste contexto, para os autores, os dados são heterogêneos, espalhados entre instituições, projetos, laboratórios, pequenos grupos de pesquisas e pesquisadores individuais.

Esses dados precisam de uma curadoria adequada, não devem ser perdidos por falta de gestão, pois são recursos valiosos que podem enriquecer novos estudos. No que se refere ao valor dos dados, Silva (2019) destaca que eles são fundamentais para a ciência; Curty e Aventurier (2017) mostram que eles são primordiais para a criação de novos ciclos das investigações científicas.

6.3.2 Origem dos Dados

Os dados das pesquisas científicas podem ser de origem observacional, computacional, experimental e de registro (BORGMAN, 2015; NSB, 2005). A partir desta realidade discutida na literatura, foi possível identificar através do presente estudo que os dados das teses brasileiras da área da CI podem ser provenientes dessas 4 (quatro) categorias, conforme pode ser visualizado no Gráfico 4.

Gráfico 4 - Distribuição da origem dos dados das teses da Ciência da Informação no Brasil



Fonte: Dados da pesquisa (2021)

Os resultados evidenciam a diferença quanto aos percentuais encontrados nas categorias de origens dos dados (Gráfico 4). A maioria dos dados descobertos nos trabalhos de conclusão de curso de doutorado da CI no Brasil é de Registro, uma vez que esta categoria foi identificada em 167 teses, ou seja, 72,9% das 229 teses estudadas. Na sequência, destacaram-se os dados de origem observacional, que foram levantados em 118 casos, logo, dizem respeito a 51,5% do total da amostra.

Através do Gráfico 4, é possível observar que os dados computacionais foram constatados em 18 casos, ou seja, em apenas 7,9% dos trabalhos analisados. A origem experimental dos dados se deu ainda em menor proporção, sendo levantada em 13 metodologias, que correspondem a 5,7% dos casos analisados.

Segundo estudo realizado por Dias, Anjos e Araújo (2019), que investigaram a origem dos dados dos pesquisadores vinculados aos programas de Pós-graduação em Ciência da Informação no Brasil, em relação a serem observacionais, computacionais, experimentais e/ou uma combinação dessas categorias, foi identificado que a maioria dos dados desses

pesquisadores é de origem observacional ou observacional com correlação às demais categorias. Levando em consideração que o estudo dos autores não contemplou a categoria Registro, é possível identificar um alinhamento com o presente trabalho, assim fica demonstrado que é necessário atentar para as medidas de preservação dos dados da área da CI, principalmente pela presença significativa dos dados observacionais.

De acordo com os dados expressos no Quadro 20, é possível visualizar que um trabalho pode possuir mais de uma categoria de origem dos dados. Segundo a análise de correlação entre essas categorias, realizada no *software* QDA Miner, 4 (quatro) casos são de origem computacional e experimental, 4 (quatro) observacional e computacional, 11 de registro e computacional, 5 (cinco) experimental e observacional, 4 (quatro) experimental e de registro, e uma quantidade significativa de 77 casos correspondem às categorias observacional e de registro, que são as que mais originaram os dados das teses da CI no Brasil.

Quadro 20 - Coocorência entre a origem dos dados e a preservação e curadoria das teses da CI

	Computacional	Experimental	Observacional	Registro
Computacional	18	-	-	-
Experimental	4	13	-	-
Observacional	4	5	118	-
Registro	11	4	77	167

Fonte: Dados da pesquisa (2021)

Os dados observacionais por serem os menos reprodutíveis, Borgman (2015) os considera como os mais importantes para preservar. Sayão e Sales (2019) acrescentam que esses dados são provenientes de fenômenos e eventos únicos, só podem ser coletados uma vez, por isso necessitam estar preservados para sempre, de forma que estejam fidedignos, autênticos e íntegros, para isso os autores apontam o caminho do arquivamento em repositórios confiáveis, que possibilitem que ao longo prazo os dados se mantenham preservados.

O conhecimento sobre a origem dos dados se faz necessário para promover a compreensão sobre em que medida deve ser adotada a curadoria dos dados voltada a cada categoria. Não foi possível fazer a correlação das categorias de origens dos dados com a etapa preservar, pois conforme será discutido adiante, não foi identificado que os pesquisadores da CI, no âmbito das teses, adotaram alguma medida de preservação. Desta forma, é possível deduzir a existência de uma situação crítica, que afeta o potencial de reutilização futura dos dados por outros pesquisadores.

Sayão e Sales (2019) revelam que os dados experimentais são oriundos de experimentos que podem ser precisamente reproduzidos, logo, não necessitam de um arquivamento

indefinido como necessitam os dados observacionais. Porém, os autores alertam que as condições experimentais nem sempre são reproduzíveis, e, na maioria dos casos, o custo de refazer o experimento é maior do que gerenciar os dados. Desta forma, o relatório NSB (2005) mostra que por questões de custos e reprodutibilidade é fundamental a preservação desses dados.

Em relação aos dados da categoria computacional. Nesse caso, a preservação dos dados de saída pode não ser uma prioridade, entretanto é preciso que o próprio modelo e um conjunto robusto de metadados referente a ele estejam arquivados (NSB, 2005).

Os dados apresentados permitem inferir a existência de um cenário preocupante, pois os dados observacionais que são os que mais necessitam de preservação, também não estão preservados, assim como os dados de registro. Os pesquisadores da área da CI no Brasil necessitam adotar estratégias de acordo com as categorias que originam os seus dados, de modo que no futuro, eles possam ser reusados por outros pesquisadores. A seguir, é apresentada uma discussão sobre cada uma dessas categorias de acordo com os dados encontrados nos trabalhos estudados.

6.3.2.1 Categoria de origem dos dados – Registro

De acordo com Borgman (2015), o registro de qualquer fenômeno ou atividade humana pode remeter a dados para a pesquisa. Neste contexto, podem fazer parte deste tipo de dados: documentação de governos, empresas públicas e atividades privadas; livros e outros textos; materiais de arquivo; documentação na forma de áudio e gravações de vídeo, entre outros. Diante deste cenário, serão elencados alguns segmentos de textos recuperados das teses que demonstram pertencer a esta categoria, conforme apresentados no Quadro 21.

Quadro 21 - Dados de Registro da Ciência da Informação no Brasil

Características	Exemplos de segmentos de textos
Documentação de governos	É utilizada a pesquisa bibliográfica, na busca pela relação entre conceitos, características e ideias; e a pesquisa documental, por meio da análise de documentos organizacionais (normas, relatórios, manuais) e governamentais (legislação). (Amostra_89, p. 29)
Documentação de empresas públicas ou privadas	Durante a primeira etapa, foi realizada uma pesquisa documental na qual foram analisados os documentos institucionais para identificação dos processos internos da Engenharia da Petrobras relativos à inovação. Foram examinados documentos normativos que descreviam e definiam as normas e procedimentos

	para execução do processo, bem como informações gerais sobre os projetos e demais documentos das unidades organizacionais envolvidas. Todos estes documentos puderam ser acessados pela intranet da Engenharia. (Amostra 116, p.80)
Livros e outros textos	A pesquisa de tese se baseou no meio-modo formal na coleta e interpretação dos dados provenientes, na maioria, de documentos escritos, como a referencial do tema encontrado em livros, artigos e jornais. (Amostra 183, p.30)
Materiais de arquivo	Utilizaram-se também as informações do arquivo do Colégio Pedro II. Informações como as que estão contidas nos registros administrativos de escolas exigem cuidado especial, pois são dados de documentos oficiais que são coletados e organizados para um público restrito. (amostra 176, p. 101)
Documentação em áudio e gravações de vídeos	O material coletado por registro em áudio, vídeo e imagens foi devidamente tratado e compartilhado com os grupos para auxiliar no processo de construção dos itens exibidos na Mostra Cultural. (Amostra 543, p. 82)

Fonte: Dados da pesquisa (2021)

De modo geral, foram categorizados como dados de registros, os dados obtidos por meio de atas, relatórios, legislação, memorandos, organogramas, planos, dados estatísticos, arquivos, tabelas cedidas ou encontradas em algum *site*, fontes históricas, projetos políticos pedagógicos, padrões, catálogos de bibliotecas ou de universidades, matrizes curriculares, ementas, programas de disciplinas, documentação que pode ser considerada de governo e/ou documentação de empresas públicas ou privadas.

Além do levantamento realizado na literatura para a coleta de dados em livros e artigos, cujo levantamento ocorreu por meio de bases de dados nacionais e internacionais, incluíram-se os dados coletados tendo como fontes enciclopédias, dicionários, manuais e dados de outros estudos. Conforme apresentado por Borgman (2015), fazem parte dos dados de Registro, livros e outros textos.

A discussão proposta permite deduzir que existe uma coocorrência entre a categoria de Registros e os dados que são coletados por meio da pesquisa documental e da pesquisa bibliográfica, cuja realidade pode ser ilustrada no Quadro 22.

Quadro 22 - Coocorrência entre a origem de Registro e a coleta de dados

Tipo da coleta de dados	Registro (nº de casos)
Pesquisa bibliográfica	107
Pesquisa de campo	69
Pesquisa documental	104

Fonte: Dados da pesquisa (2021)

Os dados expressos no Quadro 22 foram obtidos ao analisar os textos das metodologias das teses da CI no *software* QDA Miner, versão 6.0. Por meio dele, é possível visualizar que 107 metodologias que são categorizadas como sendo de Registro têm relação com a pesquisa bibliográfica e 104 com a pesquisa documental, e um quantitativo menor de 64 trabalhos ocorre simultaneamente com a pesquisa de campo. Conforme informado anteriormente, foi identificado que 167 metodologias fazem parte da categoria de Registro. Ao fazer a correlação, observamos um número maior se somarmos as 3 (três) correlações, o que implica dizer que em um mesmo trabalho que é caracterizado como Registro, dados podem ter sido coletados por meio de uma pesquisa bibliográfica e uma pesquisa documental e/ou uma pesquisa de campo. Esta coocorência foi expressa neste trabalho, uma vez que é notória a relação da origem dos dados com a sua coleta (Quadro 22).

6.3.2.2 Categoria de origem dos dados – Observacional

Os dados de origem observacional compreendem registros históricos que não podem ser replicados, necessitando de uma preservação adequada (BORGMAN, 2015; NSB, 2005). Este fato gera uma alerta para a Ciência da informação, pois foi observado que os dados observacionais fazem parte de forma expressiva das metodologias das teses brasileiras na área da CI, ao serem identificadas em um pouco mais da metade dos casos, com um percentual de 51,5%, o que corresponde a 118 trabalhos.

Foram definidos alguns critérios com base na literatura para categorizar as sentenças dos textos como pertencentes à categoria observacional. Segundo Silva (2019), nas ciências sociais as observações dizem respeito ao comportamento de diversos grupos ou indivíduos.

Marconi e Lakatos (2019), ao discutirem sobre os fundamentos da metodologia científica, mostram que a observação direta intensiva ocorre por meio das técnicas de coleta de dados: observação e entrevista. Neste contexto, a observação pode ser assistemática, sistemática, não participante, participante, individual, em equipe, em laboratório, na vida real ou naturalista. Já a entrevista pode ser padronizada ou estruturada, não padronizada, semiestruturada ou livre. A observação direta intensiva inclui também o grupo de foco. No caso da observação extensiva, para os autores, ela se faz através de questionário, formulário, medidas de opinião e atitudes e técnicas mercadológicas.

Michel (2015), ao falar sobre a metodologia científica nas ciências sociais, complementa que a observação direta intensiva acontece quando há o contato direto e pessoal com a fonte e ela é extensiva quando esse contato não é necessário para adquirir informações. Outra maneira

de realizar a observação direta do campo é através de dados terciários, relatados por terceiros, cuja comprovação não pode ser obtida. Nesse contexto, as entrevistas, os questionários e outras formas de observação fazem parte da observação direta.

As sentenças dos textos referentes à categoria observacional foram categorizadas a partir da discussão dos autores (MARCONI; LAKATOS, 2019; MICHEL, 2015; SILVA, 2019). Alguns exemplos dos dados coletados estão expressos no Quadro 23.

Quadro 23 - Dados observacionais da Ciência da Informação no Brasil

Características	Exemplos de segmentos de textos
Comportamento de diversos grupos ou indivíduos	O segundo aspecto diz respeito à condição de se investigarem variáveis que ultrapassam a fala ou depoimento dos informantes. Foram considerados, também como elementos de análise, os comportamentos demonstrados pelos entrevistados, como também o "não dito" verbalmente, mas expresso em gestos, olhares, postura, constrangimentos. Essas variáveis só poderiam ser alcançadas por meio de métodos e técnicas que envolvam o contato direto e pessoal do pesquisador. (Amostra 19, p. 68)
Entrevistas	Nos casos em que foi localizada pouca informação, foi necessário realizar entrevistas. Nestas, os principais pontos observados foram analisados qualitativamente, pois tratava-se de uma pesquisa exploratória que pretendeu explorar um determinado campo para proporcionar maior familiaridade com o problema, tornando-o mais explícito. (Amostra 02, p. 134)
Questionários	Nesse sentido, foram criados três questionários distintos, enviados via e-mail e/ou disponibilizados em link para que os entrevistados (pesquisador / editor / gestor de repositório no Brasil e em Portugal) pudessem respondê-los online, com o intuito de identificar os principais elementos constitutivos dessa realidade, na perspectiva desses atores. Buscamos entender como cada um deles aborda e compreende o direito de autor em suas implicações econômicas, sociais e morais. (Amostra 114, p. 33)
Observação	A observação participante foi realizada durante os sete dias do evento, com registros escritos em diário de campo e, também, alguns gravados em áudio. Durante o mesmo período foram realizadas gravações de entrevistas tanto contextuais quanto semiestruturadas com um total de 143 visitantes, num total de 102 arquivos. (Amostra 85, p. 97)

Fonte: Dados da pesquisa (2021)

De modo geral, foram categorizados, como dados observacionais, os dados que foram provenientes de entrevistas, questionários e de vários tipos de observação que foram encontrados nos textos das metodologias pesquisadas: assistemática, sistemática, descritiva, comportamental, observação em laboratórios, entre outros.

Através dos segmentos de textos recuperados e utilizados para representar a categoria de origem observacional, infere-se que há coocorrência entre os dados observacionais e a pesquisa de campo. As técnicas que envolvem a observação direta estão presentes neste tipo de pesquisa, bem como possibilitam constatar como a teoria estudada e as variáveis propostas ocorrem em situações concretas e no ambiente que os fatos acontecem. Desta forma, compreendem dados advindos da vida real que trazem sentido à análise em pesquisas sociais, em que o objeto principal é o ser humano e o seu comportamento inserido em um determinado contexto social. (MICHEL, 2015). Para constatar essa realidade, das 118 metodologias que possuem dados de origens observacionais, 97 delas, que correspondem a 82,2%, são dados obtidos a partir da pesquisa de campo.

6.3.2.3 Categoria de origem dos dados – Computacional

Os dados originados da categoria computacional geralmente são identificados nas ciências físicas e da vida, porém também podem ser encontrados nas ciências sociais e humanas (BORGMAN, 2015). Esta realidade pode ser visualizada na presente pesquisa, embora não apareça em uma quantidade expressiva quando comparada aos dados de origem de Registro e Observacional; a categoria computacional pôde ser encontrada na CI, que é uma área do conhecimento pertencente às Ciências Sociais. Alguns segmentos dos textos analisados estão exemplificados no Quadro 24.

Quadro 24 - Dados computacionais da Ciência da Informação no Brasil

Características	Exemplos de segmentos de textos
Execução de modelos	O Modelo de Extração de Relações Semânticas que se apresenta nesta tese foi idealizado, de antemão, para ser uma construção teórica de um modo de explicitar relacionamentos entre conceitos de documentos acadêmicos, a fim de construir estruturas semânticas. Contudo, durante a elaboração do Modelo, compreendeu-se que sua implementação computacional apoiaria a investigação sobre a melhor maneira de extrair relações semânticas. Dessa forma, o Modelo se expandiu para um protótipo computacional. (Amostra 235, p. 24)
Simulação de computador	Nessa perspectiva, a aplicabilidade no plano das redes de conteúdos deu-se pela construção de redes, a partir da coleta e simulação dos dados relativos aos Indicadores de Qualidade da Educação Superior do INEP. Para a coleta e tratamento dos dados da pesquisa, propõe-se a utilização de software estatístico e de análise de redes sociais. (Amostra 457, p.89)

Fonte: Dados da pesquisa (2021)

A partir do relatório NSB (2005), que mostra que os dados computacionais são resultados da execução de um modelo ou simulação de computador, é que categorizamos os dados dos trabalhos analisados. Desta forma, os exemplos expressos utilizam ferramentas computacionais para originar os dados.

6.3.2.4 Categoria de origem dos dados – Experimental

O experimento é amplamente utilizado nas ciências naturais; o método tem o seu valor, sendo responsável por diversas contribuições que possibilitam o crescimento da ciência, porém, segundo Gil (2019), o experimento é pouco usado nas ciências sociais, o que é justificado pelo fato do cientista não possuir o poder de introduzir modificações nos fenômenos que pretende investigar. Essa realidade foi verificada no presente estudo, pois dados de origem experimental foram os menos identificados nos trabalhos pesquisados. Foram extraídos alguns segmentos dos textos que podem ser verificados no Quadro 25.

Quadro 25 - Dados experimentais da Ciência da Informação no Brasil

Características	Exemplos de segmentos de textos
Realização de experimentos	Foram realizados três experimentos de categorização utilizando RNA. No primeiro experimento foram utilizados o ano de publicação dos artigos e suas referências como dados de entrada da RNA. Neste experimento, o ano foi representado pelo seu valor absoluto. No segundo experimento, o ano foi representado pelo número correspondente aos seus dois últimos dígitos divididos por 10. No terceiro experimento, o ano de publicação dos artigos não foi utilizado como atributo. (Amostra_21, p.66) Propõe-se uma metodologia, para a aplicação de experimentos, que integre os princípios de indexação e a modelagem das nanopublicações. (Amostra_74, p.75). O objetivo geral deste estudo é verificar o efeito da interação do solucionador com o enunciado sobre o tempo utilizado e o escore alcançado. Uma vez que se cuida em observar relação de causa e efeito entre variáveis, adotou-se o método experimental, indicado para o estudo de tais tipos de relações. (Amostra_321, p. 149)

Fonte: Dados da pesquisa (2021)

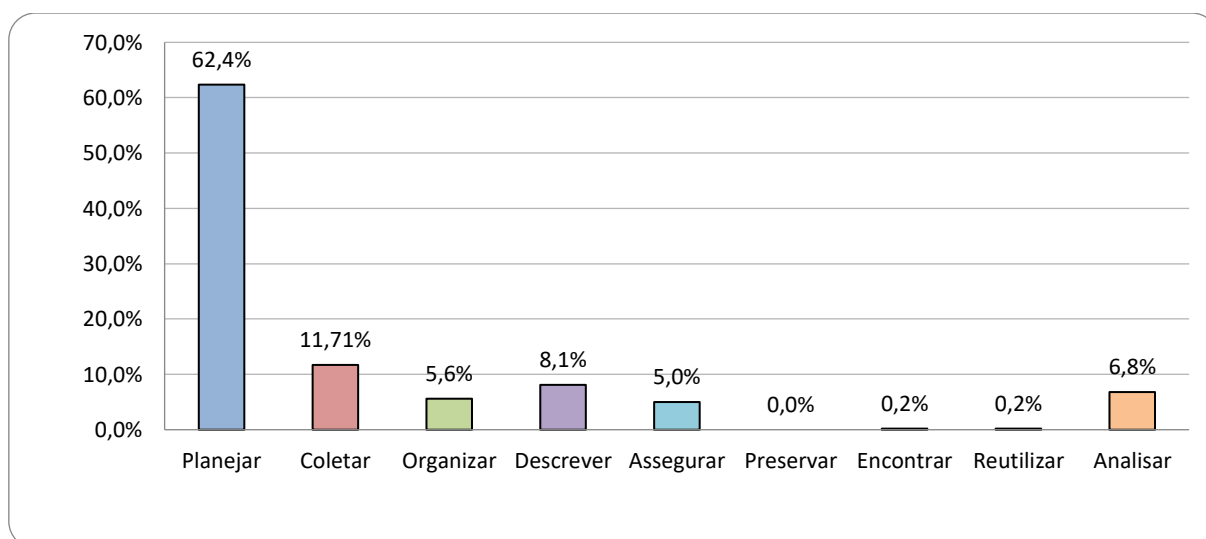
Conforme visualizado nos exemplos expressos no Quadro 25, apesar dessa categoria não ter sido comum nas metodologias, as poucas vezes em que experimentos foram realizados ocorreram por meio da aplicação do método experimental, na busca por determinar a relação

de causa e efeito, cujas investigações envolveram questões computacionais, teóricas, humanas e documentais.

6.4 INTERSECÇÕES ENTRE O CICLO DE VIDA DOS DADOS DE REFERÊNCIA E AS METODOLOGIAS DAS TESES BRASILEIRAS DA CIÊNCIA DA INFORMAÇÃO

Ao analisar a intersecção entre os elementos do modelo de ciclo de vida de dados científicos proposto no presente trabalho, baseado nas iniciativas do DataONE, DCC e DDI, com os percursos metodológicos das teses brasileiras da CI, a fim de atingir o objetivo geral desta tese, foi possível obter os resultados que podem ser visualizados no Gráfico 5.

Gráfico 5 - Intersecção entre o ciclo de vida dos dados de referência e as metodologias das teses da Ciência da Informação no Brasil



Fonte: Dados da pesquisa (2021)

Os dados mostram um alinhamento com a discussão dos autores que abordam os procedimentos metodológicos no contexto das Ciências Sociais, pois ao remeterem aos dados provenientes da pesquisa científica, esses autores enfatizam as etapas planejar, coletar e analisar. Porém, a diferença pode ser encontrada na presença de dados referentes à etapa descrever, o que será discutido adiante.

É possível visualizar que a etapa “planejar” apresentou uma maior frequência dos códigos atribuídos a cada etapa analisada com um percentual de 62,4%, seguida da etapa coletar, com 11,71%. Na sequência, aparecem as etapas com percentuais próximos: descrever com 8,1%, analisar com 6,8%, organizar 5,6% e assegurar 5%. As etapas encontrar e reutilizar

tiveram uma participação tímida na frequência dos segmentos de textos recuperados com as suas respectivas características e a etapa preservar não foi identificada. O que denota aspectos ausentes e que precisam ser mais explorados nos trabalhos científicos da CI. A seguir, essas etapas serão trabalhadas individualmente para uma melhor compreensão da realidade evidenciada no presente estudo.

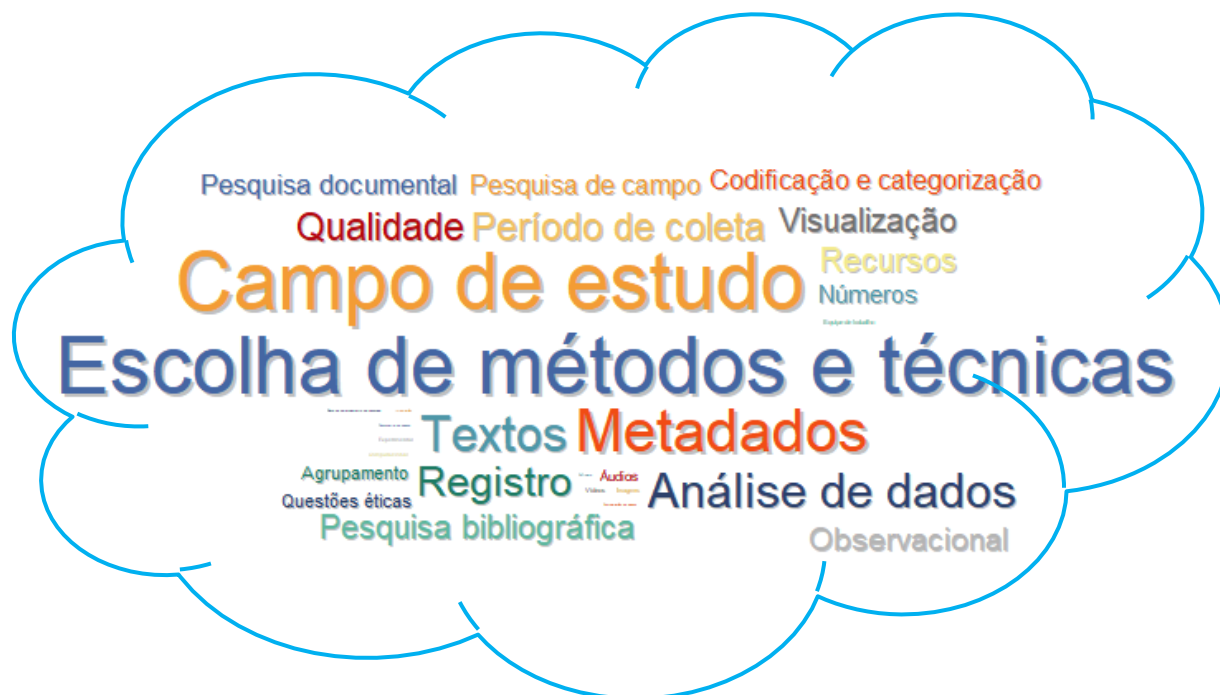
Na categoria “planejar”, foram identificados segmentos de textos referentes aos seguintes códigos: plano de gerenciamento do projeto, campo de estudo, período de coleta, escolha de métodos e técnicas, recursos, questões éticas, equipe de trabalho e tipo de dados em relação à natureza e à origem.

A categoria “coletar” recebeu 3 (três) códigos intitulados de pesquisa bibliográfica, pesquisa documental e pesquisa de campo. Foram levantados textos referentes aos códigos categorização e agrupamento da categoria “organizar”.

A identificação dos dados da categoria “descrever” ocorreu através do código metadados. Os dados da categoria “assegurar a qualidade” foram levantados através do código garantir a qualidade. A busca dos dados da categoria “preservar” foi realizada por meio do código preservação e curadoria; a categoria “encontrar”, através do código descoberta de dados; a categoria “reutilizar”, por meio do código integração; e, por fim, os códigos análise de dados e comparação foram utilizados para representar a categoria “analisar”.

A partir dos resultados correspondentes à participação dos códigos nas 229 metodologias estudadas, obteve-se, por meio do programa QDA Miner, a ilustração nuvem de palavras, que representa a participação dos códigos em relação ao número de casos, conforme Figura 20.

Figura 20 - Representação dos códigos identificados nas metodologias da Ciência da Informação



Fonte: Dados da pesquisa (2021)

Através da Figura 20, observa-se que os códigos visualizados em um tamanho de fonte maior são os que estão presentes no maior número de metodologias. Neste contexto, os códigos referentes à categoria planejar foram os que mais se destacaram. Infere-se que a etapa planejar tenha uma relação direta com as demais etapas do ciclo de vida dos dados.

6.4.1 Etapa “Planejar”

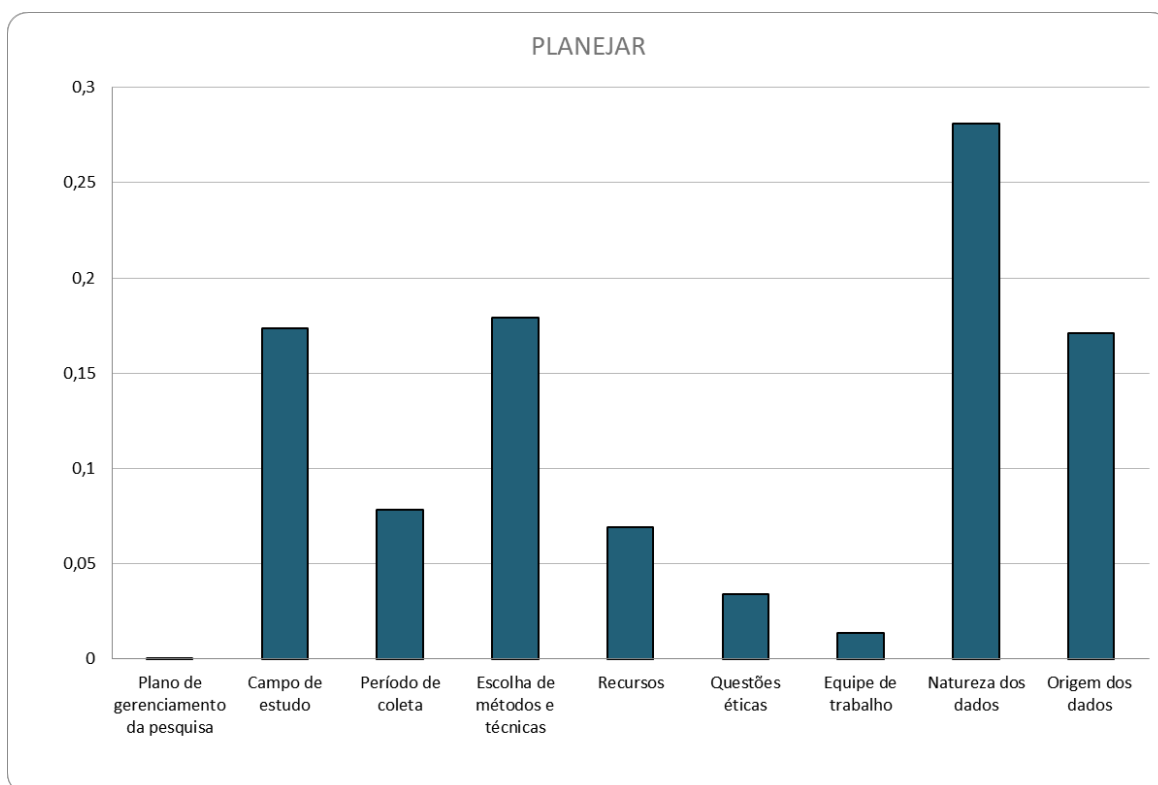
Neste momento da pesquisa, buscou-se atribuir códigos correspondentes à categoria planejar, no que se refere ao planejamento voltado aos dados científicos. Desta forma, com base no que foi levantado na literatura, observou-se que a etapa planejar envolve algumas ações apontadas por Strasser *et al.* (2012), tais como: é um momento de definir os dados da pesquisa, mostrar como eles serão coletados, os instrumentos utilizados e a definição da infraestrutura necessária. Quanto a esse quesito, Ionescu (2007) corrobora ao evidenciar ser um instante oportuno para apresentar as fontes de financiamento e a população de estudo.

Para Sales (2020), a gestão dos dados inicia com o planejamento do projeto de pesquisa e com a elaboração do plano de gestão de dados, o que envolve o trabalho do bibliotecário. Desta forma, faz parte deste contexto abordar aspectos sobre o armazenamento dos dados, dos responsáveis pelos dados e o planejamento sobre os metadados que serão utilizados com a

escolha de um padrão para descrever os dados (STRASSER *et al.* 2012). Esta etapa também envolve reflexões sobre direitos autorais e permissões de acesso (DCC, 2020). Assim sendo, essa etapa é o eixo central que possibilita organizar os dados de forma adequada, tendo em vista que nela deve ser elaborado o PGD que antevê a realização da pesquisa e relata os aspectos relacionados ao tratamento voltado aos dados científicos.

Diante da discussão dos autores supracitados, nesta etapa da pesquisa identificou-se, primeiramente, se os pesquisadores fizeram uso do plano de gestão de dados e do plano de gerenciamento do projeto; foram levantados os tipos de dados encontrados nas metodologias da CI, em relação à origem e à natureza. Foram evidenciados os métodos e as técnicas utilizadas para coleta e análise dos dados, o que também inclui os instrumentos de coleta de dados das pesquisas. Foi investigado também o campo de estudo, levantadas as questões éticas, equipe de trabalho, os recursos utilizados para oferecer suporte ao tratamento dos dados e possíveis ações sobre o planejamento da preservação e da descrição dos dados. Estas 2 (duas) últimas não foram constatadas. Os dados obtidos nos textos dos procedimentos metodológicos das teses da CI estão expressos no Gráfico 6.

Gráfico 6 - Etapa planejar e as metodologias da ciência da informação no Brasil



Fonte: Dados da pesquisa (2021)

Não foi possível recuperar nas metodologias evidências de que as teses pesquisadas utilizaram o plano de gestão de dados. Segundo a Fapesp (2020), o plano de gestão de dados passou a ser exigido por grande parte das agências fomentadoras da pesquisa públicas ou privadas da América do Norte, Austrália e de alguns países europeus (Grã-Bretanha, Holanda, Alemanha, países escandinavos) e pela própria instituição como requisito para concessão de auxílios as pesquisas. Veiga *et al.* (2019) mostram que os pesquisadores da Fiocruz também passaram a adotá-lo, o que denota uma realidade presente no cenário internacional, mas que também já chegou ao Brasil e, portanto, requer o despertar dos cientistas da informação para incluí-lo em suas práticas de pesquisa como forma de realizar a gestão dos seus dados científicos e auxiliar pesquisadores de outras áreas do conhecimento.

Também não foi identificado que os pesquisadores adotaram alguma medida prévia para a preservação dos seus dados e não apontaram a escolha de um padrão de metadados para descrevê-los.

O código que teve uma maior frequência em relação à categoria planejar foi a “natureza de dados”, seguido dos códigos “escolha de métodos e técnicas”, do “campo de estudo” e da “origem dos dados”. Essas frequências com quantificações mais significativas podem remeter ao fato de que em um mesmo trabalho foi possível atribuir o mesmo código mais de uma vez.

Em relação à natureza dos dados, por exemplo, em diversas metodologias, foi possível encontrar diferentes dados como, por exemplo, textos e números, além disso, é comum que as metodologias demonstrem, mesmo que de forma indireta, a natureza dos seus dados. O mesmo ocorreu com os demais códigos.

Diferentes métodos e técnicas foram identificados em um mesmo trabalho. Michel (2015) mostra que as técnicas que comumente são usadas nas pesquisas sociais correspondem à análise de documentos, à observação pessoal, às diversas modalidades de entrevistas e ao questionário. As técnicas evidenciadas pelo autor foram encontradas nas metodologias analisadas, principalmente a técnica de entrevista em algumas modalidades como exploratória, semiestruturada e estruturada e o questionário.

O presente estudo permitiu identificar nas metodologias analisadas os métodos que estão de acordo com Gil (2019), entre os mais utilizados nas pesquisas sociais: experimental, observacional, comparativo, estatístico e monográfico. Segundo o autor, que divide os métodos em dois grupos, tais métodos pertencem a um grupo que esclarece sobre os procedimentos técnicos necessários à pesquisa social, possibilitando uma direção para a realização da pesquisa, especialmente em relação à obtenção, ao processamento e à validação dos dados. Porém, além desses foi possível identificar os métodos dedutivo, indutivo, hipotético-dedutivo, dialético e

fenomenológico. Estes métodos, segundo Gil (2019), pertencem a um grupo que possibilita as bases lógicas da pesquisa científica.

Em relação aos métodos voltados à análise e interpretação dos dados, Richardson (2017) aponta a análise de conteúdo, a etnografia, a teoria fundamentada, a fenomenologia, a pesquisa histórica e a análise do discurso. Esta é considerada pelo autor como uma ferramenta essencial para a interpretação dos dados nas ciências sociais. De modo geral, esses métodos foram encontrados nas metodologias analisadas, porém o mais presente foi a análise de conteúdo.

Além dos métodos e técnicas apresentados, foram incluídos na codificação a bibliometria, o estudo de casos múltiplos, estudos bibliográficos, documental, pesquisa de campo, a análise de domínio, dentre outros, que foram apresentados como métodos e técnicas nos trabalhos analisados. No Quadro 26 estão exemplificadas algumas sentenças que foram encontradas no material pesquisado.

Quadro 26 - Métodos e técnicas utilizados nas teses brasileiras da Ciência da Informação

Métodos e técnicas	Exemplos de sentenças
Métodos Comparativo e estatístico	Quanto aos métodos que indicam os meios técnicos de investigação, optou-se pelos tipos comparativo e estatístico. (Amostra_275, p. 70)
Método indutivo	Quanto ao método de abordagem escolhido, esta pesquisa utilizará o método indutivo. (Amostra_224, p. 35).
Método experimental	Dentro da abordagem quantitativa, o estudo utiliza o método experimental. (Amostra_321, p. 151)
Bibliográfica, documental e estudo de caso	Serão adotados como procedimentos técnicos a pesquisa bibliográfica, a pesquisa documental e estudos de caso em duas organizações públicas que utilizam mídias e redes sociais para comunicação com os clientes. (Amostra_247, p. 125)
Questionário, entrevista e observação	Adota-se o enfoque descritivo, e a técnica utilizada são o questionário, a entrevista e a observação. (Amostra_323, p.94)

Fonte: Dados da pesquisa (2021)

A partir da coleta dos dados, foi evidenciado que a escolha dos métodos e das técnicas está relacionada com a coleta e análise dos dados, por isso é importante definir previamente os instrumentos necessários para a obtenção dos dados, assim como definir como eles serão analisados.

Para representar os demais códigos da categoria planejar, que ainda não foram exemplificados neste trabalho, através do Quadro 27 é possível visualizar alguns segmentos de textos que foram extraídos e que remetem a cada código.

Quadro 27 - Etapa planejar nas teses brasileiras da Ciência da Informação

Categoria	Códigos	Exemplos de segmentos de textos
Planejar	Plano de gerenciamento de projeto	Os processos gerenciais executados nesta pesquisa tiveram as seguintes responsabilidades: desenvolvimento de plano de gerenciamento do projeto, coleta de requisitos, definição de escopo, definição de atividades, sequenciamento de atividades, estimativa de duração das atividades, desenvolvimento de cronograma, controle de cronograma, e encerramento de projeto ou fase. (Amostra_373, p. 40)
	Campo de estudo	Diante disso, o campo empírico é composto pelo Museu da República, onde trabalhamos com o arquivo Armindo Marcílio Doutel de Andrade, a Biblioteca Nacional, onde foi selecionado o arquivo de Afonso Henriques Lima Barreto, a Fundação Casa de Rui Barbosa, com o arquivo de Américo Jacobina Lacombe, e o Arquivo Nacional, a partir do fundo Lucia Velloso Maurício. (Amostra_02, p. 136)
	Período de coleta	A coleta de dados foi realizada no dia 10 de abril de 2018, assim, os dados referem-se às informações constantes disponibilizadas pelos programas até esta data. (Amostra_209, p.110)
	Recursos	As respostas dos questionários foram transferidas para uma matriz de dados, utilizando o programa Excel. Esta matriz foi migrada para o programa estatístico da International Business Machines (IBM), o SPSS Statistics 24, que possibilitou as operações estatísticas, cruzamento de dados, criação de tabelas e gráficos. (Amostra_144, p. 139)
	Questões éticas	Após o aceite, foi enviado o Termo de Consentimento Livre e Esclarecido (TCLE) ¹³ , composto pelas diversas informações sobre a pesquisa e solicitação de autorização para uso e divulgação dos dados em contextos acadêmicos e publicações científicas, assim como foram marcados data e horário para a realização da entrevista. (Amostra_80, p. 25)

	Equipe de trabalho	Após a delimitação dos objetivos, hipóteses e das questões da pesquisa, foram avaliadas e discutidas as formas de registrar os conjuntos de dados de interesse da pesquisa e analisadas algumas experiências em conjunto com a orientadora e coorientadora. (Amostra_502, p.37)
--	--------------------	---

Fonte: Dados da pesquisa (2021)

Em relação ao plano de gerenciamento de projeto, apenas 1 (um) caso, que está demonstrado no Quadro 27, evidenciou a adoção da ferramenta. Assim, através das metodologias não foi identificado que a gestão dos dados ocorre por meio do plano de gestão de dados e não acontece pelo plano gerenciamento do projeto.

Para selecionar os trechos dos textos referentes ao campo de estudo, foram codificados os segmentos com os dados sobre o local da pesquisa, os critérios e estratégias para a escolha de quem fez parte da pesquisa (DESLANDES, 2009). Desta forma, esse código apresentou um percentual significativo em detrimento dos demais, pois nele foram incluídos os sujeitos participantes das pesquisas, o local da pesquisa, a pesquisa sobre a produção científica em uma determinada área ou documento de um domínio, as estratégias de buscas, envolvendo os critérios de inclusão e os termos de buscas, quando a pesquisa foi realizada em bases de dados selecionadas.

As buscas recuperadas em relação ao período de estudo foram obtidas no tocante ao período que os dados foram coletados, por meio de entrevistas, através de questionários, quando aplicados presencialmente, assim como o período que eles foram enviados por correio eletrônico e ficaram disponíveis para que fossem respondidos. Também foram feitas buscas pelos períodos de observações, o tempo da duração da coleta, dentre outros.

Por meio dos dados apresentados no Quadro 27, foi constatado que as questões éticas identificadas nos trabalhos estão relacionadas com diversos aspectos, incluindo a entrega e o preenchimento do Termo de Consentimento Livre e Esclarecido (TCLE), com a autorização para a divulgação de dados em outros contextos científicos. Todavia, observou-se o compromisso do pesquisador em utilizar os dados apenas para fins específicos da respectiva pesquisa realizada. Esta realidade afeta o compartilhamento e o reúso dos dados. Além dessas questões, em observância aos princípios do anonimato e da confidencialidade, os pesquisadores da CI buscaram preservar a identidade dos participantes das pesquisas e das organizações, cujas representações nos trabalhos ocorreram por meio de códigos, mantendo assim o sigilo dos dados. Outras questões apontadas foram a aprovação de Comitês de ética e pesquisa em

investigações que envolvem seres humanos, a autorização das organizações antes da realização das pesquisas e o consentimento dos entrevistados para gravação de entrevistas e para fotografar o ambiente e/ou as pessoas investigadas.

Os recursos que foram identificados nos trabalhos referem-se à utilização de *software* para a coleta, organização e análise dos dados, a exemplo do MaxQDA, SPSS, Nvivo, Atlas. Ti. Microsoft Excel, dentre outros. Também foram identificadas outros programas que possibilitam a aplicação de questionários e a transcrição de entrevistas. Foram levantadas ferramentas que possibilitam a gestão de documentos e referências bibliográficas como Mendeley e o EndNote. No processo da coleta de dados, houve gastos com transportes e alimentação, identificou-se a utilização de gravadores para aplicações de entrevistas e foi possível contar com apoio financeiro como, por exemplo, o disponibilizado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Quanto à equipe de trabalho, geralmente era composta pelos pesquisadores, orientadores e coorientadores (Quadro 27). Foi possível observar a colaboração de outros professores, ou seja, de pessoas que foram consideradas como colaboradoras por auxiliarem em situações envolvendo, por exemplo, grupos focais. Alguns trabalhos demonstraram a atuação do próprio pesquisador no processo da coleta de dados e a participação de bolsista e grupo de pesquisa.

6.4.2 Etapa “Coletar”

Para Gil (2019), os dados de uma pesquisa científica podem ser obtidos de formas variadas, através de uma pesquisa de campo, de uma pesquisa bibliográfica e/ou uma pesquisa documental. Partindo desta reflexão, entendendo que a etapa coletar compreende o momento da busca efetiva pelos dados com base no que foi planejado, buscou-se apontar como os dados foram obtidos a partir dos procedimentos revelados pelo autor. No Quadro 28, estão expressos exemplos encontrados nas metodologias que possibilitam uma melhor compreensão da realidade.

Quadro 28 - Etapa coletar nas teses brasileiras da Ciência da Informação

Categoria	Códigos	Exemplos de segmentos de textos
Coletar	Pesquisa bibliográfica	Do ponto de vista dos procedimentos técnicos, optou-se, em seu aspecto empírico, pela realização inicial de uma coleta de dados para levantar as informações sobre os colégios de aplicação das universidades brasileiras; e uma pesquisa bibliográfica, que permitiu uma análise das contribuições dos diversos autores que já abordaram as temáticas que envolvem a questão para alcançar os objetivos propostos. (Amostra 93, p.186)
	Pesquisa documental	A fim de identificar a importância do Museu estudado e de seu periódico científico, na implantação e consolidação da ciência na Amazônia, as fontes pesquisadas foram documentos de diversas instâncias, atuais ou pretéritos, tais como: relatórios, cartas, ofícios, leis, decretos, discursos, livros, artigos, mapas, fotografias, currículos, microfilmes, os próprios periódicos da instituição, entre outros. A utilização e análise deste tipo de documento caracteriza a pesquisa como documental e bibliográfica. (Amostra 135, p.46)
	Pesquisa de campo	Para a pesquisa de campo, como instrumento de coleta de dados foram utilizados questionários e entrevistas semiestruturadas. O critério utilizado na decisão da opção pela entrevista e pelo questionário foi o fato de se tratar de uma pesquisa mista, pois estes se complementam, possibilitando, por meio do conteúdo das respostas a serem obtidas nos dois instrumentos, atingir os objetivos. (Amostra 399, p. 229)

Fonte: Dados da pesquisa (2021)

Os segmentos de textos que foram codificados como pesquisa bibliográfica remetem aos estudos, cujos dados foram obtidos por meio das pesquisas acadêmicas encontradas em bases de dados de dissertações, teses, periódicos, anais de eventos, *google* acadêmico, livros, enciclopédicas, almanaques e repositórios de dados científicos. As pesquisas foram classificadas dessa forma, quando tinham por objetivo a elaboração de algum modelo e

utilizaram a literatura como parâmetro, assim como no contexto em que os pesquisadores buscaram aprofundar alguma temática abordada nos estudos investigados. Para Gil (2019), quando os dados são obtidos por meio da exploração de livros, artigos de periódico, anais de eventos e impressos variados, a pesquisa se configura como bibliográfica. Essas fontes consultadas ocorreram por meio digital e por meio de documentação impressa.

A pesquisa documental pode ocorrer quando os dados são levantados por meio de arquivos, análise de documentos ou análise de artefatos (GIL, 2019). No contexto dos trabalhos pesquisados, os dados coletados tiveram como fontes diversos tipos de arquivos eletrônicos, pessoais e de universidades, registros administrativos, memorandos, relatórios, atas, cursos, currículos, ementas, dados do Instituto Brasileiro de Geografia e Estatística (IBGE), cadernos de indicadores da CAPES, legislação, projetos políticos-pedagógicos, documentos institucionais, administrativos, e históricos, além de coleções fotográficas, indicadores de qualidades, dentre outros. Muitos desses documentos foram alcançados por meio de portais e *sites*, outros foram disponibilizados em materiais impressos, cópias de documentos ou fontes escritas.

No que concerne à pesquisa de campo, os contatos diretos envolvem pessoas que além de serem fontes de dados podem apontar possíveis caminhos para a coleta de dados, podendo ocorrer em paralelo à pesquisa bibliográfica (MARCONI; LAKATOS, 2019). Desta forma, conforme apontado por Gil (2019), os dados podem ser levantados através da observação, aplicação de questionários e entrevistas. Esta realidade foi verificada nos percursos metodológicos das teses analisadas, por meio de grupos focais, visitas técnicas e preenchimento de formulários. Neste contexto, os dados foram coletados de forma digital, quando, por exemplo, as entrevistas ocorreram por meio de *Skype* e os áudios e vídeos foram gravados. Assim também quando foram coletados de forma física, ao serem realizadas anotações no diário de campo.

Sales e Sayão (2020) revelam que a materialização dos dados pode ocorrer de forma digital ou física. A identificação dessa realidade possibilita visualizar o tratamento necessário para os dados, pois ao serem nascidos digitais ou tendo passado por um processo de digitalização, podem apresentar a fragilidade dos objetos digitais e a obsolescência tecnológica. Por isso, esses dados necessitam de um tratamento específico, com a finalidade de assegurar a sua autenticidade e integridade.

É possível que os dados físicos obtidos nas teses da CI no Brasil tenham passado pelo processo de digitalização, já que para Siebra, Borba e Miranda (2016), digitalizar os dados analógicos compreende uma prática comum entre os cientistas. Além disso, pela identificação

da presença de dados digitais nesses trabalhos, são necessários cuidados de curadoria específicos, conforme será explicado na etapa preservar.

Por meio do Quadro 29, é possível visualizar que dos 130 trabalhos que têm os seus dados coletados por meio da pesquisa bibliográfica, 44 também são coletados através da pesquisa de campo e 56 por meio da pesquisa documental. Em relação aos 107 trabalhos cujos dados foram coletados através da pesquisa de campo, 54 percursos metodológicos incluem dados que foram obtidos a partir da pesquisa documental.

Quadro 29 - Correlação entre os códigos da etapa coletar

Etapas	Pesquisa bibliográfica	Pesquisa de campo	Pesquisa documental
Pesquisa bibliográfica	130	-	-
Pesquisa de campo	44	107	-
Pesquisa documental	56	54	109

Fonte: Dados da pesquisa (2021)

Desta forma, mesmo havendo uma predominância de dados obtidos a partir da pesquisa bibliográfica, os dados da CI são coletados por meio de distintas fontes, podendo ser até combinadas. Em alguns trabalhos não foi possível constatar a forma que os seus dados foram obtidos, por apresentarem uma metodologia muito sucinta, o que dificultou o processo da coleta de dados.

6.4.3 Etapa “Organizar”

A etapa “organizar” envolve as ações referentes ao agrupamento, codificação e categorização dos dados, por isso os segmentos dos textos que estavam relacionados com a organização dos dados receberam essas respectivas codificações, conforme apresentado no Quadro 30.

Quadro 30 - Etapa organizar nas teses brasileiras da Ciência da Informação

Categoria	Códigos	Exemplos de segmentos de textos
Organizar	Categorização e codificação	Com esse material de trabalho em mãos, as entrevistas foram ouvidas novamente, lidas e relidas, e as respostas, organizadas e analisadas segundo quatro categorias de análise [...] (Amostra 257, p. 75)
	Agrupamento	Após a coleta de dados das entrevistas, dos processos de

		observação participante e da pesquisa documental, os mesmos foram agrupados, categorizados e comparados para se estabelecerem relações entre as empresas pesquisadas. (Amostra 247, p.130)
	Categorização e agrupamento	Em um segundo momento, as unidades de registro, que são os recortes das falas dos participantes da pesquisa, foram definidas e selecionadas de acordo com os objetivos estabelecidos no presente estudo. Posteriormente, foram agrupadas e classificadas em categorias e, algumas, em subcategorias de análise, por meio da operação de categorização. Ou seja, as unidades de registro foram devidamente agrupadas segundo elementos de semelhança para, depois, serem alocadas nas categorias e subcategorias estabelecidas. (Amostra 468, p.88)

Fonte: Dados da pesquisa (2021)

Em relação ao agrupamento, foi percebido que este se configura como o momento de reunir os dados que apresentam alguma similaridade, a partir deste processo é possível identificar as diferenças e agrupar as semelhanças ou os relacionados. O que também está atrelado à descrição, uma vez que os dados são agrupados de acordo com o elemento descritor.

Algumas formas de organizar os dados nos trabalhos analisados ocorreram através de quadros, planilhas do Excel, em pastas e com auxílio de ferramentas como o EndNote e o Zotero.

De acordo com as metodologias analisadas, os dados foram estruturados em categorias, sendo estas estabelecidas por meio da literatura, assim como a partir dos resultados encontrados, logo, foram criadas com base em algo para assim estruturar os dados. Em relação aos códigos, conforme abordado nas questões éticas, eles foram estabelecidos, por exemplo, para a identificação dos sujeitos da pesquisa.

Segundo Bardin (2011), o processo de categorização pode ocorrer de duas formas: o sistema de categorias pode ser fornecido previamente ou não. No primeiro caso, à medida que os elementos são encontrados e vão sendo distribuídos a elas da melhor maneira possível; no segundo caso, a classificação das categorias resultará dos elementos encontrados.

De modo geral, o agrupamento, as categorias e códigos são parte da organização do material para posterior análise. Nesse processo, a organização está atrelada à análise dos dados, uma vez que se configura como uma parte da etapa de analisar. Diante da discussão proposta,

os resultados encontrados que demonstram a correlação entre a etapa organizar e as etapas descrever e analisar estão expressos no Quadro 31.

Quadro 31 - Correlação entre as etapas organizar, descrever e analisar

-	Agrupamento	Análise de dados	Codificação e categorização	Metadados
Agrupamento	60	-	-	-
Análise de dados	53	158	-	-
Codificação e categorização	33	77	88	-
Metadados	51	114	71	140

Fonte: Dados da pesquisa (2021)

Os dados obtidos revelam que das 60 metodologias em que foram encontradas o código agrupamento, da categoria organizar, 53 delas coocorrem com o código análise de dados, da categoria analisar, e 51 com o código metadados, da categoria descrever. Os códigos da categoria organizar também coocorrem entre si em 33 casos identificados.

A codificação e categorização apresenta uma coocorrência mais expressiva com a análise de dados em 77 metodologias. Este fato pode ser justificado pela identificação da presença significativa da análise de conteúdo como técnica adotada pelos pesquisadores. E existe a correlação com metadados em 71 casos. Esta correlação demonstra que o processo de organização dos dados está atrelado aos descritores dos dados, conforme já enfatizado, além de que o ato de alocar os dados nos elementos descritores se configura como uma organização, a partir dos achados da pesquisa. Por isso, a organização deve ser incluída como etapa no processo de ciclo de vida dos dados.

6.4.5 Etapa “Descrever”

Strasser *et al.* (2012), ao tratarem da etapa descrever do CVD do DataONE, mostram ser ela um momento que possibilita que os dados possam ser entendidos no futuro, ao dispor de uma documentação adequada que inclua o contexto do arquivo de dados, o que é intitulado de metadados. Assim, algumas questões necessitam ser levadas em consideração, a saber: a descrição do contexto digital com a nomenclatura do conjunto de dados; o nome do (s) arquivo(s) de dados, data da última modificação do conjunto de dados; o *software* (com a informação sobre o número da versão) utilizado para trabalhar com o conjunto de dados; e o processamento de dados usado. É preciso também ter uma descrição do pessoal e das partes interessadas; quem coletou os dados; quem deve ser considerado como fomentador. Além de

apresentar o contexto científico, justificando a razão da coleta dos dados e apresentar os dados que foram coletados, quais os instrumentos foram utilizados e as condições ambientais no momento da coleta.

A partir do material analisado, não foi possível identificar através das metodologias que as teses da CI apresentam uma documentação completa dos seus dados ou que fizeram uso de um padrão de metadados para descrevê-los. Apenas foram visualizados alguns elementos nas metodologias que contém alguma descrição dos dados, como quem coletou os dados, a parte interessada, quem foi considerado o fomentador, os instrumentos adotados para coleta, os programas utilizados no tratamento dos dados, a extração de metadados de base de dados, os formatos, os nomes dos arquivos e a linguagem dos dados. Portanto, os trechos com esses metadados foram codificados por remeterem a algum elemento pertencente a este contexto, geralmente eles foram encontrados de forma aleatória no material pesquisado. A recuperação de alguns segmentos está demonstrada no Quadro 32.

Quadro 32 - Etapa descrever nas teses brasileiras da Ciência da Informação

Categoria	Características	Exemplos de segmentos de textos
Descrever	Quem coletou os dados	A equipe que aplicou o grupo focal foi composta do pesquisador, que exerceu o papel principal de conduzir as discussões, auxiliado por um relator, que anotou os pontos discutidos para servir de base na análise dos dados coletados. (Amostra_331, p. 107)
	Descritores textuais	Com base na coleta desses dados, foi elaborada uma tabela de termos e conceitos para cada artigo do periódico e, na tabela, identificam-se: Título do artigo, autor, termos e conceitos extraídos. (Amostra_411, p. 131)
	Instrumento de coleta	Para a pesquisa de campo, como instrumento de coleta de dados foram utilizados questionários e entrevistas semiestruturadas. (Amostra_399, p. 229).
	<i>Software</i>	Para a tabulação dos dados bibliográficos devidamente organizados, foram utilizados dois softwares: o Excel da Microsoft e o IBM SPSS Statistics Desktop 20. (Amostra_25, p. 103)

Fonte: Dados da pesquisa (2021)

Os trechos codificados na categoria descrever correspondem, predominantemente, aos instrumentos utilizados para a coleta de dados como as entrevistas, os questionários, o diário de campo e de visitas, os grupos focais, entre outros; à identificação do *software* utilizado para a organização e/ou análise dos dados, como o SPSS, Matlab (Matrix Laboratory), Excel, Atlas Ti (dentre outros) e ao formato dos arquivos de dados, PDF, CSV, HTML, TXT; por fim, outros descritores aparecem em menores proporções, como quem coletou os dados e os fomentadores, o nome dos arquivos e as linguagens dos dados.

Esta realidade demonstra que os dados encontrados nas metodologias não foram adequadamente representados pelos pesquisadores da CI no Brasil, uma vez que não existe uma documentação robusta dos dados que os contextualize, o que afeta a preservação e, conseqüentemente, o compartilhamento e o reuso dos dados. Segundo Borgman (2015), os dados não possuem significado isoladamente, mas sim dentro de uma infraestrutura de conhecimento, de uma ecologia de pessoas, práticas, instituições, objetos materiais e relacionamentos.

6.4.6 Etapa “Assegurar a Qualidade”

O momento de assegurar a qualidade envolve os processos de coleta e análise de dados. (STRASSER *et al.* 2012). Ao realizar o presente estudo, foi observado que os pesquisadores da CI demonstram assegurar a qualidade dos seus dados quando na eliminação de documentos duplicados ou no descarte de alguns dados. Neste contexto, foi visto que para seleção de dados foram adotados parâmetros ou diretrizes, o que está de acordo com a etapa de avaliação e seleção do DCC (2020).

Conforme aponta Ionescu (2007) sobre a realização de checagens, foram identificados em alguns trabalhos a revisão dos dados e a releitura quando se trata de dados textuais. Para proporcionar autenticidade aos dados, foi verificado, em alguns textos que utilizaram o instrumento da entrevista, que as falas dos entrevistados foram mantidas na íntegra para evitar uma interpretação equivocada dos dados, com a manutenção das falas literais para manter os dados fidedignos. E para manter a confiabilidade dos dados, foi identificada a realização de pré-testes para validar os instrumentos de coleta os dados, o que provocou ajustes de questionários e entrevistas, buscando trazer uma maior confiabilidade ao processo investigativo. Ainda fizeram parte dessa categoria questões como limpeza de dados inconsistentes e mudanças de formatos de arquivos. Alguns exemplos estão elencados no Quadro 33.

Quadro 33 - Etapa assegurar a qualidade nas teses brasileiras da Ciência da Informação

Categoria	Características	Exemplos de segmentos de textos
Assegurar a qualidade	Migração de formato	A próxima etapa está na conversão dos arquivos contidos nos corpora, de maneira que os dados possam ser interpretados e processados pelo computador. Para isso, foram desenvolvidos dois códigos. O primeiro código ³⁰ lista os arquivos encontrados no diretório e subdiretório apontado e já organizado. O código gera um arquivo TXT com o caminho de todos os arquivos encontrados. Já o segundo código ³¹ realiza a leitura do arquivo gerado pelo primeiro código, localizando todos os arquivos no formato PDF do corpus e realizando a conversão dos textos para um único arquivo no formato TXT, um padrão interpretável pelo computador. (Amostra_255, p. 119)
	Dados descartados	Feita a normalização dos dados, ou seja, após retirar os dados inconsistentes, incompletos e alocados em local equivocado, restaram 14.333 artigos de periódicos científicos publicados por pesquisadores e estudantes membros dos grupos de pesquisa das Áreas de Informação no Brasil, considerando o intervalo de 1992 até 2015. (Amostra_375, p. 68)
	Limpeza dos dados	Depois da planilha totalmente completa, foi executada uma apurada limpeza e organização dos dados. Na sequência, para analisar a colaboração, os artigos que possuíam apenas um autor foram excluídos. (Amostra_351, p.91)
	Aplicação de pré-teste	A aplicação do pré-teste revelou a necessidade de se efetuar algumas modificações no questionário. Ao final desta etapa, o questionário foi validado e foram definidas trinta e sete questões. (Amostra_247, p. 130)
	Trabalhos duplicados	Após a seleção inicial pelo tipo de publicação, idioma e disponibilidade de texto completo, 203 artigos passaram por leitura do título e resumo visando avaliar sua pertinência. Desses 38, foram descartados. (Amostra_241, p. 130)

Fonte: Dados da pesquisa (2021)

Esse processo de assegurar a qualidade dos dados está ligado à coleta, pois a própria realização de pré-testes para ajustar o instrumento já se configura um momento de coletar dados. Outra questão importante apontada por Strasser *et al.* (2012) é a descrição das situações ocorridas no processo de coleta de dados que possam impactar na qualidade dos seus dados. De modo geral, foi possível identificar ações que contribuem com a qualidade dos dados, porém a ausência de uma documentação que apresente a descrição dos dados encontrados acaba por prejudicar a próxima etapa e pode comprometer o processo de curadoria dos dados.

6.4.7 Etapa “Preservar”

A partir da análise inerente à etapa preservar, foi observado que os pesquisadores da CI se limitam de forma predominante a armazenar os seus dados em planilhas do Excel, ou seja, em um *software* de planilhas. Foi constatado que os dados também foram gravados em CDs, pastas do computador e arquivos em dispositivos de armazenamento. Raramente, os dados são armazenados em infraestruturas disponibilizadas por bancos de dados, bases de dados e Zotero. Essa realidade prejudica a visibilidade dos dados e, pela obsolescência tecnológica, esses dados podem não mais existir em um curto período.

Sayão e Sales (2019), ao discutirem sobre o maior volume de dados produzido pela ciência que se encontra na cauda longa, mostram que, neste contexto, os pesquisadores geralmente realizam a gestão dos seus dados por meio de ferramentas *off-the-shelves*, a exemplo de planilhas eletrônicas, porém grande parte desses dados não é submetida a ações de gestão e curadoria e com o tempo desaparece em razão da obsolescência tecnológica – nos discos dos computadores particulares dos pesquisadores e nos dispositivos portáteis de armazenamento.

De modo geral, poucos segmentos de textos demonstraram que os pesquisadores armazenaram os dados em mídias, porém não houve indícios de preservação e curadoria dos dados. Os resultados obtidos mostram que a Ciência da Informação no Brasil se encontra em um cenário de dados opacos. A maioria desses dados não está publicada, está invisível, por conseguinte, não pode ser reutilizada e enriquecer novas pesquisas. Essa situação permite inferir que, de forma predominante, os dados da área, no âmbito das teses, estão restritos aos seus pesquisadores e compreendem bens que não estão acessíveis ao público.

Sant’Ana (2018) alerta que a falta de gestão adequada afeta o compartilhamento dos dados, um assunto que tem sido debatido no cenário mundial, em virtude do avanço tecnológico, das necessidades das pesquisas, das exigências dos fomentadores, editores e governos no tocante à abertura, transparência e investimento eficiente em pesquisa (VAN DEN

EYNDEN; CORTI, 2017). Assim, é preciso que a CI, especificamente no cenário brasileiro, procure desenvolver entre os seus pesquisadores uma cultura voltada à preservação dos dados científicos, cujo debate pode ser realizado na disciplina de metodologia científica, já que os dados fazem parte deste contexto, com a finalidade de possibilitar dados acessíveis e contextualizados que permitam o surgimento de novos estudos e a otimização dos recursos que são investidos nas pesquisas da área.

6.4.8 Etapa “Descobrir”

O momento da descoberta dos dados ocorre quando eles são encontrados por outros usuários, cujo processo depende da publicação (STRASSER *et al.* 2012). Assim, buscou-se identificar nos trabalhos analisados se os pesquisadores da CI publicaram os seus dados em repositórios, em publicações ampliadas ou em periódicos de dados. A pesquisa demonstrou que, em apenas 7 (sete) casos, os dados têm a possibilidade de ser descobertos através de apêndice ou anexo dos trabalhos, em páginas da internet disponíveis para uso público e em meios de publicações acadêmicas. No Quadro 34, alguns exemplos dos poucos que foram identificados nos textos dos procedimentos metodológicos podem ser visualizados.

Quadro 34 - Etapa descobrir nas teses brasileiras da Ciência da Informação

Categoria	Características	Exemplos de segmentos de textos
Descobrir	Apêndice do trabalho	Foram pensados ainda para facilitar o entendimento do trabalho, bem como de seu percurso de elaboração, um glossário e a disponibilização dos dados da pesquisa para consulta, os quais se encontram nos apêndices. (Amostra 564, p.89)
	Página da <i>internet</i>	A proposta é disponibilizar os dados coletados e resultados alcançados na pesquisa, como por exemplo os links que reúnam informações de órgãos dos governos federal e estaduais (serviços, publicações, legislação), em uma página no Portal LTI. (Amostra 294, p. 34)
	Publicações acadêmicas	Quanto à forma de comunicar os resultados obtidos, duas abordagens são possíveis. [...]. Além da disponibilização da própria tese para acessos pós-defesa, os esforços da presente tese estão sendo comunicados por meio de publicações em periódicos

		acadêmicos, bem como apresentados em congressos, seminários, conferências, encontros, consórcios de doutorado, entre outros. Visa alcançar ainda teses e dissertações que enfoquem o método DEMO, as ontologias, os modelos organizacionais e os documentos. (Amostra 229, p. 219)
--	--	--

Fonte: Dados da pesquisa (2021)

Os resultados obtidos mostram que a maioria dos dados da CI não está disponível para ser acessada por seus potenciais usuários. Possivelmente, a quantidade ínfima que pode ser descoberta não possui uma documentação apropriada, baseada em um padrão de metadados específico. Todo esse cenário prejudica a visibilidade dos dados, quando deveriam ser descobertos por meio de livros, publicações em periódicos, páginas da *Web*, dentre outros serviços *online* (BALL, 2012).

Van den Eynden e Corti (2017) indicam os repositórios de dados específicos como uma forma de tornar os dados públicos e enfatizam o repositório voltado para as Ciências Sociais, *Reshare*. Os autores apontam alguns benefícios na escolha desse tipo de publicação, como o alinhamento entre a descrição dos dados com os padrões estabelecidos, o fato dos dados estarem armazenados em um local seguro, a descoberta de recursos online, o suporte ao usuário, dentre outros.

Para divulgar e compartilhar os dados da pesquisa, Mucheroni, Silva e Paletta (2015) apontam como caminho as publicações ampliadas. Os periódicos de dados também são mais um canal para comunicar a ciência. Segundo Curty e Aventurier (2017), esta modalidade de publicação descreve, de forma detalhada, o contexto dos dados, incluindo os procedimentos metodológicos. No Brasil, a Revista Encontros Bibli passou a publicar artigos de dados.

Todo esse cenário é favorável para os pesquisadores da CI disponibilizarem os seus dados de forma que outros pesquisadores possam descobri-los, o que poderá contribuir com a credibilidade dos estudos, já que proporciona a transparência dos resultados obtidos, e a otimização dos recursos investidos, ao evitar o retrabalho na obtenção dos mesmos dados e com reconhecimento profissional pela possibilidade da citabilidade. Por isso a necessidade de debater a temática da gestão dos dados na respectiva área.

6.4.9 Etapa “Reutilizar”

Para o levantamento dos dados referentes à etapa “reutilizar”, os trabalhos que integraram dados de outros estudos foram identificados, ou seja, que fizeram o reuso de dados, cujas sentenças encontradas em 7 (sete) metodologias, receberam a codificação integrar. Alguns exemplos estão expressos no Quadro 35.

Quadro 35 - Etapa reutilizar nas teses brasileiras da Ciência da Informação

Categoria	Características	Exemplos de segmentos de textos
Reutilizar	Integração de dados	Os registros primários das coletas dos pesquisadores se encontram anotados em cadernos de campo, no Laboratório do Nuvelhas, no prédio Administrativo III do campus da UFMG. As anotações realizadas no período de março/2003 a fevereiro/2009 foram, também, registradas em uma planilha Excel. Esta planilha não era de domínio publico, mas foi disponibilizada para utilização neste trabalho. (Amostra_41, p. 172)
	Aprofundar o assunto	Diversos aspectos do uso e apropriação das TICs pelos usuários do Programa GESAC foram analisados nos estudos de Medeiros (2012), os quais, no entanto, não focalizam diferenças de gênero no uso do programa. Com a intenção de explorar este aspecto, este estudo utiliza a mesma base de dados com esta finalidade. Trata-se, portanto, de um estudo com dados secundários. (Amostra 140, p.4)
	Comparação	Na quarta etapa, almejou-se comparar as indicações para formação dos profissionais de dados de pesquisa, apresentadas pela literatura da CI através de estudos anteriores, com as indicações reconhecidas nos cursos de dados pertencentes às iSchools. (Amostra_238, p. 121)

Fonte: Dados da pesquisa (2021)

Conforme pode ser visualizado nos exemplos extraídos dos textos, dados de outros estudos foram integrados para abordar ou aprofundar questões que não foram discutidas em outros trabalhos ou realizar comparações com pesquisas anteriores para aperfeiçoar os respectivos estudos.

Levando em consideração que a obtenção de dados pode ser onerosa, Borgman (2015) revela que a reutilização de dados pode economizar recursos como dinheiro, tempo e experiência para expandir a diversidade de recursos disponíveis voltados à pesquisa.

6.4.10 Etapa “Analisar”

O presente estudo permitiu identificar que os pesquisadores da CI utilizam as análises qualitativas e quantitativas dos dados e alguns casos apresentam as duas abordagens. A análise dos dados é realizada por meio de algumas técnicas e métodos, como a análise de conteúdo, análise do domínio, análise de redes sociais, análise teórica, análise documental, análise estatística e descrita, método metateórico, método de interpretação de sentidos, método bibliográfico, dentre outros. A categorização, a codificação e a utilização de *software* auxiliaram neste momento da pesquisa. Para Oliveira e Silva (2016), a análise de dados pode ocorrer por meio de *software*, podendo incluir ferramentas analíticas e métricas especializadas.

Houve a comparação de dados em 4 (quatro) casos com outros estudos realizados. Segundo Gil (2019), o confronto de dados oriundos da própria pesquisa com dados de outros estudos está relacionado com a análise dos dados. Porém, para que haja este tipo de ação, é preciso que os dados dos pesquisadores estejam disponíveis para serem acessados e possuam uma documentação que os contextualize para servir de parâmetro para outros estudos. No Quadro 36 estão disponibilizados alguns trechos das metodologias correspondentes a essa etapa.

Quadro 36 - Etapa analisar nas teses brasileiras da Ciência da Informação

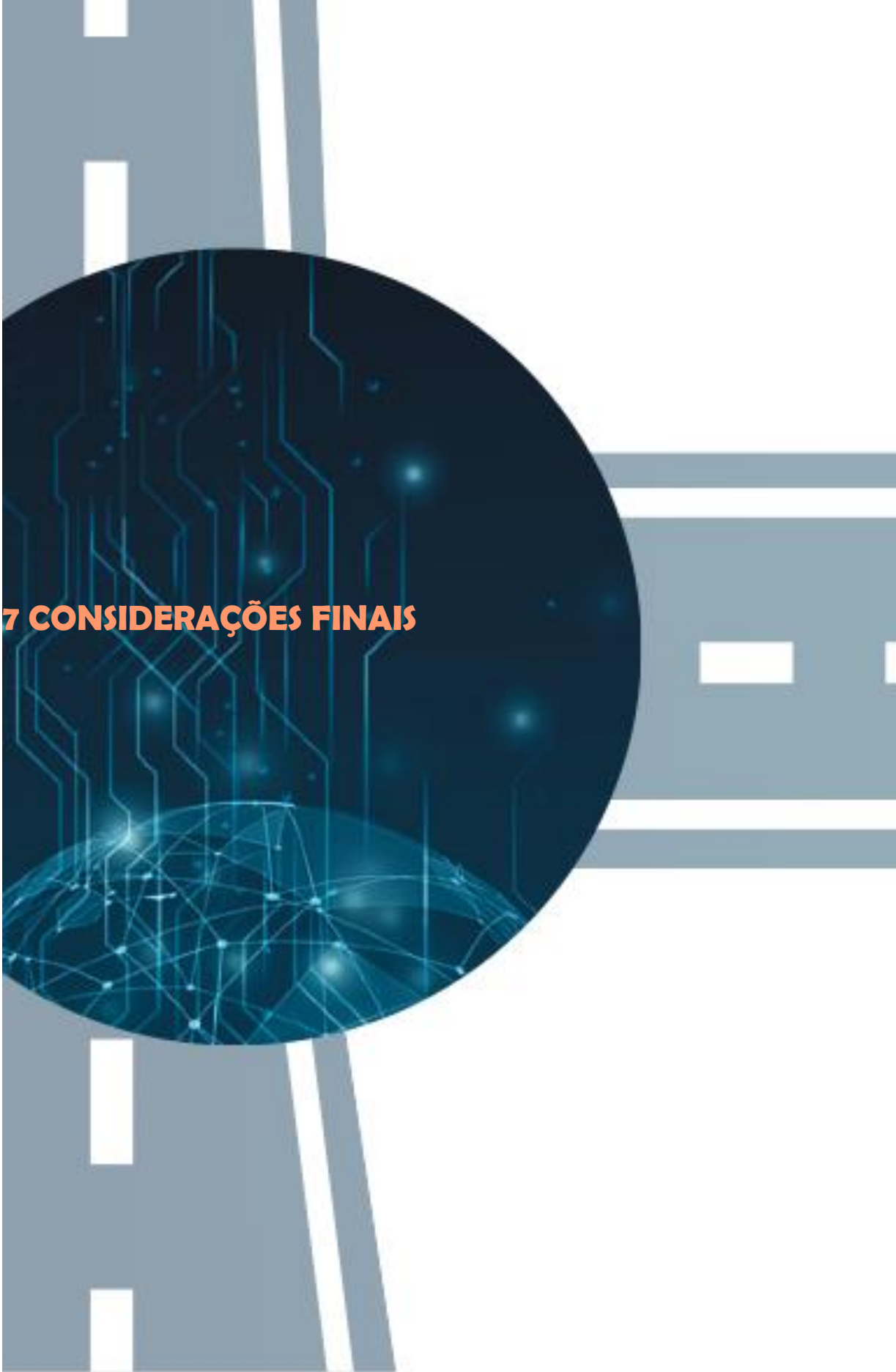
Categoria	Características	Exemplos de segmentos de textos
Analisar	Análise de conteúdo	Com o procedimento de coleta de dados concluído, a análise foi dividida em três etapas: na primeira, optou-se pela análise de conteúdo e uma análise descritiva; na segunda, utilizou-se também a análise de conteúdo para a caracterização e uma análise das variáveis que foram delimitadas conforme os objetivos propostos a, b e c que estão descritos no Quadro 3; e na terceira etapa, para cada fase do método de cenários, foram utilizadas as ferramentas pontuadas por Godet (2000), sendo o MICMAC para análise estrutural, o MACTOR para análise das

		estratégias de atores e o MORPHOL para análise morfológica. (Amostra_295, p. 101)
	Análise qualitativa e quantitativa	Ainda que a pesquisa seja qualitativa, visando auxiliar a análise, procedeu-se à análise estatística descritiva e também analítica de parte dos dados, levando em conta a frequência das respostas dos coordenadores ao questionário. (Amostra 305, p. 73)
	Comparação de dados	Para efeito de análises comparativas e norteadoras da interpretação do resultado de cruzamento das variáveis, serão considerados os resultados obtidos das pesquisas realizadas recentemente por Medeiros Neto (2012), Olinto (2007), Brasil (2008, 2005), buscando, dessa forma, traçar o perfil de uso da internet. (Amostra_140, p.78)

Fonte: Dados da pesquisa (2021)

Os achados dessa pesquisa permitem a compreensão da existência de coocorrência entre as codificações escolha de métodos e técnicas e recursos da categoria planejar e uma correlação com a etapa de organização dos dados. A partir dos resultados obtidos, depreende-se que existe uma integração entre as fases do CVD, de modo que o pesquisador precisa ter em mente que, para o bom funcionamento da gestão de dados, é preciso levar em consideração a integração entre as partes, ou seja, as decisões e ações tomadas em um momento podem afetar o ciclo como um todo.

Diante das análises e discussões apresentadas, verificou-se que os pesquisadores da CI, no âmbito das teses, mesmo não cumprindo o ciclo de vida dos dados, existem nos procedimentos metodológicos estudados a presença de ações que envolvem os elementos do CVD referência, logo, incluem ações/atividades dos CVDs do DataONE, DCC e DDI. Porém, essas ações são adotadas pelos pesquisadores de forma instintiva, pois não foram identificadas com clareza a gestão e curadoria dos dados. A própria ausência do plano de gestão de dados expõe a realidade, o que pode remeter à falta de abordagem sobre a temática por parte da literatura que trata das metodologias das Ciências Sociais. O foco dessas discussões geralmente está limitado ao planejamento da obtenção dos dados, à coleta e à análise dos dados. Assim sendo, é preciso que esses investigadores tenham uma mudança cultural na forma de lidar com os seus dados ao incluírem, nos procedimentos metodológicos dos seus trabalhos de pesquisa, as etapas do CVD de forma detalhada.



7 CONSIDERAÇÕES FINAIS

Por meio da pesquisa realizada, foi possível atingir os objetivos específicos e o objetivo geral do estudo. No primeiro momento, verificou-se que alguns programas de pós-graduação da CI apresentam consideráveis desafios no tocante ao acesso aos trabalhos de tese, além de alguns problemas na busca pela seção da metodologia nas respectivas pesquisas. Os problemas de recuperação da informação levantados indicam que a gestão dos dados científicos também é uma questão que merece uma maior atenção por parte da área, o que foi constatado no decorrer da pesquisa. Por isso, há a necessidade de documentar melhor os próprios trabalhos de tese, de forma que eles possam estar disponíveis. Além disso, é preciso existir uma seção específica de metodologia ou que ela seja bem identificada em uma subseção com indicativo numérico e com uma nomenclatura convencional. Apesar dos desafios mencionados, o primeiro objetivo específico foi alcançado ao serem levantadas as teses brasileiras da CI que estavam disponíveis.

Ao realizar a interseção entre os elementos dos ciclos de vida dos dados do DataONE, DCC e DDI, foi identificado que a primeira iniciativa foi desenvolvida com o foco nas ciências da terra; a segunda compreende um modelo de curadoria digital que pode ser aplicado não apenas para a gestão dos dados científicos; e a terceira é voltada para a área das ciências sociais. Apesar das diferenças entre as iniciativas, foi verificada a existência de correlação entre os seus elementos, o que possibilitou o alcance do segundo objetivo específico.

A partir de então, foi proposto um modelo de referência com características das três iniciativas. A finalidade foi promover um modelo mais prático, que pode ser aplicável pelos pesquisadores da CI e da grande área das Ciências Sociais Aplicadas, no desenvolvimento das suas pesquisas científicas. Assim sendo, foi alcançado o terceiro objetivo específico.

Após a análise das metodologias da CI, foi possível classificar a tipologia dos dados encontrados no tocante à natureza. Desta forma, foram classificados, predominantemente, como dados textuais, seguidos dos dados de visualizações (quadros, figuras, tabelas, gráficos, mapas e modelos). Também foram identificados dados de natureza multimídia (vídeos, áudios e fotografias), dados numéricos (medidas, resultados de levantamentos, resultados de experimentos, fórmulas, algoritmos e resultados de amostras) e uma pequena representação de dados como *software*. Assim, o estudo apontou uma diversidade significativa de dados na área, o que sugere medidas adequadas de curadoria.

Os dados das teses estudadas foram classificados como sendo de origem predominantemente de registro e observacional e em uma menor proporção como sendo computacional e experimental. Devido à presença significativa dos dados da categoria observacional, que são os mais difíceis de reproduzir, e a ausência de ações voltadas à preservação, é possível que a reutilização dos dados da área esteja prejudicada. Por isso,

recomenda-se a adoção de estratégias de acordo com as categorias que originaram os dados. A partir da classificação dos tipos de dados encontrados, alcançou-se o quarto objetivo específico.

Através da intersecção entre os elementos do ciclo de vida dos dados proposto com os percursos metodológicos das teses brasileiras da CI, a etapa planejar foi a que mais se destacou ao apresentar relação com as demais etapas do ciclo. Porém, é preciso atentar para a elaboração de um plano de gestão dos dados. Na etapa coletar, foi verificado que a CI utiliza uma combinação de fontes para obter os seus dados, se configurando como dados digitais e dados físicos. A etapa organizar se fez presente nos trabalhos analisados, pois, geralmente, os pesquisadores, após o processo de coleta, organizaram os seus dados para facilitar a compreensão no momento da análise.

Os pesquisadores da CI não utilizaram um padrão de metadados para descrever os seus dados. Essa categoria, apesar de ter apresentado um percentual considerável, não significa dizer que os pesquisadores utilizaram uma documentação robusta sobre os seus dados, uma vez que o percentual remete, predominantemente, aos instrumentos que foram utilizados no processo de coleta. Assim, orienta-se a adoção de um padrão de metadados que possibilite a descrição dos dados.

Identificou-se algumas ações voltadas a assegurar a qualidade dos dados, quando os pesquisadores os revisaram, eliminaram os dados duplicados, realizaram a limpeza dos dados, dentre outros. Um fato que chamou atenção foi a ligação com a etapa coletar, ao assegurar ajustes nos instrumentos de coleta após a realização de pré-testes, o que certamente pode contribuir com a qualidade dos dados obtidos.

A preservação não foi evidenciada na investigação, de modo a permitir que os dados possam ser reutilizados no futuro, assim como houve poucos indícios que os pesquisadores, por meio de publicações, favorecem a descoberta dos seus dados científicos. Esse cenário mostra que os dados da área não estão acessíveis e visíveis.

No que se refere ao reúso de dados de outros estudos, foi constatada uma presença tímida, apenas alguns fizeram a integração de um conjunto de dados já existente e, por fim, houve uma presença considerável da etapa analisar, que envolveu métodos e técnicas que foram previamente planejados.

O objetivo geral do trabalho foi atingido e a hipótese de que os percursos metodológicos apresentam indícios dos elementos correspondentes aos ciclos de vida dos dados foi constatada, porém, conforme verificado na literatura que trata das metodologias nas Ciências Sociais, algumas etapas possuem uma presença notável e outras uma participação pouco significativa.

Diante da realidade apontada, sugere-se que a literatura que trata sobre a metodologia nas Ciências Sociais, além de discutir sobre as fases de planejar, coletar e analisar aborde, de forma mais clara, etapas como organizar, descrever, assegurar a qualidade, preservar, descobrir e reutilizar, de modo que os pesquisadores possam explicitar cada vez mais nos seus percursos metodológicos os elementos do CVD, pois se configura como o local ideal dos trabalhos científicos para especificar os processos dos dados inerentes ao trabalho de investigação.

No contexto científico, é necessário garantir a veracidade dos dados, além de assegurar que eles sejam valorosos à medida que proporcionam novas descobertas, por isso recomenda-se que novas propostas de pesquisas contemplem a adoção de medidas por parte dos programas de pós-graduação em CI, no Brasil, e que incentivem os pesquisadores a adotar as práticas de gestão de dados por meio do CVD nos procedimentos metodológicos.

Considerando que essa pesquisa está relacionada com a Ciência da Informação, campo este pertencente à grande área das Ciências Sociais Aplicadas, sugere-se que esse estudo de diagnóstico da situação referente à gestão dos dados nos percursos metodológicos dos trabalhos científicos possa ser estendido para outras áreas do conhecimento no âmbito das Ciências Exatas e da Terra, Ciência Biológicas, Engenharias, Ciências da Saúde, Ciências Agrárias, Ciências Humanas, Linguística, Letras e Artes.

No que se refere às limitações da pesquisa, este trabalho abordou a identificação da presença dos elementos do CVD nos percursos metodológicos das teses de forma geral, no período de 2012 a 2020. Entretanto, não foi realizada uma análise por ano. Além disso, não foi verificado se as práticas dos pesquisadores em relação à gestão dos dados científicos, no âmbito dos percursos metodológicos das teses, estão alinhadas com os títulos dos respectivos trabalhos que abordam à temática em questão, o que pode ser explorado em futuras pesquisas.

REFERÊNCIAS

- ABBOT, D. **What is digital curation?**. 2008. Disponível em: <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation>. Acesso em: 20 mar. 2020.
- AGÊNCIA USP DE GESTÃO DA INFORMAÇÃO ACADÊMICA UNIVERSIDADE DE SÃO PAULO (AGUIA). **Dados de pesquisa**. 2020. Disponível em: <https://www.aguia.usp.br/apoio-pesquisador/dados-pesquisa/>. Acesso em: 16 jul. 2020.
- ANJOS, R. L. **Análise dos componentes do ciclo de vida dos dados na matriz curricular dos cursos de biblioteconomia: um estudo comparado**. 2019. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal da Paraíba, João Pessoa, 2019. Disponível em: <https://repositorio.ufpb.br/jspui/handle/123456789/19912>. Acesso em: 6 ago. 2021.
- ANJOS, R. L.; DIAS, G. A. Atuação dos profissionais da informação no ciclo de vida dos dados – DATAONE: um estudo comparado. **Informação & Informação**, [s.l.], v. 24, n. 1, p. 80-101, mar. 2019. ISSN 1981-8920. Disponível em: <http://dx.doi.org/10.5433/1981-8920.2019v24n1p80>. Acesso em: 6 abr. 2020.
- ARAÚJO, C. A. A. **O que é ciência da informação**. Belo Horizonte: KMA, 2018.
- ARAÚJO, D. G.; DIAS, G.A; ARAÚJO, W. J. Práticas de gestão de dados: uma revisão da literatura sobre o termo data life cycle. **Métodos e Pesquisa em Administração**, v. 4, n.1, p. 35-47, 2019. Disponível em: <https://doi.org/10.22478/ufpb.2525-3867.2019v4n1.44449>. Acesso em: 18 mar. 2020.
- ARAÚJO, D. G.; LLARENA, M. A. A.; SIEBRA, S. A.; DIAS, G.A. Contribuições para a gestão de dados científicos: análise comparativa entre modelos de ciclo de vida dos dados. **Liinc em Revista**, v. 15, n.2, p. 32-51, 2019. Disponível em: <http://revista.ibict.br/liinc/article/view/4686>. Acesso em: 18 mar. 2020.
- ASSANTE, M.; CANDELA, L.; CASTELLI, D.; TANI, A. Are scientific data repositories coping with research data publishing?. **Data Science Journal**, v.15, p.6, 2016. Disponível em: <http://doi.org/10.5334/dsj-2016-006>. Acesso em: 19 jun. 2020.
- ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14724: Informação e documentação: Trabalhos acadêmicos**. Rio de Janeiro, p.8. 2011.
- ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 6024: Informação e documentação: Numeração progressiva das seções de um documento**. Rio de Janeiro, p.1. 2012.
- AUER, S. Introduction to lod2. In: AUER, S.; BRYL, V.; TRAMP, C. (ed.). **Linked open data – creating knowledge out of interlinked data**. Lecture Notes in Computer Science. Springer, 2014. p. 1-17.
- BALL, A. **Review of Data Management Lifecycle Models (version 1.0)**. REDm-MED

Project Document redm1rep120110ab10. Bath: University of Bath, 2012. Disponível em: <https://purehost.bath.ac.uk/ws/portalfiles/portal/206543/redm1rep120110ab10.pdf>. Acesso em: 12 set. 2020.

BARDIN, L. **Análise de conteúdo**. Tradução: Luis Antero Reto, Augusto Pinheiro. Lisboa: Edições 70, LDA, 2011.

BARRETO, A. A. A condição da informação. **São Paulo em Perspectiva**, São Paulo, v.16, n. 3, p. 67-74, jul. 2002. Disponível em: <http://dx.doi.org/10.1590/S0102-88392002000300010>. Acesso em: 12 out. 2017.

BEZJAK, S.; CLYBURN-SHERIN, A.; CONZETT, P.; FERNANDES, P.; GÖRÖGH, E.; HELBIG, K.; KRAMER, B.; LABASTIDA, I.; NIEMEYER, K.; PSOMOPOULOS, F.; ROSS-HELLAUER, T.; SCHNEIDER, R.; TENNANT, J.; VERBAKEL, E.; BRINKEN, H.; HELLER, L. **Open Science Training Handbook**. Hannover: German National Library of Science and Technology, 2018, 204 p. Disponível em: <https://www.fosteropenscience.eu/content/open-science-training-handbook/>. Acesso em: 10 dez. 2019.

BORGMAN, C. L. **Big data, little data, no data: scholarship in the networked world**. London: The MIT Press, 2015.

CANDELA, L.; CASTELLI, D.; MANGHI, P.; TANI, A. Data journals: a survey. **Journal of the Association for Information Science and Technology**, v.66, n.9, p.1747-1762, 2015. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23358>. Acesso em: 15 mar. 2021.

COSTA, M. M. **Diretrizes para uma política de gestão de dados científicos no Brasil**. 2017. 288 f. Tese (Doutorado em Ciência da Informação) - Universidade de Brasília, Brasília, 2017.

CRUZ NETO, O. O trabalho de campo como descoberta e criação. In: MINAYO, M.C.S (org.). **Pesquisa Social: teoria, método e criatividade**. 23. ed. Petrópolis: Vozes, 2004.

CURTY, R. Abordagens de reuso e a questão da reusabilidade dos dados científicos. **Liinc Em Revista**, v. 15, n.2, 2019. Disponível em: <https://doi.org/10.18617/liinc.v15i2.4777>. Acesso em: 15 mai. 2021.

CURTY, R. G; AVENTURIER, P. O paradigma da publicação de dados e suas diferentes abordagens. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017. **Anais [...]**, Marília: UNESP, 2017. Disponível em: <https://www.brapci.inf.br/index.php/res/v/105144>. Acesso em: 8 jan. 2020.

DATAONE. **Primer on data management**: what you always wanted know. Disponível em: <https://www.DataONE.org/sites/all/documents/DataONEBPPPrimer020212.pdf>. Acesso em: 13 ago. 2019.

DATAONE. **What is DataONE?** online. 2020. Disponível em: <https://www.DataONE.org/what-DataONE>. Acesso em: 13 ago. 2019.

DATAONE. **Data Life Cycle**. 2020. Disponível em: <https://old.dataone.org/data-life-cycle>. Acesso em: 20 jul. 2020.

DATA SCIENCE JOURNAL. 2020. Disponível em: <https://datascience.codata.org/>. Acesso em: 12 jan.2020.

DDI. Data Documentation Initiative: structural reform group. **DDI version 3.0 conceptual model**. DDI Alliance. 2004. Disponível em: <https://ddialliance.org/sites/default/files/Concept-Model-WD.pdf>. Acesso em: 6 set. 2020.

DESLANDES, S. F. O projeto de pesquisa como exercício científico e artesanato intelectual. *In*: MINAYO, M. C. S. (org.). **Pesquisa Social**: teoria, método e criatividade. Petrópolis: Vozes, 2009. p. 31-60.

DIAS, G. A.; ANJOS, R. L.; ARAUJO, D. G. A gestão dos dados de pesquisa no âmbito da comunidade dos pesquisadores vinculados aos programas de pós-graduação brasileiros na área da ciência da informação: desvendando as práticas e percepções associadas ao uso e reúso de dados. **Liinc em Revista**, Rio de Janeiro, v.15, n.2, p. 5-31, nov. 2019. Disponível em: <http://revista.ibict.br/liinc/article/view/4683>. Acesso em: 15 jan.2020.

DIAS. G.A.; ANJOS, R.L.; RODRIGUES, A.A. Os princípios fair: viabilizando o reuso dos dados científicos. *In*: DIAS, G.A.; OLIVEIRA, B.M.J.F. (org.). **Dados Científicos**: perspectivas e desafios. João Pessoa: UFPB, 2019. p. 177-187.

DIAS, G. A; OLIVEIRA, B. M.J.F. (org.). **Dados científicos**: perspectivas e desafios. João Pessoa: UFPB, 2019. *E-book*. Disponível em: <http://www.editora.ufpb.br/sistema/press5/index.php/UFPB/catalog/book/359>. Acesso em: 10 jul. 2021.

DIGITAL CURATION CENTRE. **Curation lifecycle model**. 2020. Disponível em: <http://www.dcc.ac.uk/resources/curation-lifecycle-model>. Acesso em: 13 ago. 2019.

DIETRICH, P; LOISON, M; ROUPNEL, M. Articular as abordagens quantitativa e qualitativa. *In*: PAUGAM, S. **Pesquisa Sociológica**. Petrópolis: Vozes, 2015. p. 171-182.

DTL. **Jointly designing a Data FAIRPORT**. 2014. Disponível em: <https://www.dtls.nl/2014/01/20/jointly-designing-data-fairport/>. Acesso em: 25 de jun. 2020.

DUBLIN CORE™ METADATA INITIATIVE. **Innovation in metadata design, implementation & best**. Metadata Basics. 2021. Disponível em: <https://dublincore.org/resources/metadata-basics/>. Acesso em: 12 mai. 2021.

ENCONTROS BIBLI. **Sobre a Revista**. 2020. Disponível em: <https://periodicos.ufsc.br/index.php/eb/index>. Acesso em: 20 ago. 2020.

ESCOLA DE COMUNICAÇÃO E ARTES UNIVERSIDADE DE SÃO PAULO. **Teses PPGCI**. Disponível em: <http://www3.eca.usp.br/pos/ppgci/teses>. Acesso em: 22 maio 2020.

ESPÍNDOLA, P. L.; SALM JUNIOR, J. F.; ROSA, F.; JULIANI, J. P. Governança de dados aplicada à ciência da informação: análise de um sistema de dados científicos para a área da saúde. **Revista Digital de Biblioteconomia e Ciência da Informação**, v. 16, n. 3, p. 274-298, 16 ago. 2018. DOI: <https://doi.org/10.20396/rdbci.v16i3.8651080>. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8651080>. Acesso em: 11 mar. 2020.

FERGUSON, A. R.; NIELSON, J. L.; CRAGIN, M.H.; BANDROWSKI, A. E.; MARTONE, M. E. Big data from small data: data-sharing in the ‘long-tail’ of neuroscience. **Nature Neuroscience**, v. 17, n. 11, nov. 2014. Disponível em: <https://www.nature.com/articles/nn.3838>. Acesso em: 18 jun. 2021.

FORCE11. **FAIR data publishing group**. 2020. Disponível em: <https://www.force11.org/group/fairgroup>. Acesso em 25 de jun. 2020.

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO (FAPESP). **Plano de Gestão de Dados**. 2020. Disponível em: <http://www.fapesp.br/gestaodedados/>. Acesso em: 17 jan. 2020.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 7. ed. São Paulo: Atlas, 2019.

GOBEN, A.; RASZEWSKI, R. The data life cycle applied to our own data. **Journal of the Medical Library Association**, v. 103, n. 1, p. 40, 2015. Disponível em: <http://dx.doi-org.ez15.periodicos.capes.gov.br/10.3163/1536-5050.103.1.008>. Acesso em: 27 de março de 2020.

GODIM, L. M. P; LIMA, J. C. Considerações sobre o projeto de pesquisa. In: GODIM, L. M. P; LIMA, J. C. **A pesquisa como artesanato Intelectual**: considerações sobre método e bom senso. São Carlos: EdUFSCar, 2006. p. 41- 78.

GOFAIR. **FAIR principles**. 2020. Disponível em: <https://www.go-fair.org/fair-principles/>. Acesso em: 26 jun. 2020.

GOFAIR. **GO FAIR initiative**. 2020. Disponível em: <https://www.go-fair.org/go-fair-initiative/>. Acesso em: 26 jun. 2020.

GRAY, J. Jim Gray on eScience: a transformed scientific method. In: HEY, T.; TANSLEY, S.; TOLLE, K. (ed.). **The fourth paradigm**: data-intensive scientific discovery. Washington: Microsoft Research, 2009. Disponível em: <http://languagelog.ldc.upenn.edu/myl/JimGrayOnE-Science.pdf>. Acesso em: 13 jun. 2020.

GUPTA, S.; MÜLLER-BIRN, C. A study of e-Research and its relation with research data life cycle: a literature perspective. **Benchmarking**: an international journal, v. 25, n. 6, p.1656-1680, 2018. Disponível em: <https://www-emerald.ez15.periodicos.capes.gov.br/insight/content/doi/10.1108/BIJ-02-2017-0030/full/html>. Acesso em: 15 out. 2020.

HEDSTROM, M.; MYERS, J. SEAD: finding the long tail lost in big data. 2014. Disponível em: <http://sead-data.net/sites/default/files/pubs/NDS-White-Paper-The-Long-Tail-Lost-in-Big-Data-2.pdf>. Acesso em: 18 jun. 2021.

IONESCU, S. "Introduction to DDI 3.0." (PPT 2.1MB). *In: PRESENTED AT THE CESSDA EXPERT SEMINAR, Uk Data Archive*, sep. 2007. Disponível em: <https://ddialliance.org/events/workshops>. Acesso em: 10 set. 2020.

JANKOWSKI, N. W. Exploring e-science: an introduction. **Journal of ComputerMediated Communication**, v. 12, p. 549-562, jan. 2007. Disponível: <https://doi.org/10.1111/j.1083-6101.2007.00337.x>. Acesso em: 20 jan. 2021.

KANJALA, C., TODD, J., BECKLES, D., CASTILLO, T., KNIGHT, G., MTENGA, B., URASSA, M., ZABA, B. Open-access for existing LMIC demographic surveillance data using DDI. **IASSIST Quarterly**, v.40, n.2, 2017. Disponível em: <https://doi.org/10.29173/iq783>. Acesso em: 15 out. 2020.

KAUARK, F.; MANHÃES, F.C.; MEDEIROS, C.H. **Metodologia da pesquisa**: guia prático. Itabuna: Vila Litterarum, 2010.

KRATZ, J.; STRASSER, C. Data publication consensus and controversies. **F1000Research**, v.3, n. 94, 2014. Disponível em: <https://f1000research.com/articles/3-94>. Acesso em: 11 jan. 2020.

LENHARDT, W.C.; AHALT, S.; BLANTON, B.; CHRISTOPHERSON, L.; IDASZAK, R. Data management lifecycle and software lifecycle management in the context of conducting science. **Journal of Open Research Software**, v.2, n.1, 2014. Disponível em: <http://doi.org/10.5334/jors.ax>. Acesso em: 19 jun. 2020.

LOHR, S. L. **Sampling**: design and analysis. Nelson Education, 2009.

MARCONI, M. A.; LAKATOS, E. M. **Fundamentos de metodologia científica**. 8. ed. São Paulo: Atlas, 2019.

MEADOWS, A. J. **A comunicação científica**. Tradução Antonio Agenor Briquet de Lemos. Brasília: Briquet de Lemos, 1999.

MEDEIROS, C.B. Gestão de dados científicos – da coleta à preservação. **SciELO em Perspectiva**, 2018. Disponível em: <https://blog.scielo.org/blog/2018/06/22/gestao-de-dados-cientificos-da-coleta-a-preservacao/>. Acesso em: 10 jun. 2020.

MEDEIROS, J. S.; CAREGNATO, S. E. Compartilhamento de dados e e-science: explorando um novo conceito para a comunicação científica. **Liinc em Revista**, v. 8, n. 2, 2012. Disponível em: <https://doi.org/10.18617/liinc.v8i2.488>. Acesso em: 6 out. 2021.

MICHEL, M. H. **Metodologia e pesquisa científica em ciências sociais**: um guia prático para acompanhamento da disciplina e elaboração de trabalhos monográficos. 3. ed. São Paulo: Atlas, 2015.

MINAYO, M. C. S. Trabalho de campo: contexto de observação, interação e descoberta. *In: MINAYO, M. C. S. (org.). Pesquisa social: teoria, método e criatividade*. 25. ed. Petrópolis: Vozes, 2007. p. 61-78.

MINAYO, M. C. S. O desafio da pesquisa social. In: MINAYO, M. C. S. (org.). **Pesquisa Social: teoria, método e criatividade**. Petrópolis: Vozes, 2009. p.9-29.

MONTEIRO, E. C. S. A. **Direitos autorais nos repositórios de dados científicos: análise sobre os planos de gerenciamento dos dados**. Orientador: Ricardo César Gonçalves Sant'Ana. 2017. 115 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de filosofia e Ciências, Universidade Estadual Paulista, Marília, 2017. Disponível em: <https://repositorio.unesp.br/handle/11449/149748>. Acesso em: 5 jan. 2020.

MONTEIRO, E. C. S. A.; SANT'ANA, R. C. G. Plano de gerenciamento de dados em repositórios de dados de universidades. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 23, n. 53, p. 160-173, 2018. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p160/37296/>. Acesso em: 12 jan. 2021.

MORAIS, I. S.; GONÇALVES, P. F.; LEDUR, C. L.; CÓRDOVA JUNIOR, R. S.; SARAIVA, M. O.; FRIGERI, S. R. **Introdução a big data e internet das coisas (IoT)**. Porto Alegre: SAGAH, 2018. *E-book*. Disponível em: <https://integrada.minhabiblioteca.com.br/books/9788595027640>. Acesso em: 30 jun. 2021.

MOZZATO, A.R.; GRZYBOVSKI, D. Análise de conteúdo como técnica de dados qualitativos no campo da administração: potencial e desafios. **RAC**, Curitiba, v.15, n.4, p. 731-747, jul./ago.2011. Disponível em: <https://www.scielo.br/pdf/rac/v15n4/a10v15n4.pdf>. Acesso em: 4 maio 2020.

MUCHERONI, M. L.; SILVA, J. F. M.; PALETTA, C. F. Entre a publicação ampliada e a multimodalidade. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais [...]**. João Pessoa: UFPB, 2015.

NATIONAL SCIENCE BOARD (NSB). **Long-lived digital data collections: enabling research and education in the 21st century**. 2005. Disponível em: <https://www.nsf.gov/geo/geo-data-policies/nsb-0540-1.pdf>. Acesso em: 30 mar. 2020.

OLIVEIRA, A. C. S.; SILVA, E. M. Ciência aberta: dimensões para um novo fazer científico. **Informação & Informação**, [s.l.], v. 21, n. 2, p. 5-39, dez. 2016. ISSN 1981-8920. Disponível em: <http://dx.doi.org/10.5433/1981-8920.2016v21n2p5>. Acesso em: 20 abr. 2020.

ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **OECD principles and guidelines for access to research data from public funding**. Paris: OECD, 2007. Disponível em: <https://www.oecd.org/sti/inno/38500813.pdf>. Acesso em: 16 jul. 2020.

PIWOWAR, H.A.; VISION, T. J. Data reuse and the open data citation advantage. **PeerJ**, 2013. Disponível em: <https://doi.org/10.7717/peerj.175>. Acesso em: 28 jun. 2020.

PLATAFORMA SUCUPIRA. 2021. Disponível em: <https://sucupira.capes.gov.br/sucupira/public/index.xhtml>. Acesso em: 20 maio 2021.

PRÁ NETTO, A. S. D.; MORO, E. P.; FERREIRA, F. F. O Big Data. **Os 5 V's do Big Data**. 2020. Disponível em: https://www.gta.ufrj.br/grad/15_1/bigdata/vs.html. Acesso em: 15 jun. 2020.

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO IBICT –UFRJ. **Teses de doutorado**. Disponível em: <http://www.ppgci.ufrj.br/teses-de-doutorado/>. Acesso em: 26 maio 2020.

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO UFBA. **Trabalhos de conclusão**. 2020. Disponível em: <https://ppgci.ufba.br/trabalhos-de-conclusao>. Acesso em: 24 maio 2020.

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO UFMG. **Teses defendidas**. 2020. Disponível em: <http://ppgci.eci.ufmg.br/teses-defendidas/>. Acesso em: 28 maio 2020.

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO UFSC. 2020. Disponível em: <https://pgcin.ufsc.br/>. Acesso em: 27 maio 2020.

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO UNESP. **Teses** 2020. Disponível em: <https://www.marilia.unesp.br/#!/pos-graduacao/mestrado-e-doutorado/ciencia-da-informacao/teses/>. Acesso em: 23 maio 2020.

PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO & ORGANIZAÇÃO DO CONHECIMENTO. **Teses defendidas**. 2020. Disponível em: <http://ppggoc.eci.ufmg.br/defesas/teses-defendidas/>. Acesso em: 28 maio 2020.

RAUTENBERG, S.; HILD, T. A.; SOUZA, L. Curadoria digital de dados e web de dados: mantendo dados abertos conectados para estudos bibliométricos e cientométricos. **Em Questão**, Porto Alegre, v. 24, p. 29-47, 2018. Disponível em: <http://dx.doi.org/10.19132/1808-5245240.29-47>. Acesso em: 18 mar. 2020.

REPOSITÓRIO INSTITUCIONAL DA UFPB. 2020. Disponível em: https://repositorio.ufpb.br/jspui/?cnpq_page=2#facet_cnpq. Acesso em: 25 maio 2020.

REPOSITÓRIO INSTITUCIONAL DA UNB. **Navegando "FCI - Doutorado em Ciência da Informação (Teses)" por Data de defesa**. 2020. Disponível em: https://repositorio.unb.br/handle/10482/5364/browse?type=datesubmitted&sort_by=4&order=ASC&rpp=20&etal=-1&null=&offset=0. Acesso em: 5 jun 2020.

RICHARDSON, R. J. **Pesquisa social: métodos e técnicas**. 4. ed. São Paulo: Atlas, 2017.

RICE, R.; SOUTHALL, J. **The data librarian's handbook**. London: Publisher by Facet Publishing, 2016.

SALES, L. F. As 5 leis de Ranganathan aplicada à gestão de dados de pesquisa. Publicado pelo canal da Biblioteca de Ciências Humanas da Universidade Federal do Ceará. 2020. 1 vídeo (123min). Disponível em: <https://www.youtube.com/watch?v=wExu6UsLbOM>. Acesso em: 19 jun. 2020.

SALES, L. F.; SAYÃO, L. F. Uma proposta de taxonomia para dados de pesquisa. **Revista Conhecimento em Ação**, v. 4, n. 1, jan./jun. 2019. Disponível em: <https://revistas.ufrj.br/index.php/rca/article/view/26337>. Acesso em: 10 maio. 2021.

SAMPIERI, R. H; COLLADO, C. F; LUCIO, P. B. Tipos de pesquisa. *In*: SAMPIERI, R. H; COLLADO, C. F; LUCIO, P. B. **Metodologia da pesquisa**. 3. ed. São Paulo: McGraw-Hill, 2006. p. 96-115.

SANCHEZ, F. A.; VIDOTTI, S. A. B. G.; VECHIATO, F. L. A contribuição da curadoria digital em repositórios digitais. **Revista Informação na Sociedade Contemporânea**, v.1, p. 1-17, 11 jun. 2017. Disponível em: <https://periodicos.ufrn.br/informacao/article/view/12280>. Acesso em: 15 mar. 2020.

SANT'ANA, R. C. G. Ciclo de vida dos dados e o papel da ciência da informação. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Florianópolis. **Anais [...]**. Florianópolis: UFSC, 2013. Disponível em: <http://www.brapci.inf.br/index.php/article/view/0000013772/fac7d2fec2c83f1e9208204b2fed10>. Acesso em: 20 jun. 2018.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, Londrina, v. 21, n. 2, p. 116–142, maio/ago. 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940/20124>. Acesso em: 20 jun. 2018.

SANTOS, D. B.; ROCKEMBACH, M. Publicações ampliadas: aspectos da integração de dados de pesquisa. **Informação & Sociedade: estudos**, v.28, n. 2, p.257-269, 2018. Disponível em: <https://periodicos.ufpb.br/index.php/ies/article/view/38558/20806>. Acesso em: 10 jan. 2020.

SAYÃO, L. F. Gestão de dados de pesquisa: o que todo pesquisador espera da sua biblioteca. Publicado pelo canal da Biblioteca de Ciências Humanas da Universidade Federal do Ceará. 2020. 1 vídeo (124 min). Disponível em: https://www.youtube.com/watch?v=_HT_nrmwjUE. Acesso em: 14 jul. 2020.

SAYÃO, L.F.; SALES, L.F. Curadoria digital: um novo patamar para a preservação de dados digitais de pesquisa. **Informação & Sociedade: estudos**, João Pessoa, v.22, n.3, p. 179-191, set./dez. 2012. Disponível em: <https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/curadoria-digital---sayao.pdf>. Acesso em: 20 mar. 2020.

SAYÃO, L.F.; SALES, L.F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: Comissão Nacional de Energia Nuclear, 2015. Disponível em: http://carpedien.ien.gov.br/bitstream/ien/1624/1/GUIA_DE_DADOS_DE_PESQUISA.pdf. Acesso em: 10 out. 2017.

SAYÃO, L. F; SALES, L. F. Curadoria digital e dados de pesquisa. **AtoZ: novas práticas em informação e conhecimento**, [s. l.], v. 5, n. 2, p. 67-71, dez. 2016. Disponível em: <https://revistas.ufpr.br/atoz/article/view/49708/30179>. Acesso em: 5 jan. 2020.

SAYÃO L.F.; SALES, L. F. Subsídios para a construção de um modelo de avaliação de sistemas de gestão de dados de pesquisa. **PontodeAcesso**, Salvador, v.12, n.3, p.80-108, dez. 2018. Disponível em: <http://dx.doi.org/10.9771/rpa.v12i3.28965>. Acesso em: 16 jun. 2020.

SAYÃO, L.F.; SALES, L.F. A ciência invisível: os dados da cauda longa da pesquisa científica. *In*: DIAS, G. A; OLIVEIRA, B. M.J.F. (org.). **Dados científicos: perspectivas e desafios**. João Pessoa: UFPB, 2019. p. 33-52.

SAYÃO, L. F.; SALES, L. F. Afinal, o que é dado de pesquisa?. **Biblos**, v.34, n.2, 2020. Disponível em: <https://www.seer.furg.br/biblos/article/view/11875>. Acesso em: 20 abr. 2021.

SEGUNDO SILVA, S. J. ARAÚJO, W.J. Curadoria e ciclo de vida de dados. *In*: DIAS, G. A; OLIVEIRA, B. M.J.F. (org.). **Dados científicos: perspectivas e desafios**. João Pessoa: UFPB, 2019. p. 113-152.

SCHÖPFEL, J.; PROST, H.; MALLERET, C.; JUZNIC, P.; CESAREK, A.; KOLER-POVH, T. Dissertations and data. **Grey Journal (TGJ)**, [s. l.], v. 12, n. 3, p. 126–148, 2016. Disponível em: <http://search-ebscohost-com.ez15.periodicos.capes.gov.br/login.aspx?direct=true&db=lih&AN=118841427&lang=pt-br&site=ehost-live>. Acesso em: 16 mar. 2020.

SIEBRA, S. A.; BORBA, V. R.; MIRANDA, M. J. K. F. O. Curadoria digital: um termo interdisciplinar. **Informação & Tecnologia**, v.3, n.2, p.21-38, jul./dez. 2016. Disponível em: <https://periodicos.ufpb.br/index.php/itec/article/view/38408/20163>. Acesso em: 5 jul. 2018.

SIEBRA, S. A.; BORBA, V. R.; GALINDO, M.; MIRANDA, M. K. F. O.; TAVARES, L. L. L.; OLIVEIRA, J. N. N. O. Curadoria digital: além da questão da preservação digital. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Florianópolis. **Anais [...]**. Florianópolis: UFSC, 2013. Disponível em: <http://www.egov.ufsc.br/portal/conteudo/curadoria-digital-al%C3%A9m-da-quest%C3%A3o-da-preserva%C3%A7%C3%A3o-digital-0>. Acesso em 20 jun. 2018.

SILVA, F. C. C. O papel dos bibliotecários na gestão de dados científicos. **Revista Digital de Biblioteconomia e Ciência da Informação**, v. 14, n. 3, p. 387-406, 30 set. 2016. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8646333>. Acesso em: 18 mar. 2020.

SILVA, F. C. C. **Gestão de dados científicos**. 1. ed. Rio de Janeiro: Interciência, 2019.

SILVEIRA, D. T.; CÓRDOVA, F.P. A pesquisa científica. *In*: GERHARDT, T. E.; SILVEIRA, D. T. (org.). **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009.

SPRINGER NATURE. **Scientific data**. 2020. Disponível em: <https://www.nature.com/sdata/>. Acesso em 15 jan.2020.

STALL, S.; YARMEY, L.; CUTCHER-GERSHENFELD, J.; HANSON, B.; LEHNERT, K.; NOSEK, B.; PARSONS, M. Make scientific data FAIR. **Nature**, v. 570, n. 7759. p. 27-29, jun. 2019. Disponível em: <https://www-nature.ez15.periodicos.capes.gov.br/articles/d41586-019-01720-7>. Acesso em: 25 jun. 2020.

STENT, C. Using DDI to drive data governance in the 21st century. *In*: Presented at the North American Data Documentation Initiative Conference (NADDI) 2018 Annual Conference

(NADDI2018), abr. 2018, Washington, D.C.: Zenodo. Disponível em: <http://doi.org/10.5281/zenodo.1230510>. Acesso em: 20 out. 2020.

STRASSER, C.; COOK, R.; MICHENER, W.; BUDDEN, A. **Primer on data management: what you always wanted to know**. California: CDL, 2012. Disponível em: <http://escholarship.org/uc/item/7tf5q7n3#page-1>. Acesso em: 13 ago. 2019.

THOMAS, W. DDI: Metadata to support collection processes, discovery, and comparability. *In: Workshop delivered at the North American Data Documentation Initiative Conference (NADDI 2013)*, apr. 4, 2013, University of Kansas, Lawrence, Kansas. Disponível em: <http://hdl.handle.net/1808/11060>. Acesso em: 13 out. 2021.

UNIVERSIDADE FUMEC. **Edições anteriores**. Disponível em: <http://www.fumec.br/revistas/sigc/issue/archive>. Acesso em: 25 maio 2020.

UNIVERSITY OF TORONTO FACULTY OF INFORMATION. Digital Curation. **What is Digital Curation?**. 2020. Disponível em: <http://dci.ischool.utoronto.ca/what-is-digital-curation/>. Acesso em: 22 jun. 2020.

VAN DEN EYNDEN, V., CORTI, L. Advancing research data publishing practices for the social sciences: from archive activity to empowering researchers. **International Journal on Digital Libraries**, v. 18, p. 113–121, 2017. Disponível em: <https://doi.org/10.1007/s00799-016-0177-3>. Acesso em: 25 out. 2020.

VARDIGAN, M. The DDI Matures: 1997 to the Present. **IASSIST Quarterly**, v. 37, p. 45-50. 2013. Disponível em: http://www.iassistdata.org/sites/default/files/iqvol371_4_vardigan.pdf. Acesso em: 20 jan. 2021.

VARDIGAN, M.; HEUS, P.; THOMAS, W. Data Documentation Initiative: toward a standard for the social sciences. **International Journal of Digital Curation**, v. 3, n. 1, p. 107-113, 2008. Disponível em: <http://dx.doi.org/10.2218/ijdc.v3i1.45>. Acesso em: 21 out. 2020.

VEIGA, V. S. O.; MACHADO, R. R.; JORGE, V. A.; PENEDO, E. Compartilhamento de dados de pesquisa na Fiocruz: diagnóstico e percepção do pesquisador. **Ciência da Informação**, Brasília, n.3, v.48, p.319-327, set./dez. 2019. Disponível em: <http://revista.ibict.br/ciinf/article/view/4999/4464>. Acesso em: 30 jun. 2020.

VICENTE-SAEZ, R.; MARTINEZ-FUENTES, C. Open science now: a systematic literature review for an integrated definition. **Journal of Business Research**, v. 88, p. 428-436, jul. 2018. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0148296317305441>. Acesso em: 25 jun. 2020.

WILKINSON, M.D. *et al.* The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 160018, 2016. Disponível em: <https://doi.org/10.1038/sdata.2016.18>. Acesso em: 25 jun. 2020.

ZIMMERMAN, A.; BOS, N.; OLSON, J. S.; OLSON, G. M. The promise of data in e-research: many challenges, multiple solutions, diverse outcomes. *In*: NICHOLAS, W. J. (ed.). **E-Research**: transformation in scholarly practice. Nova York: Routledge, 2009. p. 222-239.

APÊNDICE A - PLANO DE GESTÃO DE DADOS

A PRESENÇA DE ELEMENTOS DE CICLOS DE VIDA DOS DADOS NO PERCURSO METODOLÓGICO DAS TESES BRASILEIRAS DA ÁREA DE CIÊNCIA DA INFORMAÇÃO: Um estudo diagnóstico

COLETA DE DADOS

Quais os dados você coletará ou criará?

O tipo de dados coletados quanto à natureza corresponde a textos, números e visualizações, cujos resultados foram representados por meio de gráficos. Os dados foram também classificados como de Registro, no tocante à origem. A princípio, foi realizada a identificação e o *download* das teses brasileiras da Ciência da Informação para analisar os componentes do ciclo de vida dos dados em seus percursos metodológicos, assim como para identificar os tipos de dados gerados, no que se refere à classificação e à origem desses dados. Os textos que foram levantados remetem a um total de 565 teses disponíveis em formato pdf. Após a importação das metodologias das referidas teses para o *software* QDA Miner, versão 6.0, que foi utilizado no estudo, os dados gerados na própria pesquisa, possivelmente no volume de KB, foram armazenados no formato *comma-separated-values* - CSV, por ter a possibilidade de ser lido por vários aplicativos e por ser um formato não-proprietário. Logo, apesar do *software* utilizado na análise de dados não ser aberto, ele permitiu gerar os dados neste formato.

Como os dados serão coletados ou criados?

Os dados foram coletados por meio de uma pesquisa caracterizada como descritiva, bibliográfica, documental, quali-quantitativa. Para realizar a análise dos dados, foi adotada a técnica de análise de conteúdo.

A princípio, as teses levantadas foram separadas por pastas de arquivos no computador nomeadas de: Teses-PPGCINF-UNB, Teses-PPGCI-USP, Teses-PPGCI-UNESP, Teses-PPGCI-UFBA, Teses-PPGCI-UFPB, Teses-PPGCI-UFGM, Teses-PPGGOC-UFGM, Teses-PPGCIN-UFSC, Teses-PPGCI/IBICT-UFRJ, Teses-PPGCI-UFF, Teses-PPGSIGC-FUMEC, correspondentes aos programas em que foram encontradas teses disponíveis e por subpastas separadas por ano. Neste processo, de acordo com a ordem que as teses foram baixadas, elas

foram enumeradas de 1 a 565, para a seleção da amostra. Após esse momento, foram separadas as metodologias das 229 teses obtidas no processo amostral e agrupadas em uma nova pasta nomeada de “Metodologias-teses selecionadas”, em cada arquivo foi intitulado de Amostra_(nº), com a numeração correspondente. Na sequência, os arquivos foram importados em pdf para o *software* de análise qualitativa.

Os dados propriamente ditos foram coletados a partir da formação de categorias relacionadas aos elementos do ciclo de vida dos dados de referência. Para cada categoria, foram atribuídos códigos referentes às suas características. Para a categoria planejar, foram criados os códigos: plano de gestão de dados, plano de gerenciamento do projeto, campo de estudo, período de coleta, escolha de métodos e técnicas, recursos, questões éticas, equipe de trabalho, planejamento da descrição, planejamento da preservação, natureza dos dados e origem dos dados. A categoria coletar apresentou 3 (três) códigos: pesquisa bibliográfica, pesquisa de campo e pesquisa documental. A categoria organizar envolveu 2 (dois) códigos: Agrupamento e codificação e categorização. Na categoria descrever foi utilizado o código metadados. O código usado para representar a categoria assegurar a qualidade foi garantir a qualidade. Na etapa preservar, foi utilizado o código intitulado de preservação e curadoria. Na categoria encontrar, foi utilizado o código descobrir; na reutilizar, o código integrar; e, por fim, na categoria analisar, foram criados os códigos análise de dados e comparação de dados.

Para garantir a qualidade do trabalho, foi realizada a releitura dos 229 textos e das sentenças recuperadas. Neste momento, foram mantidas as sentenças que estão alinhadas com os códigos e descartadas as que não se enquadravam ou apresentavam duplicidade de informação. Além disso, algumas partes dos textos não foram lidas pelo *software*, porém, para garantir que todas as sentenças correlacionadas com as categorias fossem codificadas, o material original foi consultado e essas sentenças foram incluídas, de forma manual, no programa.

O *software* que foi utilizado não é licenciado de forma aberta, foi utilizada a versão teste gratuita por 30 dias e depois, em outro *laptop*, por mais 30 dias. Neste contexto, foi realizado o *backup* diário das planilhas geradas contendo informações relevantes sobre as categorias, os códigos, as metodologias que possuem os elementos, os trechos do texto que atestam a realidade, assim como os tipos de dados relacionados à natureza e origem dos dados das teses.

DOCUMENTAÇÃO E METADADOS

Que documentação e metadados acompanharão os seus dados?

Os dados serão acompanhados de uma documentação com as seguintes informações: quem criou os dados, os colaboradores, o editor, título do conjunto de dados, data da criação, idiomas das metodologias selecionadas, formato do conjunto de dados que será disponibilizado, extensão do arquivo, o assunto – que corresponde às palavras-chaves –, a fonte, o tipo de dados quanto à natureza e a origem e questões de direitos.

ÉTICA E CONFORMIDADE LEGAL

Como você gerenciará problemas éticos?

As teses utilizadas no presente estudo se limitarão às que se encontram disponíveis em acesso aberto na *Web*, ou seja, nos sites dos seus programas, repositórios e/ou bibliotecas digitais.

Como você gerenciará questões de direitos autorais e direitos de propriedade intelectual (DPI)?

A proprietária dos direitos autorais será da própria autora de todos os dados coletados na pesquisa. Os dados serão licenciados com uma [Licença Creative Commons - Atribuição 4.0 Internacional \(CC BY 4.0\)](#).

ARMAZENAMENTO E BACKUP

Como os dados serão armazenados em backup durante a pesquisa?

No decorrer da atividade da pesquisa, à medida que ocorreu a coleta de dados, eles foram armazenados em pastas e planilhas no *laptop* pessoal e no *google drive*, de maneira frequente, e a responsável pelo armazenamento foi a própria autora do trabalho.

Como você gerenciará o acesso e a segurança?

Após a aprovação do trabalho pela banca examinadora de doutorado, os dados da

pesquisa serão armazenados no repositório de dados Figshare, pela própria pesquisadora. Após o período de embargo de 1 (um) ano, os dados serão disponibilizados em acesso aberto para serem reutilizados em novos estudos.

SELEÇÃO E PRESERVAÇÃO

Quais dados são de valor a longo prazo e devem ser retidos, compartilhados e / ou preservados?

Os dados utilizados na pesquisa serão mantidos para validar os seus resultados, assim como para possibilitar novos estudos. Os dados que foram descartados no decorrer da análise, por apresentarem duplicidade de informação, não foram preservados.

Qual é o plano de preservação de longo prazo para o conjunto de dados?

Os dados serão preservados no repositório de dados Figshare, sem custo adicional. Os elementos descritores do conjunto de dados seguirão as orientações do repositório. Os dados brutos, utilizados na pesquisa, serão disponibilizados em planilha eletrônica no formato csv, por serem estáveis e não proprietários. O acesso aos dados pelos possíveis usuários (pesquisadores que investigam sobre a temática da gestão de dados científicos e os pesquisadores da área da Ciência da Informação) ocorrerá após o período de embargo de 6 (seis) meses, tempo necessário para realização de artigos e publicações, com uma licença considerada livre, por permitir que os futuros usuários possam utilizar o material em qualquer formato e fazer adaptações, transformações e novas construções com qualquer finalidade, desde que se inclua a atribuição dos direitos autorais.

COMPARTILHAMENTO DE DADOS

Como você compartilhará os dados?

Os dados serão compartilhados através do repositório de dados Figshare, após o período de embargo de 1 (um) ano.

Os dados também poderão ser disponibilizados por meio da publicação de artigos de dados (*data papers*) em um periódico de dados voltado para a área da Ciência da Informação.

Os dados serão disponibilizados para a comunidade interessada em investigar sobre a temática. Será considerada a aplicação de um identificador persistente, para que os dados possam ser encontrados de forma confiante e eficiente.

São necessárias restrições ao compartilhamento de dados?

Os dados só serão compartilhados após o período de embargo de 1 (um) ano, antes deste período, eles serão de uso exclusivo da pesquisadora e do seu orientador, para possíveis publicações.

RESPONSABILIDADES E RECURSOS

Quem será responsável pelo gerenciamento de dados?

A implantação do plano de gestão de dados é de responsabilidade da autora do trabalho, assim como a sua revisão para possíveis adequações necessárias. Neste contexto, a própria autora foi responsável por capturar os dados, produzir os metadados, zelar pela qualidade dos dados, fazer o armazenamento, *backup* dos dados, arquivamento e compartilhamento dos dados.

Quais recursos você precisará para entregar seu plano?

Foi realizado o *download* da versão teste de um *software* voltado à análise de conteúdo. Foi necessário o conhecimento técnico para operacionalizá-lo, sendo preciso um treinamento para utilização de tal *software* através de vídeos disponibilizados na *internet*.

APÊNDICE B – LISTA CÓDIGO/REFERÊNCIA

Código	Referência
Amostra_02	MACEDO, Patricia Ladeira Penna. Um estudo sobre o princípio da ordem original em arquivos pessoais . 2018. Tese (Doutorado em Ciência da Informação) – Universidade Federal Fluminense, Niterói, 2018.
Amostra_19	VIANNA, Jaqueline Abreu. O trabalho mediado por TIC - Tecnologias de Informação e Comunicação e seus efeitos sobre o trabalhador. 2012. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2012. Disponível em: http://hdl.handle.net/1843/ECID-8XLKUL . Acesso em: 22 jul. 2021.
Amostra_21	MEIRELES, Magali Rezende Gouvêa. Categorização de documentos a partir de suas citações: um método baseado em redes neurais artificiais . 2012. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2012. Disponível em: http://hdl.handle.net/1843/ECID-92APE4 . Acesso em: 22 jul. 2021.
Amostra_26	FREITAS, Adolfo Júlio Porto de. A informação como fundamento para formação de novos padrões de cultura cívica no âmbito do Orçamento Participativo-(OP) . 2013. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2013. Disponível em: http://hdl.handle.net/1843/BUBD-9EFGDG . Acesso em: 21 jul. 2021.
Amostra_41	MAIA, Regina Maria da Costa Smith. Intercomplementação do sistema categorial de Ranganathan e do modelo entidade-relacionamento de Chen para a modelagem conceitual: uma aplicação no domínio do biomonitoramento do Projeto Manuelzão/UFGM nas águas da Bacia do Rio das Velhas . 2013. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2013. Disponível em: http://hdl.handle.net/1843/ECIC-9CQMKL . Acesso em: 22 jul. 2021.
Amostra_47	TEIXEIRA, Clotildes Madalena de Avelar. Informação, educação patrimonial e museus no ciberespaço: diálogos . 2014. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2014. Disponível em: http://hdl.handle.net/1843/BUBD-9WETEQ . Acesso em: 22 jul. 2021.
Amostra_68	BARBOSA, Daniel Mendes. Um modelo de gestão da informação e do conhecimento para o contexto da avaliação de cursos de graduação . 2016. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2016. Disponível em: http://hdl.handle.net/1843/BUOS-AWCNP9 . Acesso em: 21 jul. 2021.
Amostra_69	PACHECO, Katia Lucia. Obra e instâncias na organização da informação musical: estudo da adequação do modelo conceitual FRBR . 2016. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2016. Disponível em: http://hdl.handle.net/1843/BUBD-AKENHX . Acesso em: 21 jul. 2021.
Amostra_74	PAULA, Lorena Tavares de. Nanopublicação e indexação: interlocuções semânticas, pragmáticas e discursiva em aplicações metodológicas . 2016. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2016. Disponível em: http://hdl.handle.net/1843/BUBD-AE7LJW . Acesso em: 21 jul. 2021.
Amostra_78	DUTRA, Daniel Marcos Resende. Efeitos das fontes externas de informação no desempenho inovador das empresas . 2017. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas

	Gerais, Belo Horizonte, 2017. Disponível em: http://hdl.handle.net/1843/BUBD-B49HL5 . Acesso em: 21 jul. 2021
Amostra_79	RODRIGUES, Kátia de Oliveira. Fatores que influenciam o comportamento de citação de docentes-pesquisadores do campo da Cancerologia brasileira . 2017. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2017. Disponível em: http://hdl.handle.net/1843/BUOS-AV3HA9 . Acesso em: 22 jul. 2021.
Amostra_80	ADAO, Sheila Margareth Teixeira. 1ª CNARQ: a Conferência Nacional de Arquivos e a construção de uma política nacional para os arquivos brasileiros . 2017. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2017. Disponível em: http://hdl.handle.net/1843/BUBD-AXWLKC . Acesso em: 22 jul. 2021.
Amostra_85	GANDRA, Tatiane Krempser. Práticas informacionais dos visitantes do Museu Itinerante Ponto UFMG . 2017. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2017. Disponível em: http://hdl.handle.net/1843/BUOS-AV2EZJ . Acesso em: 21 jul. 2021.
Amostra_89	SANTOS, Silvana Aparecida Silva dos. Arquivos Universitários: a gestão eficaz e a preservação digital. Um estudo dos sistemas acadêmicos e de recursos humanos da Universidade Federal de Minas Gerais-UFMG . 2018. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2018. Disponível em: http://hdl.handle.net/1843/BUOS-B5EJY9 . Acesso em: 21 jul. 2021.
Amostra_91	LEMOS, Ariane Barbosa. Comunicação organizacional e gestão do conhecimento: práticas, interface e modelagem . 2018. 175 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2018. Disponível em: http://hdl.handle.net/1843/VAFA-BB5GWS . Acesso em: 20 jul. 2021.
Amostra_93	AGUIAR, Niliane Cunha de. O letramento para a competência informacional em bibliotecas escolares: estudo a partir dos projetos políticos-pedagógicos dos colégios de aplicação das universidades federais brasileiras . 2018. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2018. Disponível em: http://hdl.handle.net/1843/BUOS-B6PFSK . Acesso em: 22 jul. 2021
Amostra_106	SILVA, Luiz Carlos da. Desafios da implementação da política pública de arquivos: as funções arquivísticas e o acesso à informação na prefeitura municipal de Vitória, ES . 2020. 263 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2020. Disponível em: http://hdl.handle.net/1843/33628 . Acesso em: 20 jul. 2021.
Amostra_114	OLIVEIRA, Jaury Nepomuceno. Acesso livre e direito de autor: a comunicação científica eletrônica na ciência da informação no Brasil e em Portugal . 2013. Tese (Doutorado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013. Disponível em: http://ridi.ibict.br/bitstream/123456789/671/1/oliveira2013.pdf . Acesso em: 21 jul. 2021.
Amostra_116	RIOS, Patrícia Andréa do Prado. Difusão intrafirma da inovação em organizações baseadas em projetos: o caso da Petrobras . 2013. 189 f. Tese (Doutorado em Ciência da Informação) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013. Disponível em: http://ridi.ibict.br/handle/123456789/901 . Acesso em: 21 jul. 2021.
Amostra_119	SOUZA, Joice Cleide Cardoso Ennes de. Banco de imagens: abordagem teórica conceitual de representação de fotografias para uso na publicidade . 2013. Tese

	(Doutorado em Ciência da Informação) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013. Disponível em: http://ridi.ibict.br/bitstream/123456789/895/1/Tese-JoiceCardoso.pdf . Acesso em: 21 jul. 2021.
Amostra_126	MORAES, Julia Nolasco Leitão de. Museu, informação artística e “poesia das coisas” : a divulgação artística em museus de arte. 2014. Tese (Doutorado em Ciência da Informação) - Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2014. Disponível em: http://ridi.ibict.br/handle/123456789/807 . Acesso em: 20 jul. 2021.
Amostra_134	LOVÓN CANCHUMANI, Roberto Mario. Domínios científicos na UFRJ : mapeamento de áreas de conhecimento. 2015. 185 f. Tese (Doutorado em Ciência da Informação) – Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015. Disponível em: http://ridi.ibict.br/handle/123456789/799 . Acesso em: 20 jul. 2021.
Amostra_135	BENCHIMOL, Alegria Célia. Resgate e ressignificação da pesquisa no Museu Paraense Emílio Goeldi : presença e permanência de cientistas estrangeiros (1894-1914) na produção científica de autores atuais (1991-2010). 2015. Tese (Doutorado em Ciência da Informação) - Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015. Disponível em: http://repositorio.ibict.br/bitstream/123456789/847/1/Tese%20Alegria%20Benchimol%20.pdf . Acesso em: 21 jul. 2021.
Amostra_140	MACIEL, Ariane Durce. Gênero e inclusão digital : uso e apropriação das TICs pelos usuários do programa federal GESAC. 2015. Tese (Doutorado em Ciência da Informação) - Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015. Disponível em: http://ridi.ibict.br/bitstream/123456789/878/1/TeseArianeMaciel.pdf . Acesso em: 21 jul. 2021.
Amostra_141	Santos, Anne Danielle Soares Clinio dos. Novos cadernos de laboratório e novas culturas epistêmicas : entre a política do experimento e o experimento da política. 2016. 240 f. Tese (Doutorado em Ciência da Informação) – Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016. Disponível em: https://ridi.ibict.br/handle/123456789/943 .
Amostra_144	CRUZ, Emília Barroso. Acesso e gestão da informação governamental do tipo arquivístico : aspectos técnicos e a visão de gestores e servidores de duas instituições governamentais. 2016. Tese (Doutorado em Ciência da Informação) - Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016. Disponível em: http://ridi.ibict.br/handle/123456789/882 . Acesso em: 21 jul. 2021.
Amostra_153	NISENBAUM, Moisés André. A transposição científica da Física para a Ciência da Informação . 2017. Tese (Doutorado em Ciência da Informação) - Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2017. Disponível em: http://ridi.ibict.br/handle/123456789/958 . Acesso em: 21 jul. 2021.
Amostra_164	DOURADO, Stella Moreira. O almanaque enquanto documento de informação e comunicação popular escrita : a coleção da família Carneiro Rezende. 2018. 163 f. Tese (Doutorado em Ciência da Informação) - Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2018. Disponível em: http://ridi.ibict.br/handle/123456789/965 . Acesso em: 20 jul. 2021.
Amostra_175	MIGLIOLI, Sarah. A surdez como ciência no Brasil : parâmetros de organização e representação do conhecimento. 2019. (Doutorado em Ciência da Informação) –

	Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2019. Disponível em: http://ridi.ibict.br/handle/123456789/1001 . Acesso em: 21 jul. 2021.
Amostra_176	SANTOS, Nádia Bernuci dos. Diferenças de gênero na apropriação das Tecnologias de Informação e Comunicação : um olhar a partir do ambiente do ensino médio do Colégio Pedro II. 2019. 177 f. Tese (Doutorado em Ciência da Informação) – Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2019. Disponível em: http://ridi.ibict.br/handle/123456789/1017 . Acesso em: 21 jul. 2021.
Amostra_183	LEE, Patrícia Pui Yue. A Memória Científica e Tecnológica da Biblioteca do Instituto Nacional de Tecnologia (INT) sob a perspectiva do regime de informação . 2019. 174 f. Tese (Doutorado em Ciência da Informação) – Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2019. Disponível em: https://ridi.ibict.br/handle/123456789/1090 . Acesso em: 21 jul. 2021.
Amostra_200	INOMATA, Danielly Oliveira. Redes colaborativas em ambientes de inovação : uma análise dos fluxos de informação. 2017. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Santa Catarina, Florianópolis, 2017. Disponível em: https://repositorio.ufsc.br/xmlui/handle/123456789/182585 . Acesso em: 21 jul. 2021.
Amostra_203	SEMLER, Alexandre Ribas. Ciência da informação em contextos de e-science : bibliotecários de dados em tempos de Data Science. 2017. 165 f. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Santa Catarina, Florianópolis, 2017. Disponível em: https://repositorio.ufsc.br/handle/123456789/185593 . Acesso em: 22 jul. 2021.
Amostra_209	GARCIA, Cristiane Luiza Salazar. Fragmentos teóricos de domínios de pesquisa da ciência da informação : perspectiva metateórica para gestão do conhecimento e competência em informação. 2018. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Santa Catarina, Florianópolis, 2018. Disponível em: https://repositorio.ufsc.br/handle/123456789/198504 . Acesso em: 22 jul. 2021.
Amostra_224	MELLO, Josiane. Proposta de modelo para a preservação e curadoria digital de objetos digitais de centros de pesquisas oncológicas . 2020. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Santa Catarina, Florianópolis, 2020. Disponível em: https://repositorio.ufsc.br/handle/123456789/216123 . Acesso em: 22 jul. 2021.
Amostra_225	SILVA, Luciana Mara. Repositório institucional e o ecossistema da ciência aberta : mecanismos de funcionamento. 2020. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Santa Catarina, Florianópolis, 2020. Disponível em: https://repositorio.ufsc.br/handle/123456789/219442 . Acesso em: 21 jul. 2021.
Amostra_229	COELHO, Kátia Cardoso. Princípios ontológicos dos atos dos documentos aplicados à modelagem demo : uma investigação no contexto da ontologia corporativa. 2017. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2017. Disponível em: http://hdl.handle.net/1843/ECIP-AZWNKV . Acesso em: 21 jul. 2021.
Amostra_235	SILVA, Patricia Nascimento. Dados governamentais abertos : métricas e indicadores de reuso. 2018. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2018. Disponível em: https://repositorio.ufmg.br/handle/1843/BUBD-AYNG4U . Acesso em: 21 jul. 2021.
Amostra_238	GOMES, Gracielle Mendonça Rodrigues. Expansão das fronteiras da educação na Ciência da Informação : a formação de profissionais para atuação no contexto

	dos dados de pesquisa. 2019. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2019. Disponível em: http://hdl.handle.net/1843/31349 . Acesso em: 21 jul. 2021
Amostra_241	ALMEIDA, Fernanda Gomes. Suporte à gestão de dados de pesquisa : uma ampliação dos serviços oferecidos pelas bibliotecas. 2019. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2019. Disponível em: http://hdl.handle.net/1843/31596 . Acesso em: 21 jul. 2021
Amostra_247	DUTRA, Frederico Giffoni de Carvalho. Monitoramento e uso de informações sobre clientes e marcas em mídias e redes sociais digitais : um estudo em empresas públicas. 2020. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2020. Disponível em: http://hdl.handle.net/1843/34091 . Acesso em: 21 jul. 2021.
Amostra_255	SOUZA, Marcos de. O comportamento de termos da Ciência da Informação por meio da modelagem de tópicos . 2020. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2020. Disponível em: http://hdl.handle.net/1843/34292 . Acesso em: 21 jul. 2021
Amostra_257	REIS, Ronara Cristina Bozi dos. Universidade, território e inovação : construção de identidade na economia da informação e do conhecimento. 2020. Tese (Doutorado em Sistemas de Informação e Gestão do Conhecimento) - Faculdade de Ciências Empresariais, Universidade FUMEC, Belo Horizonte, 2020. Disponível em: https://repositorio.fumec.br/xmlui/handle/123456789/425 . Acesso em: 22 jul. 2021.
Amostra_269	SILVA, Elieny do Nascimento. A responsabilidade social da biblioteconomia nas ações de extensão universitária . 2017. Tese (Doutorado em Ciência da Informação) - Universidade Federação da Bahia, Salvador, 2017. Disponível em: https://repositorio.ufba.br/ri/handle/ri/25357 . Acesso em: 20 jul. 2021.
Amostra_275	SOBRAL, Natanael Vitor. Pesquisadores em doenças tropicais negligenciadas no Brasil : produção científica e convergências com o Plano Nacional de Saúde (2016 a 2019). 2019. Tese (Doutorado em Ciência da Informação) - Instituto de Ciência da Informação, Universidade Federal da Bahia, Salvador, 2019. Disponível em: http://repositorio.ufba.br/ri/handle/ri/31156 . Acesso em: 22 jul. 2021.
Amostra_293	CÂMARA, Rafael Silva da. Institucionalização científica da economia política da informação : contributos sociais e cognitivos na produção científica do campo da Ciência da Informação no Brasil. 2020. Tese (Doutorado em Ciência da Informação) - Universidade Federação da Paraíba, João Pessoa, 2020. Disponível em: https://repositorio.ufpb.br/jspui/handle/123456789/18301 . Acesso em: 20 jul. 2021.
Amostra_295	ANDRADE, Robéria de Lourdes de Vasconcelos. Perspectivas da publicação de livros em formato digital pelas editoras universitárias na América Latina . 2020. Tese (Doutorado em Ciência da Informação) - Universidade Federal da Paraíba, João Pessoa, 2020. Disponível em: https://repositorio.ufpb.br/jspui/handle/123456789/18133 . Acesso em: 22 jul. 2021.
Amostra_305	ARANHA, Sofia Cristina Adjuto Daher. Estudo sobre o fomento à pesquisa em dengue no Brasil : contribuições para a aproximação entre a geração e o uso do conhecimento. 2012. 269 f. Tese (Doutorado em Ciência da Informação) - Universidade de Brasília, Brasília, 2012. Disponível em: https://repositorio.unb.br/handle/10482/11919 . Acesso em: 21 jul. 2021.

Amostra_307	AFONSO, Alexandre Ribeiro. B2: Um sistema para indexação e agrupamento de artigos científicos em português brasileiro utilizando computação evolucionária. Brasília, 2013. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, 2013. Disponível em: https://repositorio.unb.br/handle/10482/15480 . Acesso em: 21 jul. 2021.
Amostra_315	CRUZ, Felipe Lopes da. O impacto da satisfação das necessidades de informação na tomada de decisão inerente ao planejamento estratégico de uma organização pública federal. 2013. Tese (Doutorado em Ciência da Informação) - Universidade de Brasília, Brasília, 2013. Disponível em: https://repositorio.unb.br/handle/10482/15150 . Acesso em: 21 jul. 2021.
Amostra_321	PIMENTA, Shirley Guimarães. A relação entre informação e a estrutura de conhecimento na resolução de problemas: estudo experimental. 2013. Tese (Doutorado em Ciência da Informação) — Universidade de Brasília, Brasília, 2013. Disponível em: http://repositorio.unb.br/handle/10482/15714 . Acesso em: 21 jul. 2021.
Amostra_323	LEMONS, Wilda Soares. Mediação da informação técnica para produtores de leite da região oeste goiana em programas de formação. Brasília, 2013. Tese (Doutorado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília. Disponível em: http://repositorio.unb.br/handle/10482/13676 . Acesso em: 21 jul. 2021.
Amostra_331	SOUZA, Emílio Evaristo de. A organização da informação e o ensino técnico do DF: um modelo para promover o feedback para professores nas salas de aula. 2015. Tese (Doutorado em Ciência da Informação) — Universidade de Brasília, Brasília, 2015. Disponível em: http://repositorio.unb.br/handle/10482/19587 . Acesso em: 21 jul. 2021.
Amostra_338	ARAUJO, Lauro César. Uma linguagem para formalização de discursos com base em ontologias. 2015. Tese (Doutorado em Ciência da Informação) — Universidade de Brasília, Brasília, 2015. Disponível em: http://dx.doi.org/10.26512/2015.11.T.19319 . Acesso em 21 jul. 2021.
Amostra_351	VILARINHO, Tatiane Ferreira. Padrões de comunicação científica nas ciências sociais: a interdisciplinaridade entre economia, ciência política, sociologia e antropologia. 2015. Tese (Doutorado em Ciência da Informação) — Universidade de Brasília, Brasília, 2015. Disponível em: http://dx.doi.org/10.26512/2015.11.D.20504 . Acesso em: 21 jun. 2021.
Amostra_373	ALBUQUERQUE, Fernando Antônio de Araújo Chacon de. Arcabouço de arquitetura da informação para ciclo de vida de projeto de vocabulário controlado: uma aplicação em Engenharia de Software. 2017. Tese (Doutorado em Ciência da Informação) — Universidade de Brasília, Brasília, 2017. Disponível em: http://repositorio.unb.br/handle/10482/31288 . Acesso em: 21 jul. 2021.
Amostra_375	MOREIRA, Jonathan Rosa. As áreas de informação no Brasil: um estudo bibliométrico sobre características dos grupos de pesquisa e a produção científica de seus membros. 2017. 354 f. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Brasília, 2017. Disponível em: https://repositorio.unb.br/handle/10482/23676 . Acesso em: 20 jul. 2021.
Amostra_399	MALHEIROS, Tania Milca de Carvalho. Produtos e serviços de informação para pessoas com deficiência visual. 2019. Tese (Doutorado em Ciência da Informação) — Universidade de Brasília, Brasília, 2019. Disponível em: http://repositorio.unb.br/handle/10482/34969 . Acesso em: 21 jul. 2021.
Amostra_404	SOUZA, Flávia Moreno Alves de. Conhecimento, atitude e prática no contexto da epidemia do HIV/aids: uma abordagem da Ciência da Informação. 2020. Tese (Doutorado em Ciência da Informação) — Universidade de Brasília, Brasília, 2020. Disponível em: https://repositorio.unb.br/handle/10482/38646 . Acesso em: 21 jul. 2021.

Amostra_411	SILVA, Irisneide de Oliveira Souza. A organização e a representação do conhecimento no domínio da arquivística . 2012. 193 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2012. Disponível em: http://hdl.handle.net/11449/103377 . Acesso em 21 jul. 2021.
Amostra_413	SOUZA, Patrícia da Silva Moreno e. SRDigital : proposta de um modelo baseado na linguagem natural e controlada como instrumentos de apoio ao agente computacional do processo de referência. 2012. 125 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências de Marília, Universidade Estadual Paulista, Marília, 2012. Disponível em: http://hdl.handle.net/11449/103367 . Acesso em: 20 jul. 2021.
Amostra_429	ARBOIT, Aline Elis. O processo de institucionalização sociocognitiva do domínio de Organização do Conhecimento a partir dos trabalhos científicos dos congressos da ISKO . 2014. 285 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista Júlio de Mesquita Filho, Marília, 2014. Disponível em: http://hdl.handle.net/11449/123389 . Acesso em: 20 jul. 2021.
Amostra_436	ELLIOTT, Ariluci Goes. A Fé documentada : perspectivas metodológicas de organização da informação fotográfica sobre romarias de Juazeiro do Norte – Ceará. 2014. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista Júlio de Mesquita Filho, Marília, 2014. Disponível em: http://hdl.handle.net/11449/114035 . Acesso em: 20 jul. 2021.
Amostra_440	MARTINS, Gracy Kelli. Institucionalização cognitiva e social da organização e representação do conhecimento na ciência da informação no Brasil . 2014. 184 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências de Marília, Universidade Estadual Paulista, Marília, 2014. Disponível em: http://hdl.handle.net/11449/114036 . Acesso em: 21 jul. 2021.
Amostra_457	MARTINS, Valéria dos Santos Gouveia. Indicadores de bibliotecas universitárias : criação de índice de desenvolvimento baseado em redes de conhecimento compartilhado. 2016. 160 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2016. Disponível em: http://hdl.handle.net/11449/137835 . Acesso em 21 jul. 2021.
Amostra_463	PÖTTKER, Luciana Maria Vieira. Arquitetura para recuperação de objetos de aprendizagem : uma abordagem baseada em agentes inteligentes e relevance feedback. 2017. 176 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2017. Disponível: http://hdl.handle.net/11449/150090 . Acesso em: 22 jul. 2021.
Amostra_468	SILVA, Marli Vitor da. O comportamento de busca de informação : uma análise a partir da teoria do capital social de Nan Lin. 2017. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2017. Disponível em: http://hdl.handle.net/11449/150016 . Acesso em: 22 jul. 2020.
Amostra_483	RANIERI, Clemilton Luís Bassetto. A competência em informação como elemento inovador no apoio às micro e pequenas empresas : uma modelagem teórico-prática aplicável aos programas de capacitação do SEBRAE/SP. 2018. 260 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2018. Disponível em: http://hdl.handle.net/11449/153805 . Acesso em: 21 jul. 2021.
Amostra_499	NAKANO, Natália. Princípios do design da informação na curadoria digital de ambientes virtuais de aprendizagem sob a perspectiva da ciência da informação . 2019. 165 f. Tese (Doutorado em Ciências da Informação),

	Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2019. Disponível em: http://hdl.handle.net/11449/181518 . Acesso em: 21 jul. 2021.
Amostra_502	PADUA, Mariana Cantisani. Design da informação e interação : compartilhamento de informações em ambientes digitais de museus. 2019. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2019. Disponível em: http://hdl.handle.net/11449/181306 . Acesso em: 22 jul. 2020.
Amostra_543	PAIVA, Simone Borges. Oficinas intergeracionais : saberes e fazeres da experiência, mediação cultural e significação. 2015. Tese (Doutorado em Ciência da Informação) - Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2015. Disponível em: doi: 10.11606/T.27.2015.tde-13112015-092819. Acesso em: 22 jul. 2021.
Amostra_557	AGUIAR, Francisco Lopes de. Dspace e archivematica : concepção e criação de um protótipo de repositório digital aplicado no domínio da SBPC. 2018. Tese (Doutorado em Ciência da Informação) - Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2018. Disponível em: doi:10.11606/T.27.2018.tde-11092018-095527. Acesso em: 2021-07-14.
Amostra_564	SOUZA, Soraia Herrador Costa Lima de. Ágoras pós-modernas : uma análise dos mecanismos de participação social nos espaços informacionais digitais. 2019. Tese (Doutorado em Ciência da Informação) - Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2019. Disponível em: doi:10.11606/T.27.2019.tde-07112019-160134. Acesso em: 22 jul. 2021.